Pisco Med Publishing

*Original Research Article*

# The Forecast of Housing Price in Xi'an Based on Big Data Analysis

**Zhiyuan Guo***

Department of Architecture and Civil Engineering, City University of Hong Kong, Kowloon, Hong Kong 999077, China. E-mail: guozhiyuan123456@hotmail.com

***Abstract:*** Based on the statistical data, the forecast for housing price of Xi'an city is made by identifying 13 different kinds of indicators. The multi-variable regression model and SPSS are used to analyze the data in linear and non-linear way respectively. R2 for both methods are over than 0.9. So we can get the similar conclusion that housing price in Xi'an will not increase dramatically recently and keep stable.

***Keywords:*** Housing Price; Forecast; Big Data

## 1. Introduction

Based on the crossover development of data collection, storage and analysis, big data technology is recognized by the academic community as an important factor of production that penetrates into various fields. With the rapid rise in the bandwidth of the mobile network, the availability of cloud computing and the Internet, more sensor equipment and mobile terminals are connected to the interconnection network, and the interconnection and intercommunication between the devices have been realized previously. The data generated from this cannot be estimated. In the field of real estate, the process of commercial housing sales also produces abundant data. How to effectively dig the hidden value behind the data and analyze and use these data to predict the trend of future housing price is a hot topic in the research of today's real estate industry.

On January 6, 2014, the state officially approved the establishment of Xi Xian new district in Shaanxi province. At this time, Xi Xian new district becomes a national new district and the seventh largest new district in China. In February 2018, Xi'an was identified as the national central city, and Xi'an was also the starting point of the silk road. Under the joint promotion of several national strategies, the rapid development of Xi'an also laid a solid foundation for its real estate market. Of course, when it develops rapidly, all kinds of problems can't be avoided. How to make the real estate market in Xi'an, which is the central city and the ancient capital of 13 dynasties, develop better and faster becomes an important problem to be solved. It plays an important role for Xi'an's economic development. Therefore, it is necessary to use big data to predict future housing prices in Xi'an[1].

## 2. Literature review

Joe Peek and James A. Wilcox[2] studied the influence of the first baby boomers on housing price in America. Through the simulation analysis, they thought the housing price would fall when the first American baby boomers became adults. Although the population was large, the actual income of these young people was relatively low. However, when the first baby boomers were getting old, due to the increment of their income


doi: 10.18282/i-s.v3i1.353

and the improvement of their purchasing power, the housing price would be enhanced and it would make the housing price higher than the original level.

Cost is the important factor of impacting the housing price, and the increment of cost must boost the housing price increasing. Shi Qun and Fu Jiangbo[3] worked out the trend of housing price in Nantong through the analysis of Nantong housing cost of each factor price trend. Wang Limei[4] analyzed the reason of increasing housing price in China from the angle of housing price composition. He thought the land price, artificial cost, the increment of price of building materials and China's imperfection of industrial tax system resulted in continuously high housing price in China. Sheng Guangheng and Li Xinyong[5] thought the land price and expenses of taxation accounting for 50% of housing price in China led to the high level of housing price.

# 3. Establishment of database

In order to establish our database to predict the housing price in Xi'an, we carefully looked up the Yearbook of Xi'an from 2000 to 2017[6]. There were many plenty of information and data in the Yearbook in the almost past 20 years, and we chose 14 indicators among them according to the theories and professors' suggestions. The indicators are selling price indices of residential buildings (1999 is 100), GDP (100 million), per capita GDP (Yuan), the value added in construction industry (100 million), total population (10000), population of urban area (10000), vacancy of residence (10000 $m^2$), floor space of buildings under construction (10000 $m^2$), floor space of buildings completed (10000 $m^2$), per capita annual disposable income of urban-absolute number (Yuan), average household size, average number of employed persons.

# 4. Model establishment

When designing a model to analyze and predict the price of residential buildings, variables should be selected carefully. First of all, dependent variables and independent variables must have the inner linkage, which means the dependent variables must be used to explain the performance of independent variables. Secondly, the number of dependent variables must be suitable for the problem conditions. The more variables we involved in the model, the more general solution we will get to describe the independent variables, but if too much variables are included to calculate the predicting equation, the variance of course will become larger and larger. The best way to deal with this issue is modifying the prediction model to ensure the important variables which with higher significance should be included and used to predict the result.

As we identified previously, 14 indicators may affect the residential pricing of Xi'an city (Shown in the **Appendix 1**). But not all these indicators are independent from each other. So the multi-variable regression model is considered to be used based on the previous government statistical big data.

## 4.1 Linear regression modeling

The step 1 is to definite the corresponding relationship among the variables. Here the simple linear relation is assumed between residential pricing and each indicator.

The step 2 is to analyze the data using "stepwise" method in SPSS. The linear relationship among these dependent variables may exist due to the economic relation and the limitation of source data. The data collected for this research is not quite adequate because of the limited data uploaded on the website, there is no doubt that some of the variables are not independent. So during model establishment, linear relationship among variables must be taken into consideration. The "stepwise" method here is used to avoid this kind of problem.

The variable which is the most significant is figured out, then other variables will be involved one by one, meanwhile, the test of significance needs to be done after involving one variable. When the significance is no longer increase any more as involving more variables, the "stepwise" stops automatically. The result of stepwise regression analysis is shown as below in **Table 1** (The process of stepwise method is shown in the **Appendix 2**).

**Table 1**. Model summary

| Model | R | $R^2$ | Adjusted $R^2$ | Standard error of estimate | D.W. |
|---|---|---|---|---|---|
| 1 | 0.978[a] | 0.956 | 0.953 | 5.97198 | |
| 2 | 0.989[b] | 0.977 | 0.974 | 4.44037 | 2.490 |
| a. Predictor | (constant) | l | | | |
| b. Predictor | (constant) | l,m | | | |

The table shows that 2 steps of regression have been made. In the first step, indicator l, which means per capita annual disposable income of urban, has been involved. The R is 0.978, R square is 0.956, and adjusted R square is 0.953. Secondly, indicator m—average household size, has been involved. The R value, R2 and adjusted R2 have increased. No more indicators are involved, that is to say these two indicators contribute most to the residential pricing.

**Table 2** shows the ANOVA result, the significance of F value is 0, much less than 0.05.

**Table 2.** ANOVA

**ANOVA**

| Model | | Sum of square | df | Mean square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 12390.175 | 1 | 12390.175 | 347.409 | .000[b] |
| | Residual | 570.633 | 16 | 35.665 | | |
| | Total | 12960.808 | 17 | | | |
| 2 | Regression | 12665.055 | 2 | 6332.527 | 321.173 | .000[c] |
| | Residual | 295.753 | 15 | 19.717 | | |
| | Total | 12960.808 | 17 | | | |

a. Dependent: a

b. Predicator: (constant), l

c. Predictor: (constant), l, m

Then we can analysis the coefficients of the variables and test the significance. From the **Table 3** we can know that the constant in this model is 220.406, coefficient of per capita annual disposable income of urban is 0.002, coefficient of average household size is -43.907, both of the significance of these two indicators are less than 0.05. So the statistical prediction equation is:

$$Y = 220.406 + 0.002 * l - 43.907 * m;$$

**Table 3.** Regression analyzing result and coefficients

| Coefficients | | | | | | |
|---|---|---|---|---|---|---|
| Model | | Unstandardized | | Standardized | t | Sig. |
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 91.001 | 2.838 | | 32.060 | .000 |
| | l | .002 | .000 | .978 | 18.639 | .000 |
| 2 | (Constant) | 220.406 | 34.722 | | 6.348 | .000 |
| | l | .002 | .000 | .996 | 25.337 | .000 |
| | m | -43.907 | 11.759 | -.147 | -3.734 | .002 |

a. Dependent variables: a (Residential pricing)

## 4.2 Model testing of linear regression model

As the P-P diagram shows in **Figure 1**, all points locate in both sides of the line, the distribution curve is a standard normal deviation in the regression standardized residual diagram.

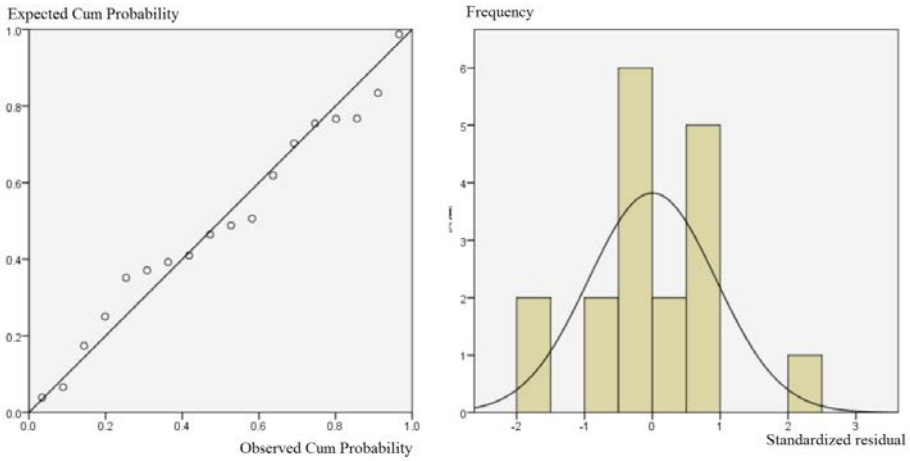


**Figure 1.** P-P Diagram and regression standardized residual.

## 4.3 Non-linear regression modeling

Similar as the previous analyzing procedures, a non-linear regression is made to modify the proper model. Assuming that the residential pricing and the indicators have an exponential relationship, so the equation can be written as:

$$Y = a_0 \cdot x_a^{a1} \cdot x_b^{a2} \cdot x_c^{a3} \cdots x_n^{an};$$

Where Y is the residential pricing for each year, $a_0$, $a_1$, $a_2$, …, $a_n$ are the coefficients of each indicator, $x_a$, $x_b$, $x_c$, …, $x_n$ are the independent variables which may affect the dependent variable Y. The exponential relationship of them can be written as:

$$\ln Y = \ln a_0 + a_1 \ln x_a + a_2 \ln x_b + a_3 \ln x_c + \cdots + a_n \ln x_n;$$

Taking logarithm for each variables and analysis through SPSS[7] as the procedures we did when analyzing the linear regression relationship. The result of stepwise regression analysis is shown as below in **Table 4**.

**Table 4.** Model summary for non-linear regression

| Model | R | $R^2$ | Adjusted $R^2$ | Standard error of estimate | D.W. |
|---|---|---|---|---|---|
| 1 | 0.991[a] | 0.983 | 0.982 | 0.028 | 1.542 |

a. Predictor (constant) l

The R square is a little bit higher than the R square we got through linear regression method, **Table 5** shows the ANOVA result, the significance of F value is 0, much less than 0.05.

**Table 5.** ANOVA

**ANOVA**

| Model | | Sum of square | df | Mean square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 0.701 | 1 | 0.701 | 917.100 | .000[b] |
| | Residual | 0.012 | 16 | 0.001 | | |
| | Total | 0.713 | 17 | | | |

a. Dependent: ln(a)

b. Predicator: (constant), ln(xd)

Then we can analyze the coefficients of the variables and test the significance. From the **Table 6** we can know that the constant for non-linear regression is 3.750, coefficient of "the value added in construction industry" is 0.208, the statistical non-linear regression prediction equation is:

$Y = 42.521 \times x_d^{0.208}$;

Where Y is the residential pricing, $x_d$ is the added value in construction industry of each year.

**Table 6.** Non-linear regression analyzing result and coefficients

**Coefficients**

| Model | | Unstandardized | | Standardized | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 3.750 | 0.039 | | 97.331 | .000 |
| | Ln($x_d$) | 0.208 | 0.007 | 0.991 | 30.284 | .000 |

a. Dependent variables: ln(a) (Residential pricing)

As the P-P diagram shows in **Figure 2**, all points locate in both sides of the line, the distribution curve is a standard normal deviation in the regression standardized residual diagram.

## 4.4 Model testing of non-linear regression model

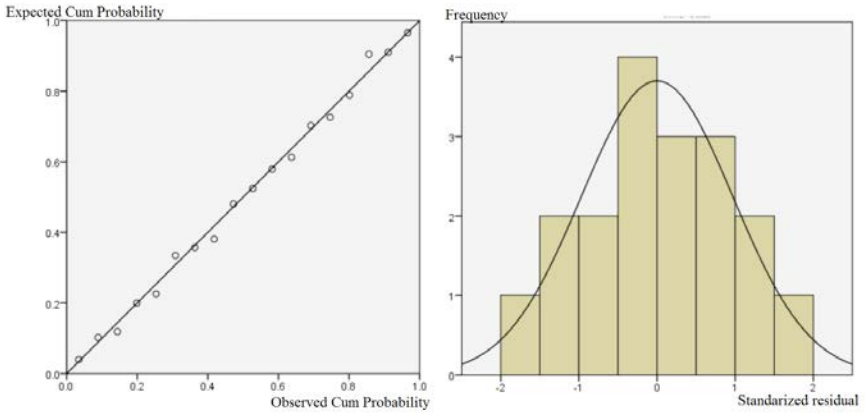


**Figure 2.** P-P Diagram and regression standardized residual of non-linear regression model.

The exponential relationship between residential pricing and value added in construction industry expresses a mathematical relation between these two variables. The pricing will increase as the value added exponentially. This equation shows the relation between government adjustment methods and pricing.

# 5. Conclusion analysis

## 5.1 The aspect of demand

For the linear regression model, we mainly focus on two factors which might influence the housing price in Xi'an. They are "per capita annual disposable income of urban-absolute number (Yuan)" and "average household size". Although the Xi'an government encourages talents to settle down and has introduced a series of policies on talent introduction recently, the talents introduced are almost young single people or young couple. It will lead to the little decrement of average population of each household. In other word, the demand of commercial house market in Xi'an will go down relatively. In the meantime, the development of economy in Xi'an has slowed down, so the per capita annual disposable income of urban will not increase significantly in the short term. The conclusion of it is that the housing price in Xi'an for the coming years will not increase dramatically.

## 5.2 The aspect of supply

For the non-linear regression model, the most important factor is "the value added in construction industry". The value added in construction industry is related to the government regulation, and its effect on housing price is exponential. In recent years, real estate industry has developed rapidly in Xi'an, and the developers are scrambling to build commercial house in Xi'an, so that it leads to the shortage of land resources. And with the reasonable government regulation, the development of commercial housing market in Xi'an could be slow down and the housing price will remain stable and not increase significantly.

# References

1. Ding J. The impact factor of commodity housing's price analysis in Xi'an and the price prediction [MSc thesis]. Xi'an: Xi'an University of Science and Technology; 2009. p. 8–18.
2. Peek J, Wilcox JA. The baby boom, "pent-up" demand, and future house prices. Journal of Housing Economics 1991; 1: 347–367.
3. Shi Q, Fu J. Analyze the current housing price trends in our city from the cost composition: A survey of housing prices in Nantong city (in Chinese). Marketing Week 2007; (3): 35–36, 47.
4. Wang L. From the composition of the real estate price to discuss the cause of excessively high housing price (in Chinese). Contemporary Manager 2006; (11): 228.
5. Sheng G, Li X. Commercial house: Looking at house prices from the cost structure (in Chinese). Economic Forum 2004; (5): 136–137.
6. Statistical Yearbook of Xi'an. Available from: http://navi.cnki.net/KNavi/YearbookDetail?pcode=CYFD&pykm=YXATJ&bh=.
7. Mi H, Zhang W. Practical modern statistical analysis methods and application of SPSS (in Chinese). Beijing: Contemporary China Publishing House; 2004. p. 121–150.

# Appendix 1

| Indicators | Descriptions |
|---|---|
| | (all data are cited in the web[6] from 2000 to 2017) |
| a | Selling price indices of residential buildings (1999 is 100) |
| b | GDP (100 million) |
| c | per capita GDP (Yuan) |
| d | The value added in construction industry (100 million) |
| e | total population (10000) |
| f | population of urban area (10000) |
| g | vacancy of residence (10000 $m^2$) |
| h | floor space of buildings under construction (10000 $m^2$) |
| i | residential buildings (10000 $m^2$) |
| j | floor space of buildings completed (10000 $m^2$) |
| k | residential buildings (10000 $m^2$) |
| l | per capita annual disposable income of urban-absolute number (Yuan) |
| m | average household size |
| n | average number of employed persons |

# Appendix 2

**Correlation**

| | | a | b | c | d | e | f | g | h | i | j | k | l | m | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Pearson correlation** | a | 1.000 | .965 | .966 | .962 | .968 | .941 | .593 | .955 | .974 | .848 | .854 | .978 | -.024 | .903 |
| | b | .965 | 1.000 | .998 | .998 | .945 | .951 | .757 | .996 | .989 | .927 | .919 | .987 | .156 | .817 |
| | c | .966 | .998 | 1.000 | .999 | .945 | .943 | .752 | .998 | .994 | .922 | .917 | .991 | .161 | .824 |
| | d | .962 | .998 | .999 | 1.000 | .937 | .937 | .754 | .998 | .993 | .924 | .918 | .989 | .171 | .812 |
| | e | .968 | .945 | .945 | .937 | 1.000 | .977 | .649 | .932 | .943 | .830 | .830 | .948 | -.054 | .939 |
| | f | .941 | .951 | .943 | .937 | .977 | 1.000 | .724 | .936 | .927 | .873 | .864 | .933 | -.021 | .873 |
| | g | .593 | .757 | .752 | .754 | .649 | .724 | 1.000 | .773 | .710 | .827 | .801 | .681 | .411 | .438 |
| | h | .955 | .996 | .998 | .998 | .932 | .936 | .773 | 1.000 | .993 | .931 | .924 | .987 | .180 | .802 |
| | i | .974 | .989 | .994 | .993 | .943 | .927 | .710 | .993 | 1.000 | .906 | .906 | .995 | .138 | .840 |
| | j | .848 | .927 | .922 | .924 | .830 | .873 | .827 | .931 | .906 | 1.000 | .997 | .883 | .328 | .679 |
| | k | .854 | .919 | .917 | .918 | .830 | .864 | .801 | .924 | .906 | .997 | 1.000 | .881 | .317 | .693 |
| | l | .978 | .987 | .991 | .989 | .948 | .933 | .681 | .987 | .995 | .883 | .881 | 1.000 | .124 | .852 |
| | m | -.024 | .156 | .161 | .171 | -.054 | -.021 | .411 | .180 | .138 | .328 | .317 | .124 | 1.000 | -.207 |
| | n | .903 | .817 | .824 | .812 | .939 | .873 | .438 | .802 | .840 | .679 | .693 | .852 | -.207 | 1.000 |
| **Significance (single tail)** | a | . | .000 | .000 | .000 | .000 | .000 | .005 | .000 | .000 | .000 | .000 | .000 | .463 | .000 |
| | b | .000 | . | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .269 | .000 |
| | c | .000 | .000 | . | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .262 | .000 |
| | d | .000 | .000 | .000 | . | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .249 | .000 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | e | .000 | .000 | .000 | .000 | . | .000 | .002 | .000 | .000 | .000 | .000 | .000 | .416 | .000 |
| | f | .000 | .000 | .000 | .000 | .000 | . | .000 | .000 | .000 | .000 | .000 | .000 | .468 | .000 |
| | g | .005 | .000 | .000 | .000 | .002 | .000 | . | .000 | .000 | .000 | .000 | .001 | .045 | .034 |
| | h | .000 | .000 | .000 | .000 | .000 | .000 | .000 | . | .000 | .000 | .000 | .000 | .238 | .000 |
| | i | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | . | .000 | .000 | .000 | .293 | .000 |
| | j | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | . | .000 | .000 | .092 | .001 |
| | k | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | . | .000 | .100 | .001 |
| | l | .000 | .000 | .000 | .000 | .000 | .000 | .001 | .000 | .000 | .000 | .000 | . | .312 | .000 |
| | m | .463 | .269 | .262 | .249 | .416 | .468 | .045 | .238 | .293 | .092 | .100 | .312 | . | .204 |
| | n | .000 | .000 | .000 | .000 | .000 | .000 | .034 | .000 | .000 | .001 | .001 | .000 | .204 | . |
| **Number of cases** | a | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | b | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | c | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | d | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | e | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | f | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | g | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | h | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | i | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | j | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | k | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | l | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | m | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |
| | n | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 |