Real Estate Prediction of House Price in Text of Multivariate Linear Regression

Yongxuan Zhang[1, *]

[1]Mathematics Department, College of Sciences, North Carolina State University, Raleigh, United States

*Corresponding author: Yzhan249@ncsu.edu

Abstract. During the recent decades, the estate market of China reflects dramatic increasing speed which shocks the whole world. Among various indicators, house price is one of the important indexes reflecting the fast raising price. In this study, the elements which connect to the house price are investigated. To be specific, this paper picks up the typical indexes to construct the model through the Multifactorial linear regression approach. Based on the method, this research discoveries the relationship between the house price with the other related macro-economic index new housing index, citizen wages, resident population, and currency of country. According to the analysis, there is slightly connection between them, and this connection is varying between different cities. While providing the reference to the following research, it also presents that the different development level for Chinese city have the different route to development. Thus, when enterprise or government make the decision or policy, it is necessary to consider its own conditions, not just copy other's experience. These results shed light on guiding further exploration of the city development.

Keywords: Estate prediction; Python; OLS; House Price.

1. Introduction

The city development is so faster in china for the recent decades. Most of the cities changed quite a lot compare to the 20 years ago. Urbanization is the only way which must be passed when a country expects to develop itself further in the modern times [1]. The house price as one of most important signals, bring influence to all other city policy and planning. Hence, until now, many cities and countries had done many researches to expect found out the critical elements for it. Singapore had carried out this kind of research on the new housing price index and house price [2]. Besides the Asia, the Unite State also evaluate this topic and find it's related to the expected inflation rates [3].

For an estate development, the sales forecast is one of most critical topics through the whole development lifetime. From the bidding lands completed, the sales prediction is started, in different styles, some in markets. Some in the relationship with customer. All the estate company combined to the estate industry, and for better this research have the indexes and prediction as reference to make the budget and plan for next incoming invest and business activities. Therefore, this house price is meaningful for the whole estate industry. This is why one wants to make the prediction for it. In addition, in the history, method has been developed step by step from basic linear prediction to the higher-level machine learning [4, 5].

In addition to the OLS method, some researcher had tried many other methods to build up different models to forecast the estate related price or indexes. For example, the ARIMA model is a generalization of an autoregressive moving average model. Researcher had used ARIMA method to make the self-regression on the new housing price index to make the short-term prediction and Xi'an city [6]. Its full name is cellular neural networks model. It is a parallel computing paradigm which like the neural networks, with the difference that communication is allowed between neighboring units only [7]. There was trail done to use the CNN model to prediction, and paper author found that compared with other traditional methods, the CNN work can get a better result by experiments put the actual data of property transaction into using [8]. The lightgbm model is a gradient boosting framework that uses tree-based learning approach method. It is recently designed to be distributed and efficient with the many advantages for used as a machine learning algorithm [9]. The previous research evaluated this method via using it predict housing price, and make the conclusion that this

method is a hybrid regression method. It is simple but performs much better than several other old comparable methods ascribed to using the generalization to enhance the performance [10]. For the OLS approach, the full name is ordinary least squares and is a standard approach via the regression analysis to approximate the solution of overdetermined systems by minimizing the sum of the squares of the residuals which made from each individual equation results. This method is simple and easy to approach, it balances the time cost and efficiency which fits the normal usage level. It has been adopted for the prediction in this paper and meanwhile taking further treatment to make the results to reach higher accuracy.

This article is in order to estimate the clues that to predict house price for the future based on the exist macro-economic data. The motivation is to provide the solid evidence and good prediction for commercial usage. With more accurate prediction, the capital of estate company can calculate and balance the cash flow and make the fully use of the free cash flow. The rest part of the paper is organized as follows. The Sec. 2 will introduce the methodology this article implements to research for the factors. The Sec. 3 will present the results of the analysis and discussion. Eventually, a brief summary will be given in Sec. 4.

2. Methodology

This article mainly uses the multivariate linear regression method, which is the most common method to treat the multiple factors to the targets. It can be process fasters with less computation resource.

2.1 Data

The data is the most critical input for the whole model construction. Due the recently days, the protection for the data is more and more strictly. This article gets the data from the download at the public information data. For this topic, the typical variables selected for major macro-economic data are listed as follows:

- House price. Since in the past ten years, the house price has great changes compare to all other indicators. Meanwhile it combined with the new housing price index in major cities. This index reflexes the new housing price condition in index format. In addition, it supposed to be the house price change rate for the cities. With this in mind, this paper uses it to calculate the house price data for each month.
- Average Salary/Wages in major cities. This is a easy understandable variables, which reflect the buying power of the citizen.
- Residential population in major cities, which means that the one lives city for long-time, and gets the local city's residential identity.
- M2 data for major cities, is a measure of the money supply of individual city government, = including cash, checking deposits, and the money which is easily-convertible.

Since there's multiple cities in different raw data when get them from database, it needs additional treatment to make them able to be used. Thereofore, the common cities are selected which existed in all the datasets, thus there won't be blank error for the regression.

The data is composited by several major cities, and the data panel need for analysis is from the January of 2010 to the December of the 2020, while some data only have the annual values. In the acquired data, all twelve-month filled with same value. Therefore, the treatment is using the linear interpolation method to get the different for each month, assuming every month in the year are changes with even ratio Original annual value assign to the December of each year. Then, one needs to make the linear interpolation for range from the December value in previous year to current year's December.

2.2 Algorithm and Theory

For Multiple Linear Regression, it's a statistical analysis method that uses regression analysis in mathematical statistics to determine the interdependent quantitative relationship between two variables or more ones [11]. To check the regression model, this study takes the Kolmogorov–Smirnov to check if the residual is normal distributed or not. For the definition in statistics field, the Kolmogorov-Smirnov test (also called as K-S test or KS TEST) is a continuous (or discontinuous) non-parametric test for the equality of one-dimensional probability distribution, which usually used to compare one sample with a reference probability distribution (single sample K-S test). It is named after its creators Andrey Kolmogorov and Nikolai Smirnov [12]. Kolmogorov Smirnov statistics quantitatively describes the distance between the empirical distribution function of a sample distribution and the cumulative distribution function of a reference distribution. The null distribution of the zero distribution of the statistic is calculated based on the null hypothesis. This methodology had been integrated in python scipy package, could be used as function.

For the Box Cox transformation, it is a generalized power transformation method proposed by box and Cox in 1964. It is a data transformation commonly used in statistical modeling, which is used for continuous response variables do not meet the normal distribution. With the box Cox transform, the correlation between unobservable error and prediction variables can be reduced to a certain level. The key features of box Cox transform will introduce a parameter and estimate it by the data itself, and then determine the form of data transform that should be adopted. Box Cox transform can significantly improve the normality, symmetry and variance equality of data, and is effective for many actual data. Since for some data, there's no efficient data to be normal distributed. Thus, here used Box-cos transformation to reform the properties of data in somehow.

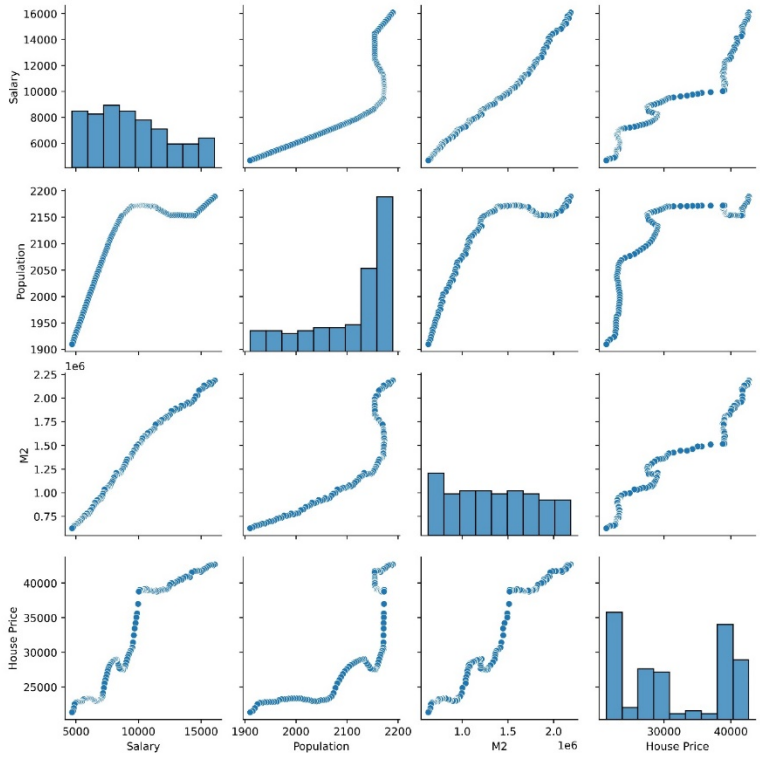


Fig. 1 Beijing Correlation analysis

3. Result & Discussion

3.1 Correlation analysis

Correlation analysis is applied to dig the relationship between different variables. For the first level cities, Beijing, Shanghai, Guangzhou, Shenzhen, the related X parameters with Y house price pair

plot as below. From the sub scatter chart and histogram, one can imply the correlation between each two variables. The results are given in Figs. 1-4.

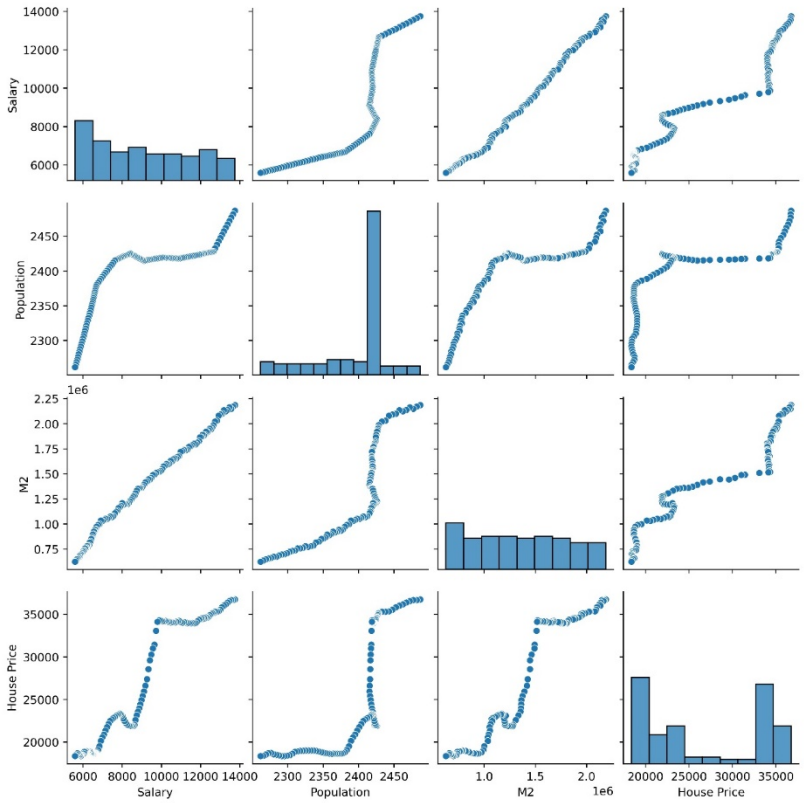


Fig. 2 Shanghai Correlation analysis

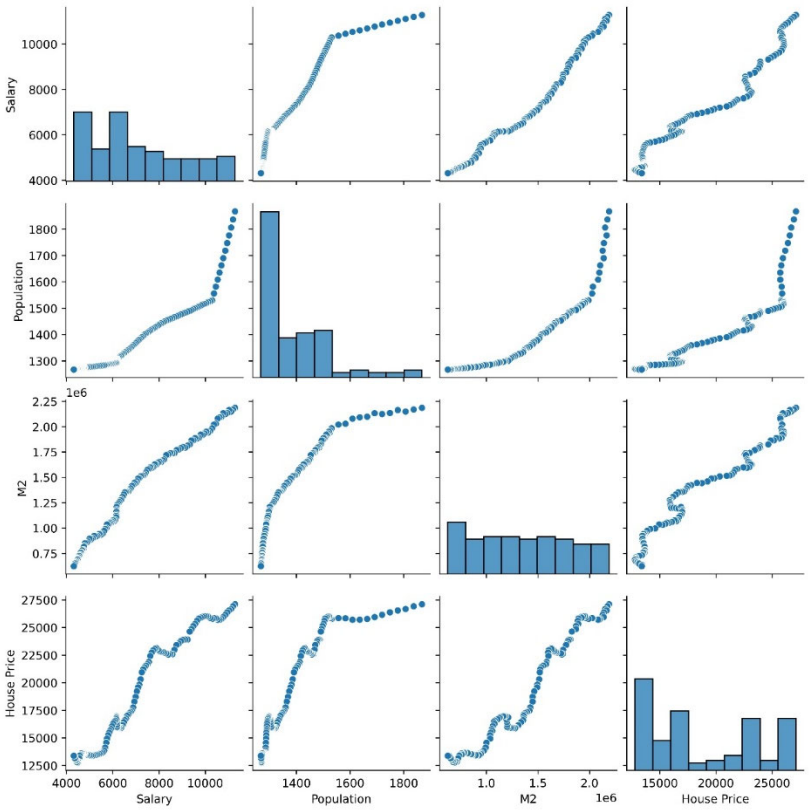


Fig. 3 Guangzhou Correlation analysis

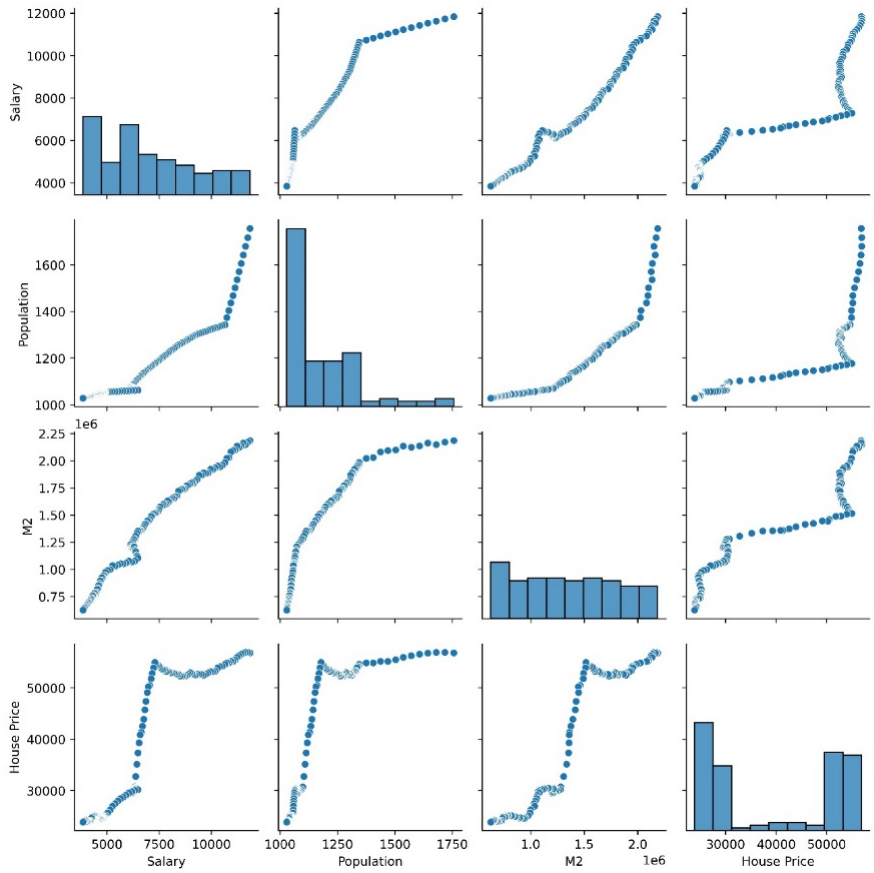


Fig. 4 Shenzhen Correlation analysis

3.2 Comparison of different models

For different model with different theory will cause different results. After calculation, one may find out the regulation which hidden under the macro-economic variables. For major first level cities, the regression formulae are given in Table. 1. In this paper, the differences of coefficients and metrics of different models are analyzed. Compare to the normal raw data and normalization data, and especially the variable new housing building index, the original one is less fitting. There's different in the final fitting values. with accumulated new housing building index, the mean value of all cities increases 0.02 approximately.

Table. 1 The regression results

Models, y: House Price, x1: Salary, x2: Population, x3: M2	R^2	MSE	MAE
Beijing: $y = 16.3450 + 5.1361 \times 10^{-150} \cdot x_1 + 1.7385 \times 10^{-75} \cdot x_2 + 2.6232 \times 10^{-147} \cdot x_3$	0.66	0.082	0.22
Shanghai: $y = 6.3971 + 3.1138 \times 10^{-157} \cdot x_1 + 3.2690 \times 10^{-79} \cdot x_2 + 6.0174 \times 10^{-154} \cdot x_3$	0.64	0.0041	0.057
Guangzhou: $y = 1.9766 + 7.0257 \times 10^{-2} \cdot x_1 + 0 \cdot x_2 + 3.0147 \times 10^{-6} \cdot x_3$	0.99x	4.19E-07	0.00055
Shenzhen: $y = 65.2588 - 8.3907 \cdot x_1 - 6.3537 \times 10^{-12} \cdot x_2 + 5.8699 \times 10^{-3} \cdot x_3$	0.98	0.22	0.39

3.3 Explantion

The reason for this is because the model structural is based on the multiplies factors. More similar trend at these individual factors, the result will fits bestter. The correlationship reflect the way two variables changes with each other. As calculated the selected variables have high value more than 0.9 with other variables, it means they reflect same trend, that could predict the house price better.

3.4 Limitation

After all, this article done the most works depend on the available data and logic, and it still has some limitations on this research.

- Due to the limit time, this study not processes all the cities in China for full samples. In this case, it will lead the specific sample to be ignored, then the result will not cover all the cities in china.
- The tools only used python, not implemented any other software, the algorithm depend on the modules built inside, not validated from code level.
- For some city, the fitting R^2 not reach the expect level. Some policy reason needs to be considered in futures. For example, the big cities have strictly buying permit, this affect the prices intention changes, and supposed to be investigate separately as another topic.
- In addition, on the other way, from the total amount, big cities forecast is more important than other cities, as the trade volume is quite big. Current simulation didn't add the weight at this aspect.

4. Summary

In conclusion, this research proposes a regression model for the new housing price index model with the selected variables in specific aspects in terms of the critical data of major cities in China. According to the analysis, the new housing price index has some object relationship with the local salaries, population, and macro-economic M2. Besides, this relationship is varied from different cities. This present that these exists huge unbalance between different level cities in China. Thus, the factor's influence limits between first level cities and other level cities are quite different. Nevertheless, generally, the basic rules are more population and M2 will bring the raise trend for the new housing price index, that bring the hint for estate industry that most estate invest prefer to invest on the raising cities. On the other hand, different company culture will have different interpret for this result, some will continue focus on the big city which have larger market which support by the great population. Some company may just position to the smaller one, but have big potential to attract more peoples. Anyway, this article didn't analysis data in many algorithms, and some details are ignored as just use the available variables. In the future, some mass data which from the real time feed could be considered, then the prediction could be improved. Overall, these results offer a guideline for key economic index prediction, take this research as a foundation, then more works and model take these results as benchmark for further prediction on the all-other important index.

References

[1] Wang F. The Correlation Study of Urbanization and Real Estate Price. Journal of Henan University of Science & Technology (Social Science), 2010.

[2] Jiang L, Phillips P, Yu J. New methodology for constructing real estate price indices applied to the Singapore residential market. Journal of Banking & Finance, 2015, 61(DEC.SUPPL.2): S121-S131.

[3] Jonathan H. Mark, Michael A. Goldberg. Alternative housing price indices: an evaluation. Real Estate Economics 12.1 (1984): 30-49.

[4] Bishnu S Atal. The history of linear prediction. IEEE Signal Processing Magazine 23.2 (2006): 154-161.

[5] Madhuri C H et al. House price prediction using regression techniques: a comparative study. 2019 International conference on smart structures and systems (ICSSS). IEEE, 2019.

[6] Yuan Fang. Analysis and prediction of price index of newly built houses in Xi'an Based on ARIMA model. China Science and technology information, 2013 (11): 2

[7] Leon O. Chua, and Lin Yang. Cellular neural networks: Theory. IEEE Transactions on circuits and systems 35.10 (1988): 1257-1272.

[8] Yong Piao, Ansheng Chen, and Zhendong Shang. Housing price prediction based on cnn. 2019 9th international conference on information science and technology (ICIST). IEEE, 2019.

[9] Ke Guolin, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30 (2017).

[10] Truong Quang, et al. Housing price prediction via improved machine learning techniques. Procedia Computer Science 174 (2020): 433-442.

[11] Uyan K G K, N Güler. A Study on Multiple Linear Regression Analysis. Procedia-Social and Behavioral Sciences, 2013, 106:234-240.

[12] Hubert W Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American statistical Association 62.318 (1967): 399-402.