Proceedings of the 2nd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/15/20230943

The Second-Hand House Price Prediction Using Multiple Linear Regression Model

Yuhan Jin[1,2,a,*]

[1] *Binghamton University, Vestal NY13902, USA*
[2]*Henan, Zhengzhou, Tongtai road, 450000, China*
a. yuhanjin1202@163.com
**corresponding author*

Abstract: Analyze the influencing factors of second-hand housing and build the relevant model using the data on second-hand housing prices in 2017-2018. The R-code analysis is used to construct a prediction model of house prices, and the main factors affecting their changes are obtained. According to the significance test, the model meets the expectation and is feasible. Finally, it is concluded that the most noticeable impact on housing prices is room distribution and ladder ratio, and the least obvious is trade time.

Keywords: second-hand house price forecast, multiple linear regression model, regression analysis

1. Introduction

With the development of the economy, living standards in Beijing and other big cities have improved rapidly, and people's living costs have increased significantly. Whether to buy a house has become the most concerning issue. With the commercialization of housing and the gradual opening of the registered residence, buying a home has become an important symbol of integration into the city. The house represents a person's financial ability and social status in today's society. In recent years, as people under the influence of the epidemic face the dilemma of money shortage and employment difficulties, it is difficult for people to buy houses. In today's social context, real estate is depressed due to various factors, such as the economy, so second-hand housing has gradually become a topic of increasing concern. Second-hand houses can be occupied more quickly without spending much money on fine decoration and furniture purchases. Therefore, it is necessary to predict and effectively analyze the factors that affect the second-hand housing price for a macro analysis of the current second-hand housing market. In the early stage of a second-hand house purchase, it is essential to predict the accurate house price. For purchasers, it is necessary to budget the cost in advance, saving much time to see the house in person. For housing agents, the estimated price is conducive to their later sales.

According to market research, a city's development potential and location are the added value of house prices. The central city and the core area represent the scarcity of houses and the subsequent rising space. In the exact location, the house price is also affected by the unique factors of the house itself. For example, house type, area, age, facilities, etc. The house price is determined by the land value and its added value. Multiple linear regression is the best choice to solve such problems more simply, effectively, and efficiently.

DOI: 10.54254/2754-1169/15/20230943

Therefore, this study takes the data on second-hand housing prices and their influencing factors in Beijing from 2002 to 2018 as an example and uses multiple linear regression to establish a prediction model of second-hand housing prices and analyze the main factors affecting the prices of second-hand housing.

This paper is divided into five parts. The second part of the paper will review the classic literature, such as previous studies on housing prices. The third part of the paper will focus on the multiple linear regression model. The fourth part of the paper will discuss the advantages and disadvantages of the model or its application. Finally, the fifth part of the paper will be summarized to discuss the limitations of this study and the direction of future research.

2. Related Works

Wei Fang, Heng Xiao, and Xiangchen Ren predicted and analyzed the logistics demand based on the linear regression model by analyzing the data set of China from 1993 to 2007 and analyzing the relationship between GDP and the total cost of social logistics. The results show that, from the perspective of the economic phenomena simulated by the model, with the growth of time, the contribution rate of logistics demand to GDP is becoming larger and larger, so the supporting effect of the logistics industry on the national economy is becoming more and more apparent [1].

D. Nguyen, N. A. Smith, and C. P. Rose' analyzed blog, telephone conversation, and 2011 breast cancer online forum, and using multiple linear regression to predict the author's age from the text, the experiment used blog and telephone conversation data to develop the model and used forum data to test and finally concluded that content characteristics and style characteristics are favorable indicators of a person's age [2].

Yong Zhou, Yu HuanXiao, and Runsheng Huang used multiple linear regression analysis models to predict the grain yield in Guangxi. He analyzed the relationship between the planting area (x_1), rainfall (x_2), unit area yield (x_3), and grain yield (y). The results showed that the most significant factor affecting grain yield was planting area, followed by unit area yield, and precipitation had a minor effect on grain yield. The regression equation between grain yield and planting area, grain yield per unit area, and rainfall is $y = -517.759 + 0.158x_1 + 0.382x_2 - 0.001x_3$ [3].

Yu Kang predicted the gas drainage volume by analyzing the drainage data of Jiulishan Mine from 2006 to 2010 and using multiple linear regression. Conclusion through the correlation analysis between the influencing factors and the annual drainage volume, it is concluded that the main factors affecting the gas drainage volume of the mine are the total length of adequate coal holes, the length of drilling holes per ton of coal, and the percentage of 100mm diameter drainage pipeline in the entire length of drainage pipeline. Through regression test, it is confirmed that the model established by multiple linear regression can be used to predict the gas drainage volume of Jiulishan Mine [4].

E. Cakra and B. Distiawan Trisedya used the linear regression model to forecast the stock price based on sentiment analysis. The stock price depends on the demand for stocks, but there are no specific variables to help predict demand. The stock price depends to some extent on new information, that is, how people view social networks every time. Support vector machine (SVM), naive Bayes, decision tree, random forest, and neural network were used to analyze Indonesian stock prices in 2015. Finally, the relationship between sentiment analysis and stock price is obtained, and the error with the actual stock price is obtained [5].

Eric R. Edelman and Sander M. J. van Kuijk predicted the total operation time by analyzing the whole operation time of six academic hospitals in the Netherlands from 2012 to 2016 and using the multiple linear regression model and established the regression model through the relationship between eSCT, operation type, ASA classification, anesthesia type and total operation time. Finally, the linear regression model for predicting TPT based on eSCT, operation type, patient ASA classification,

DOI: 10.54254/2754-1169/15/20230943

and anesthesia type is superior to the current practice of using ACT standard duration or fixed ratio between eSCT and TPT [6].

Paikun, T. Kadri, and R. D. Hudayani Sugara applied a multiple linear regression model to estimate budget construction housing larger than 29 square feet and smaller than 290 square feet in Indonesia and used neural networks and regression analysis to build a model to predict the error value between the cost of a house with a floor area of 290 square feet and the actual price, deriving linear regression models for easy and quick solutions to housing budget problems [7].

S. Acharya, A. Armaan, and A. S. Antony predicted graduate admissions by analyzing the UCLA graduate data set and compared four regression models: multiple linear regression, support vector regression, decision tree, and random forest to achieve the optimal model. The final results show that higher test scores, GPA, and other factors often lead to greater access. And the multiple linear regression model is the model with the slightest error [8].

Xinxin Hu predicted the GDP of Yunnan Province by analyzing the GDP and related data of Yunnan Province from 2007 to 2016 and using multiple linear regression and established a model to explore the relationship between the GDP of Yunnan Province and associated variables. The model can be used to forecast the GDP of Yunnan Province through autocorrelation tests and other tests. Conclusion through the correlation analysis between the influencing factors and the GDP of Yunnan Province, it is concluded that the main factors affecting the GDP of Yunnan are industrial added value, agricultural, forestry, and animal husbandry added value, and construction added value [9].

Ran Zhao analyzed Boston's housing prices in the 1970s based on the regression method and tested the significance of the regression model and coefficient. Conclusion Through the correlation analysis between the influencing factors and housing prices, it is concluded that the main factors affecting housing prices in Boston are the average number of rooms per house, the proportion of urban students and teachers, and the percentage of population decline [10].

3. Method

3.1. Linear Regression

In statistics, linear regression is a regression analysis that models the relationship between one or more independent variables and dependent variables. Generally, in multiple linear regression, there are numerous factors affecting y; assuming that there are $x_1, x_2, \ldots x_n$ factors, the following linear relationship can be considered:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1}$$

Because multiple linear regression is fitted by least square approximation, which is widely known. If you want to know more details, refer to [11].

3.2. MSE

Mean square error (MSE) is the most commonly used regression loss function. The calculation method finds the fair sum of the distance between the predicted and actual values. The formula is shown below.

$$MSE = \frac{\sum_{i=1}^{n}(f(x)-y)^2}{n} \tag{2}$$

DOI: 10.54254/2754-1169/15/20230943

The range of MSE is $[0, \infty)$. When the predicted value is the same as the actual value, MSE equals 0. The greater the error, the greater the MSE value. Therefore, the lower the MSE value, the higher the model's accuracy.

3.3. MAE

Mean absolute error (MAE) is another regression loss function. MAE is the mean value of the sum of the absolute value of the difference between the predicted value f (x) and the target value y. The formula is shown below.

$$MAE = \frac{\sum_{i=1}^{n}|f(x)-y)|^2}{n} \tag{3}$$

MAE is more stable and insensitive to outliers as a loss function, but its derivative is discontinuous and has low solution efficiency.

3.4. R Square

R square is a statistical measure of the proportion of variance of dependent variables affected by independent variables in the regression model.

$$R^2 = 1 - \frac{\sum_i(y_i-\hat{y}_i)^2}{\sum_i(y_i-\bar{y}_i)^2} = 1 - \frac{Total\ Variation}{Unexplained\ Variation} \tag{4}$$

If the fraction part is 0, it means if and only if there is no deviation.

If the fraction part is 1, that means if and only if the deviation amplitude is the same as the variance, your predicted value and average value have the same effect.

If the fraction part is more significant than 1, the prediction effect cannot even reach the average level.

3.5. F Statistic

F statistics determines whether there is a significant difference between the two variables during regression analysis. The larger the F statistic, the more influential the difference between sample averages relative to the difference in the sample. Therefore, the more significant the F statistic, the greater the difference between the averages.

4. Result

Analysis of the influencing factors of second-hand housing prices in Beijing

4.1. Characteristics of Houses in Beijing

Beijing is the capital of China and the center of China's economic development. Excellent geographical location is an inevitable condition for high housing prices. In addition, people living in Beijing can have more abundant resources than other provinces, access to more excellent units, and so on. The living environment also brings people different wealth.

4.2. Dataset

Data can be viewed in [12]. This data includes information on second-hand housing prices and related influencing factors in Beijing from 2011 to 2017. The data consists of more than 290000 houses and 26 elements affecting house prices. It includes the longitude and latitude of the house's geographical

DOI: 10.54254/2754-1169/15/20230943

location, unit price, total price, total area, number of rooms, construction time, decoration degree, construction time, house type, floor, ladder ratio, whether there is an elevator, whether there is a subway, community average, etc. Due to the extensive data, to understand the trend of second-hand housing prices in Beijing in recent years and the main influencing factors, the data from 2017 to 2018 were extracted. To ensure the visualization and value of the data, the data were preprocessed as follows before the experiment. First, because the data on decoration degree and house type is vague, the specific situation cannot be determined, so it has been deleted from the data. Second, the unit price, total price, and area are in positive proportion, so the unit price data is removed. Third, the number of kitchens is single, and there is only exist 0 or 1, which cannot be analyzed from a linear perspective. Therefore, this parameter is removed. In addition, to ensure that the data value is within the specified range, all parameters have been standardized. After preprocessing, the new dataset has 42617 entries and 13 parameters. It includes price, longitude and latitude, area, number of rooms, construction time, trade time, renovation conditions, ladder ratio, subway, community average, etc.

4.3. Related Work

The R-code software analyzes the data, establishes a linear regression model, and draws a scatter plot. Since there is too much data, 400 samples are taken according to the principle of randomness to display the relationship between respective variables and dependent variables in the scatter plot as much as possible, then ensure the probability of each parameter in the data being selected is the same.

4.4. Prediction of the Results of Second-hand Housing Prices in Beijing

Linear regression model

Use linear regression to establish a model, and assume that there is a linear relationship between all variables, establish a multiple linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{12} x_{12} \quad (5)$$

x_1 represents longitude, x_2 represents latitude, x_3 means trade time, x_4 represents a square, x_5 represents living rooms, x_6 represents drawing rooms, x_7 represents community average, x_8 represents bathrooms, x_9 represents construction time, x_{10} represents renovation conditions, x_{11} represents ladder ratio, x_{12} represents subway. y indicates second-hand house price， β_0 is a constant，β_i is a parameter.

Statistical analysis is realized by r-code. The Independent variable selects second-hand house price(y), dependent variable selects upper segment $x_1 \ldots x_{12}$.

Evaluation

Table 1: Evaluation metric for the following content.

Evaluation Metric	Value
Mean Square Error (MSE)	70781213
Mean Absolute Error (MAE)	5961.18
F-statistic	2.604e+04 on 12 and 42604 DF
R-square	0.88

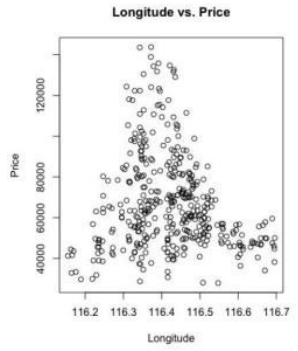
DOI: 10.54254/2754-1169/15/20230943

Table 2: The coefficient for different parameters.

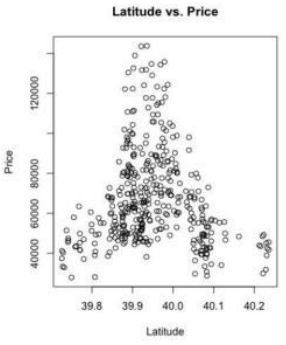
Model	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.150e+07	1.146e+06	10.035	*** p<0.01
Longitude	2.354e+02	3.588e+02	0.656	p>0.1
Latitude	-2.419e+03	4.148e+02	-5.831	*** p<0.01
Trade Time	-5.648e+03	5.675e+02	-9.953	*** p<0.01
Square	-8.199e+01	2.138e+00	-38.352	*** p<0.01
Livingroom	-5.911e+02	8.073e+01	-7.322	*** p<0.01
Drawing Room	6.954e+02	1.086e+02	6.406	*** p<0.01
Community Average	1.034e+00	2.112e-03	489.574	*** p<0.01
Bathroom	2.322e+03	1.396e+02	16.635	*** p<0.01
Construction Time	-1.389e+01	5.218e+00	-2.663	*** p<0.01
Renovation Condition	3.974e+02	4.271e+01	9.306	*** p<0.01
Ladder Ratio	1.721e+03	2.652e+02	6.492	*** p<0.01
Subway	4.365e+02	8.924e+01	4.891	*** p<0.01

Plot

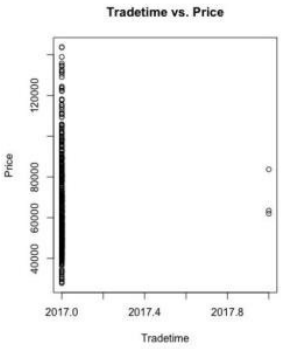
To clearly show the relationship between respective and dependent variables and test the model's accuracy, scatter plots of each independent variable and dependent variable y on the following figures.



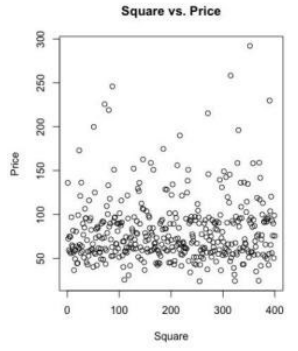
(a)



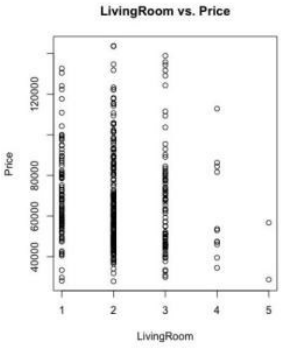
(b)



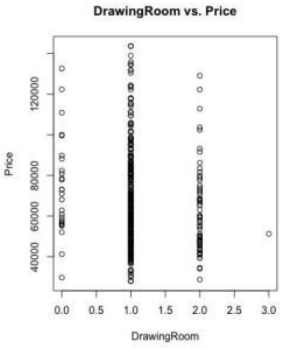
(c)



(d)



(e)



(f)

DOI: 10.54254/2754-1169/15/20230943

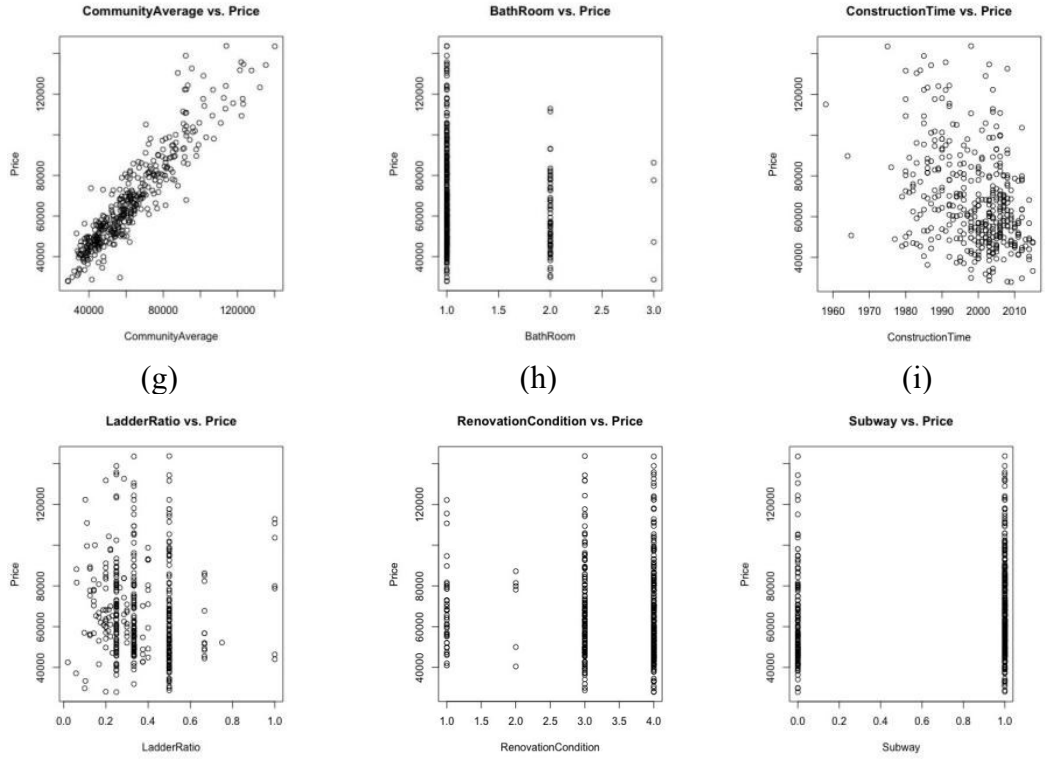


Figure 1: Group of scatter plot.

(a) Scatter plot of Second-hand House Price and Longitude
(b) Scatter plot of Second-hand House Price and Latitude
(c) Scatter plot of Second-hand House Price and Trade Time
(d) Scatter plot of Second-hand House Prices and Square
(e) Scatter plot of Second-hand House Price and Living Room
(f) Scatter plot of Second-hand House Price and Drawing Room
(g) Scatter plot of Second-hand House Prices and Community Average
(h) Scatter plot of Second-hand House Price and Bathroom
(i) Scatter plot of Second-hand House Price and Construction Time
(j) Scatter plot of Second-hand House Price and Ladder Ratio
(k) Scatter plot of Second-hand House Price and Renovation Conditions
(l) Scatter plot of Second-hand House Prices and Subway

5. Discussion

Through the analysis of the data on second-hand housing prices and their influencing factors in Beijing from 2017 to 2018, a scatter plot is drawn to analyze the relationship between variables further. A Scatter plot is an intuitive method to reflect the relationship between variables. It can usually directly reflect whether there is a linear correlation between two variables, favorable linear distribution or negative linear distribution [13].

The scatter plot shows that the relationship between price and community average is positively linear. The more expensive the average cost of the community, the higher the house price. Analyzing the relationship between construction time and price from the overall scattered plot distribution, the

DOI: 10.54254/2754-1169/15/20230943

later the construction time is, the younger the age of the house is. The figure shows that the house price with the later construction time is increasing, and the place with a younger generation is more popular. Several points on the right side of the scatter plot are separated from the scattered group. This is a normal phenomenon because the house price is also affected by the market economy, and there will be a specific range of fluctuations. More specifically, the scatter of construction time under the same price is horizontal distribution, and the price under the same construction time is vertical. This indicates that building age's impact on house prices is a weak negative correlation. That is, the construction time does not determine the cost of the house but also depends on other factors, such as the location and area of the house.

Next, we will analyze the relationship between geographical location and housing prices. According to common sense, the house price in the center is higher because the site brings people a faster and more convenient life. It can be seen from the scatter plot that the points are relatively dense in the middle part. Combining longitude and latitude, we can find that the houses with the highest price are in the downtown area. It is evident that the closer the house's location is to the city center, the higher the house price will be. House prices are not only affected by longitude and latitude, so there is no apparent linear relationship between them.

The point distribution of the room area is very messy, and it isn't easy to see the relationship between the site and the house price because the 50 square meters house in the city center and the 150 square meters house in the suburb may have the same price.

Through the analysis of second-hand housing data in Beijing, the prediction model of price and related influencing factors is obtained, and the regression model is obtained:

$$y = 1.152 \times 10^7 - 2412x_2 - 5648x_3 - 81.93x_4 - 595.4x_5 + 696.9x_6 + 1.034x_7 + 2352x_8 - 13.99x_9 + 398.9x_{10} + 1723x_{11} + 440.4x_{12} \quad (6)$$

It can be seen from the model that the most apparent factors affecting second-hand housing prices are the number of rooms, the ladder ratio, and elevators. The trade time has the most negligible impact.

6. Conclusion

Based on the multiple linear regression model, we explained the problem of second-hand housing price forecasts in Beijing. Because houses represent a person's economic-financial ability and social status nowadays, house price prediction has tremendous significance for budgeting the house purchase in advance and subsequent investment. This paper first focuses on the necessity and importance of predicting housing prices. Then review the relevant classical literature and analyze the data before the study. Then the significance analysis and model construction are carried out. Finally, discuss the final results of the model and analyze relevant factors. In the results obtained at this stage, this paper focuses on establishing the secondary housing price model in Beijing and analyzing its influencing factors. The results are as follows: the most objective factors affecting secondary housing prices are the number of rooms, the ladder ratio, and elevators and the trade time has the most negligible impact.

Because of the control of some conditions, this study has some limitations. In this study, we used the data on second-hand housing prices and their influencing factors in Beijing from 2017 to 2018. We use relatively little data, which will lead to the research results being limited to some periods. More data collection and in-depth discussion are needed to obtain more widely used methods. In addition, there may be more accurate means and technologies to improve the accuracy of this model in the future, and we still have much work to do. To improve our model as much as possible.

DOI: 10.54254/2754-1169/15/20230943

References

[1] *Wei Fang, Heng Xiao, Xiangchen Ren. (2009). Logistics demand forecasting analysis based on the linear regression model. Productivity research, 12, 94-110.*

[2] *Nguyen, D., Smith, N. A., & Rose, C. (2011, June). Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities (pp. 115-123). doi:10.1184/R1/6473069.*

[3] *Yong Zhou, Yu HuanXiao, Runsheng Huang. (2011). Forecast the grain yield of Guangxi based on the multiple linear regression model. Journal of Southern Agriculture, 42(9), 1165-1167.*

[4] *Yu Kang, Zhaofeng Wang, Jun Liu. (2012). Application of multiple linear regression in the prediction of gas drainage. Coal engineering, (3), 59-60.*

[5] *Cakra, Y. E., & Trisedya, B. D. (2015, October). Stock price prediction using linear regression based on sentiment analysis. In 2015 international conference on advanced computer science and information systems (ICACSIS) (pp. 147-154). IEEE.*

[6] *Edelman, E. R., Van Kuijk, S. M., Hamaekers, A. E., De Korte, M. J., Van Merode, G. G., & Buhre, W. F. (2017). Improving the prediction of total surgical procedure time using linear regression modeling. Frontiers in medicine, 4, 85. doi:10.3389/fmed.2017.00085*

[7] *Kadri, T., & Sugara, R. D. H. (2017, November). Estimated budget construction housing using linear regression model accessible and fast solutions accurate. In 2017 International Conference on Computing, Engineering, and Design (ICCED) (pp. 1-6). IEEE. DOI: 10.1109/CED.2017.8308095.*

[8] *Acharya, M. S., Armaan, A., & Antony, A. S. (2019, February). A comparison of regression models for prediction of graduate admissions. In 2019 international conference on computational intelligence in data science (ICCIDS) (pp. 1-5). IEEE. DOI: 10.1109/ICCIDS.2019.8862140.*

[9] *Xinxin Hu. (2019). Influencing Factors and Regression Analysis of GDP in Yunnan Province. Statistics and Application, 8, 581.*

[10] *Ran Zhao. (2020). Analysis of the Correlation between Housing Price Data in Boston Based on the Regression Method. Statistics and Application, 9, 335.*

[11] *Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley Series in Probability and Statistics, vol. 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679.*

[12] *Housing price in Beijing. (2018). Kaggle. Retrieved October 12, 2022, from https://www.kaggle.com/datasets/ruiqurm/lianjia*

[13] *Yi, M. (2021). Scatter Plots | A Complete Guide to Scatter Plots. Chartio. Retrieved October 12, 2022, from https://chartio.com/learn/charts/what-is-a-scatter-plot/*