Research on the Prediction and Influencing Factors of House Price Based on Regression Analysis Model

Jiaxin Zhang[1, *]

[1]Urban Construction College, Beijing University of Technology, Beijing, China

*Corresponding author: hanjing.zh@ccb.com

Abstract. To analyze what factors may have an impact on house price, this research used the house sales information from king county in U.S from May 2014 to May 2015 including house prices and 16 other factors to see whether they are correlated with house price or not. To minimize the error the research pre-processes the initial data and runs an endogeneity analysis among the independent variables. The research used single regression models along with multiple regression models to analyze the data. The research found out that the square feet of the living and the surrounding environment as well as the grade of the house are highly correlated with the house price. The research then suggests that the government should pay attention to these factors when building a house and the real estate company should adjust their marketing strategy based on this model in order to promote the overall value of the house.

Keywords: House price; Regression model; Factors analysis.

1. Introduction

1.1 Background

With the development of urbanization in China, the house sales market and the real estate market have been growing rapidly, which caused an increase in house price [1]. While there are many different factors that can influence the house price, one of the most important factors is the property of the house which is highly related to a comfortable and healthy living environment for the resident. As a developed country, the urbanization of U.S cities is much higher and also much earlier than China.So, U.S is a very effective sample to predict the future trend of Chinese house market, by analyzing one of the largest counties in U.S (king county), the research hopes to find out what factor significantly affects the house price in U.S, and how should the government or the real estate company adjust their marketing strategy according to it.

1.2 Related research

1.2.1 The difference between U.S and China

The fact that the Chinese market has unique characteristics has reached consensus among people, Ouyang Wenhe & Zhang Xuan argues that in terms of the type of system, U.S implement diversified land ownership. As long as it is in the range of restriction and law, Land in various forms of ownership can be freely bought or sold, thus makes the U.S have a market-oriented foundation. While China Implement socialist public ownership of land, the land is not allowed to be privatized. Therefore the price and the supply of the land depends on the policy of the government [2]. The research done by Liu Lin & Zhang Hongrui (2018) also shows that affected by the current stage of economic development, China's real estate policy changes more frequently and more intense so the impact of housing prices by policies is more significant, Unlike the real estate market in U.S which is dominated mostly by supply and demand and market laws [3].

1.2.2 The two dimensions of real estate

There are much research focusing on the two-dimension part of the real estate. Li Jian & Deng Ying indicates that the house (real estate) can be seen as both substance asset and a financial asset, so it will damage the economy if the house price fluctuates [4]. And in Li Xiaoxia, Yang Lin's research, the real economy refers to the economic activities of production, sale and provision. While

the virtual economy refers to the holding and trading activities of virtual capital [5]. The research then argues that although the relationship between real economy and the virtual is interdependent and mutually restrictive, the real asset is really the fundamental base, without which, the virtual cannot exist. In terms of the house, although the house price is high in correlation with the virtual economy, we should focus more on the property of the house itself because that is the firm base. We should see that the development of the virtual economy must be compatible with the development of the real economy, advanced development of the virtual economy cannot drive the super-rapid development of the real economy, but will cause a bubble economy which will damage the real economy eventually [5]. As the research of Guo Lingxia, Peng Qianying, Chen Lulu, Chen Ying & Yan Rui shows that the real estate bubble will affect the economy, for example in In Japan, the massive housing bubble after 1990 caused a two-decade economic depression in the country [6]. Many experts agreed that the housing bubble is more of a negative impact rather than a positive impact.

1.2.3 Geographical factors

According to Aini Zhong, Xing & Yu Jingmin. (2022), Housing prices are a highly complex problem under the influence of multiple factors, and their change is non-linear, dynamic and accidental. Although the traditional research of the house price mainly focuses on Characteristics and trends of the sales, most of which is based on the results of house price statistics and experience in the previous period. But on a much smaller scale, however, traditional research methods are not applicable anymore[1]. They came to the conclusion that residents not only consider the internal nature of the house itself, but also pay more attention to transportation, public services, and future planning related to the Geographical location of the house [1]. Another example is the research of Mao Yating (2022) which also shows that the pursuit of a better living environment has led to the capitalization of the waterfront of the house, whether the house has a waterfront has a huge impact on the price overflow rate [7].

1.3 purpose of the research

The research hopes to finds out the factors that significantly impact house price by using regression strategies. The purpose of this research is to build a model that can not only explain the previous trend of the house price but also forecast the future flow in a certain degree, so the estate company or government can use it as a reference when assessing a house and use it to adjust their cooperate and marketing strategy. In addition, although the market between U.S and China is very different, It is still a useful example for China to reference along the development of Chinese urbanization.

2. Data and methods

2.1 Data description

This dataset is from Kaggle, and contains house sale prices for King County. It includes houses sold between May 2014 and May 2015. There are totally 21614 sets of house information in this dataset, there are multiple factors in this dataset which include the house sold price, number of bedrooms, number of bathrooms, the square feet of living space, square feet of the lot, the waterfront, the condition of the house, view, grade, square feet of the place above ground, square feet of the basement, Zip code, latitude, longitude and the year of the house being built.

2.2 Method and Strategy

There are many variables that are correlated with each other, some of the relationships are confimed and can be described by certain mathematic equations. While others however, the relationship between their variables is non-deterministic, so the effective way to describe this relationship is to build regression models which is basically some simple functions, such as polynomials, to approach this relation [8].

This research uses a linear regression model to analyze the dataset. both single linear regressions and multiple linear regressions have been used in our research. The single variable linear regression model is as follows, equation (1):

$$\text{House price} = C + \beta \cdot x + \varepsilon \tag{1}$$

In the model, C stands for the constant variables including all the controlled variables, β is the coefficient of x, while x stands for the factors that may have an impact on house prices. And ε stands for the error in this model.

Multiple linear regression models are as follows, equation (2):

$$\text{House price} = C + \beta 1 \cdot x1 + \beta 2 \cdot x2 + \beta 3 \cdot x3 + \cdots + \varepsilon \tag{2}$$

While in this model, C and ε stand for the same meaning as above, the difference is that all the independent variables are added in the same equation this time to see the conjoint influence of them together on the house price

2.3 Accuracy analysis

2.3.1 Data pre-processing

Before running the model, it is necessary to pre-process the Data. In this research, the Dataset contains a total amount of 21614 sets of data, some of them have a very extreme price which is not suitable for our data analysis as illustrated in Fig. 1, by removing the extreme ones from the dataset based on their difference between them and the average price (540182) can increase the accuracy of the analysis.

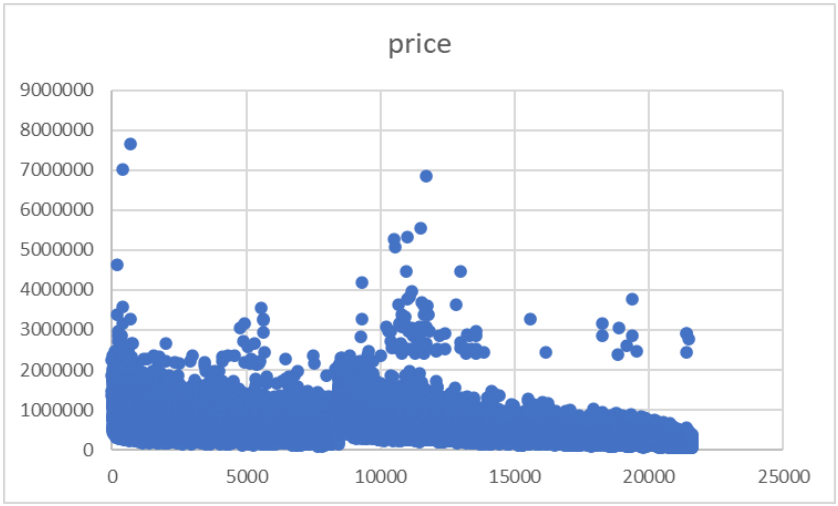

Fig. 1 Distribution of house price scatter plot（Photo credit: Original ）

Additionally, the house price can be massively huge but the factors however are not. Thus, the unit are not correspondingly matched. Therefore, this research uses the following equation (3) to match all the units of the data and set their range between 0 and 1 [9]:

$$Xi' = \frac{Xi - Xmin}{Xmax - Xmin} \tag{3}$$

2.3.2 About single variable regression model

While using the single variable regression model, the coefficient of x is the key factor to decide the correlation level of x to y. If the coefficient(β) is less than 0.1, then x is not strongly correlated with y. On the opposite, if the coefficient is very large, then x is a high correlation with y.

2.3.3 About multiple linear regression model

While using the multiple linear regression model, P-value and R-square are the vital factors. P-value shows whether x has a significant impact on y and R-square shows how well our model fits with the dataset.

3. Results

In the research, the data can be mainly separated into three types: the factor about the house itself, the factors of the surrounding environment and another factor. The following results will be discussed based on these three types.

3.1 The impact of the house itself on price

For this part, we will discuss the impact of house's own property on house prices. The property of the house includes the number of bedrooms, number of bathrooms, square-feet of the living space, square-feet of the lot, number of floors, condition of the house, square-feet of above ground and square-feet of the basement. We analyzing these factors individually and the resulting Table 1 is as follows:

Table 1. House's own property

Factors	Coefficient
bedrooms	1.472224
bathrooms	0.716385
sqft_living	1.361041
sqft_lot	0.522427
condition	0.032397
floors	0.168156
sqft_above	0.88481
sqft_basement	0.368541

Based on the result chart, the number of bedrooms, number of bathrooms and square-feet of living and square-feet above shows a high and also positive coefficient which indicates that customers consider these factors more important than others when assessing the house. While square-feet of the lot, square-feet of the basement shows a relatively moderate positive coefficient means that they are the less important value. At last, the condition of the house and the number of floors has a small coefficient, their correlation with house price is weak. So, they will be removed in the next multiple linear analysis.

3.2 The impact of geographic location and surrounding environment on house price

The geographic location of the house and the surrounding environment are also very critical. With the development of the economy, people pay more attention to the environment around the house to have a better quality of living. These types of factors include waterfront, view, Zip code, latitude and longitude. And the resulting Table 2 is as follows:

Table 2. Surrounding environment

factors	coefficient
waterfront	0.2948
view	0.259814
Zipcode	-0.0232
lat	0.205693
long	0.038724

The resulting Table illustrates that all these factors' coefficients are not very large which shows that the geographic location and the surrounding environment are not highly correlated with house price. But some of them like waterfront, view and also latitude is still moderate related and positive. While the others, Zip code, and longitude have no relation to the house price at all, therefore they will be removed in the following multiple regression.

3.3 Some other factors

There are also some other factors worth studying in this dataset. Including the grade of the house, the year the house was built and the year the house was renewed. And the result in Table 3 are as follows:

Table 3. Other factors

factors	coefficient
grade	0.936927
yr_built	0.030029
yr_renewed	0.08183

Based on this sheet, the grade of the house has a relatively high and positive coefficient while the year the house was built and the year the house was renewed have a low coefficient. That means that the grade is more related to the house price.

3.4 Multicollinearity problems

Multicollinearity refers to the phenomenon of linear correlation between independent variables, The reasons for multiple correlations between independent variables mainly include four aspects: (1) there is a common trend of economic variables over time; (2) Modeling with cross-sectional data; (3) A large number of lagging variables are used in the model; (4) Improper selection of variables due to the limitations of understanding during modeling [10]. It is certain that multicollinearity problems exist in this model since there are many independent variables that can interact with each other. So, the researchers decided to run a regression individually between these independent factors to see if the coefficient is very high or the P-value and R-square of them show any significance. The result came out to be that between square feet of living and square feet of the above also basement is very high, the one explanation is that the square feet of living include the rest of two factors, so if the square feet of living space expands, the square feet of the basement and square feet of above also increases. In addition, the coefficient of stationing is 1.36 which is much higher than the rest of the two in terms of the impact on the house price. So, the researchers decided to remove the sqft of above and basement only keep the sqft of living.

Another interaction between independent factors is view and grade, apparently, a good view of the house can increase the over-all grade of the house, while the grade has a higher coefficient of 0.93 compared to view's coefficient 0.26 in terms of the impact on house price, the researcher decides to remove the view factor in the following multiple linear model.

3.5 Multiple linear regression

When analyzing the data with multiple linear regression model, it is necessary to remove the unrelated factors that are mentioned in the previous single regression models. The remaining related factors are bedroom numbers, bathroom numbers, square feet of living, square feet of the lot, grade of the house and latitude.

The resulting Table 4 and Table 5 are as follows:

Table 4. The result of multiple linear regression model

	Coefficients	Standard error	t Stat	P-value
Intercept	-0.25828	0.004506	-57.3206	0
bedrooms-	-0.22719	0.024513	-9.26819	2.07E-20
bathrooms-	-0.0573	0.008968	-6.38879	1.71E-10
sqft_living-	0.944026	0.015892	59.4025	0
sqft_lot-	0.03158	0.022073	1.430696	0.152532
waterfront-	0.24229	0.007021	34.50768	5.1E-254
grade-	0.451197	0.009003	50.11849	0
lat-?	0.173934	0.002455	70.85177	0

Table 5. The result of multiple linear regression model

Multiple R	0.799598
R Square	0.639357
Adjusted R Square	0.63924

As we can see in the result sheet, the adjusted R Square is 0.639 which is decent. It illustrates that the multiple regression model line fits well with the data. In terms of the factors, the p-value is very small except for sqft_lot which has an p-value of more than 0.1. Thus, we can come to the conclusion that they are all statistically significant in this model except sqft_lot. In addition, the sqft of living and grade has a relatively high coefficient while the others don't.

So, the final model turns out to be the following, equation (4):

$$house\ price = -0.2583 - 0.2272 \times bedroom + 0.944 \times sqft_{living} + 0.242 \times waterfront + 0.451 \times grade + 0.174 \times latitude + \varepsilon \quad (4)$$

4. Discussion and conclusions of the results

4.1 Analysis of the results and the potential reason

For single regression models, about the property of the house itself, the coefficient of the bedrooms, bathrooms, sqft of living, sqft of lot and sqft of above are high while the coefficient of the condition of the house, floor number and the sqft_basement are low. That is probably because consumers consider roomy and spacious space more valuable than the condition and floor numbers of the house.

In terms of the geological location and the surrounding environment, the zip code and the longitude show no relationship to the house price which is plausible. But the waterfront, view and along with latitude show some correlation with the house price. The potential reason for it is that people favor the house with a good view as well as a water view, so the house which has a waterfront or a good view will gain an extra price addition, this might also explain why the latitude is surprisingly related to the house price. According to the map of the king county, there are many rivers and lakes in the west, therefore from the west to the east, the house changes from waterfront house to non-water front house, and resulting in a difference in house price.

In terms of the other actors, the year in which the house was built or renewed shows no relationship. However, the grade of the house has a coefficient of 0.9 which means that people will reference the grade as an important factor when assessing a house.

For multiple linear regression model, the adjust R-square is 0.64 and the p-value is very small shows that the line fits well with the model. The coefficient of the factors is also worth studying, the square feet of living have a coefficient of 0.94 while the water front and grade have a relatively medium coefficient. Therefore, the square feet of the living place is the dominant factor in the multiple regression model, while the waterfront and grade of the house are also taken into account by

customers. But the others, however, have small or even negative coefficients. This is very surprising because in common sense, usually the greater number of bedrooms a house has, the more expensive the house price is. The potential explanation of this can be in multiple regression models, the independent variables mutually interact with each other and affect each other so the result will make a big difference.

4.2 Limitations and suggestions

Based on the former analysis, the model of this research can extrapolate the house price effectively. However, the model still has many limitations. First is that it can't predict the specific price of a certain type of house as the dataset has been processed in order to analyze the data more precisely causing the model of the research cannot fit the realistic situation very well. Second is that although the dataset contains a large number of house data over a long period of time, it is only constrained to one particular county in U.S, therefore, it can't show us the full picture of U.S or The World.

However, the model of this research can provide you with a reference that which factors should the estate company or the government consider when estimating a house's price. The resulting equation (5) also provides prediction models of how much will the factors impact prices. To provide further suggestions, the government, needs to pay attention to the living square feet of the house while building it, they also need to create a good environment of the surrounding including a good view or a waterfront. As for the real estate companies, they should adjust their marketing strategy when selling the house, they need to prominent a big living space of the house as well as to show the customers the view of the house in order to promote the overall value of the house, they also need to pay attention to the house's grade since the customers see it as an important factor when considering buying a house.

5. Conclusions

The research implemented a literature review that is relevant to the house price. Then it indicated the purpose of this research and described the data and method used in this research. The research is based on the data set of the house sales information in king county of U.S between May 2014 and May 2015, it used two kinds of model: The single regression model focus on the correlation based on the coefficient while the multiple regression model pays attention in p-value and R-square.

Firstly, the research chose to pre-process the data by their differences from the average since there are many extreme data in this dataset that will affect the research. Secondly, the research runs single regression model with individual factors one at a time. Thirdly, the research analyzes the endogeneity between the independent factors. Finally, based on the results, the research removed the factors that are not related to the price or highly related to other factors. And then the research run multiple regression model with all the factors that remained

The research came out with the following results:

1) In the single regression model, the number of bedrooms and bathrooms, the sqft of living, the sqft of the lot, sqft of above, sqft of the basement, the water front, the view of the house, the latitude and also the grade highly correlates with the house price.

2) In the multiple regression model, the number of bedrooms and bathrooms, the square feet of living, the water front, the grade and also the latitude are all significantly related to the house price. The square feet of living has a high coefficient illustrating that it may impact the house price more than the other factors. While the negative coefficient of bedroom and bathroom numbers is likely because of the interaction among all the factors

3) The government should consider square feet of living and the surrounding environment when building a house, the real estate company, however, should focus on mentioning the roomy space of the living and the beautiful view of the surrounding house as well as the good grade of the house specifically to their customers in order to make more money.

However, there are limitations to this research due to the small scale of the data and the potential interaction between each factor that is not considered in this research. Thus leaving more space worth studying in this field.

References

[1] Aini Zhong, Xing & Yu Jingmin. Study on housing prices and their influencing factors in Shenzhen from a geographical perspective—— Based on random forest model. Urban Survey,2022, (02): 66-70.

[2] Ouyang Wenhe & Zhang Xuan. A Comparative Study on the Development Path of Real Estate in China and the United States. Journal of Hebei University of Economics and Business,2022, (01), 66-71.

[3] Liu Lin & Zhang Hongrui. An Empirical Comparative Study on the Influencing Factors of Housing Prices in China and the United States. China Price, 2018, (12): 57-60.

[4] Li Jian & Deng Ying. A Study on the Monetary Factors Driving the Rise of Housing Prices: An Empirical Comparative Analysis Based on the Bubble Accumulation Period in the United States, Japan and China. Journal of Financial Research, 2011, (06): 18-32.

[5] Li Xiaoxi, Yang Lin. Virtual Economy, Bubble Economy and Real Economy. Finance and Trade Economics,2000, (06): 5-11.

[6] Guo Lingxia, Peng Qianying, Chen Lulu, Chen Ying & Yan Rui. A Literature Review of Research on Real Estate Bubbles at Home and Abroad. Journal of Aba Normal University,2021, (04): 89-9.

[7] Mao Yating. (2022).The Effect of Waterfront Vision on Residential Prices (Master's Thesis, East China Normal University).

[8] Shi Ruiping. (2009). Research Based on Univariate Regression Analysis Model (Master's Thesis, Hebei University of Science and Technology).

[9] Gao Xiaohong, Li Xingqi. Comparison of Dimensionless Methods in Multiple Linear Regression Models. Statistics & Decision, 2022, 38(06): 5-9.

[10] Ma Xiongwei. Multicollinearity diagnostic method and empirical analysis in linear regression equation. Journal of Huazhong Agricultural University (Social Science Edition), 2008, (02): 78-81+85.