

Research on the Impact of the Rent in Shanghai Based on Multiple Linear Regression Model

Yichuan Zhang*

School of mathematical science, Tongji University, Shanghai, 200092, China

*Corresponding author: 2051092@tongji.edu.cn

Abstract. In today's era, renting has become one of the most important ways of living for young people. Therefore, finding the main influencing factors of rent has important guiding value for both tenants and the government. By establishing a multiple linear regression model, this paper analyzes about 2,900 rental data collected from Shanghai, and obtains the relationship between Shanghai housing rent and some main influencing factors. Some variables are processed to be incorporated into a multiple linear regression model. In the process of establishing the model, the variables that have a significant impact on rent are screened out by significance testing. The model shows that the zone where the house is located and the type of lease have a greater impact on rent, while other factors such as the size of the house have less impact on rent. These results can help tenants make better choice about which to rent and help the government to regulate rental prices.

Keywords: House rent, Multiple linear regression model, Price estimate.

1. Introduction

1.1. Background

With the development of society, renting is accepted by more and more people as a more economical life style. Since most tenants will consider rent first when renting and it is known to all that the level of rent is related to many factors, the impact of rent is an important issue to be researched. Finding what causes the increase in price is helpful for tenants to choose the most affordable and suitable accommodation. A suitable house can both save money and avoid wasting space, which is especially important for employees with low salaries. Therefore, it is important to find what properties of the house itself (area, the number of rooms, the number of bathrooms, etc.) affect rent the most. Not only that, zone in which the house is located is also an important factor. Some people choose to work in the center of the city for higher salaries, but rents in the center of the city will rise accordingly. Therefore, finding out the difference in rent in different zones can help people choose where they work. On the other side, it will help the landlord or the government to make a more reasonable price.

1.2. Related Research

The linear regression model is one of the most widely used techniques to analyze the factors that influence the change in price or something difficult to predict only by one factor. There is amount of research under the linear regression model. Xie and Wan collected the data in the financial report of 55 A-share listed companies and used a multiple linear regression model to research the correlation between stock pricing and some financial variables. They drew a conclusion that multiple linear regression is a reliable tool for analyzing stock pricing [1]. Zheng et al divided the employee and the different industries they work in into several groups, each group is a variable added to the multiple linear regression model to consider the impact on GDP. The GDP is strongly related to the region according to the model [2].

The price prediction is also an application of linear regression model, among which the analysis and prediction of house price are meaningful. Jiang used the data of regional revenue, fixed investment, GDP per capita, etc. to find the connection between these factors and the rent in different cities in Zhejiang province. The multiple linear regression model was used to clarify the impact of each factor and predict the rent [3]. In order to establish a model of real estate prices, Zhong elaborated

on a linear regression model to estimate the real estate prices by analyzing data such as location, structure and traffic, providing an effective prediction of real estate prices [4]. Zhang established a model by linear regression to study the commercial house price in Xi'an. The study showed that the price has a strong linear correlation with GDP, the total population in a certain area and some other factors. So, the conclusion was that it is reliable to predict the house price by linear regression model [5]. Dang and Yang studied the house prices in Tangshan and divided the factors into demand factors and supply factors. After a multiple linear regression model was established, a multicollinearity test was done to correct the model, the result was that the demand factor has a more important influence than the supply factor [6]. Dai and Li studied the Second-hand housing prices in a certain area in Chengdu and their influencing factors, including continuous variables like area, non-continuous variables like the number of bedrooms, the floor, whether the house is furnished, whether the house is lifted, etc. All the variables were used in the multiple linear regression model to describe the price. The conclusion showed that all the variables are significant and related to the price change, though some factors have interaction [7]. In the same way, Li and Zhuang focused on the rent in Shijiazhuang. In the research, they found there is multicollinearity among the variables. So, the multiple linear regression model was adjusted to find some main factors influencing the rent. By LASSO regression, the rent varies a lot in different zones and for the house itself, the significant influencing factors are the area, hall, low floor, elevator, etc. [8].

1.3. Objection

The objective of the research is to establish a multiple linear regression model of rent. In this paper, data are collected based on the housing rental market in Shanghai, aiming to find out the influencing factors of housing rental in Shanghai. Through the model, it is expected to find a linear expression between different variables and rent. In the model, there will be a coefficient in front of each independent variable which indicates its effect on the dependent variable. After obtaining the corresponding coefficients before each variable, it can be known which one of the variables has a greater impact on the rent. The model can help the tenant to obtain the main factors affecting the rent and help the government to better regulate the rent price.

2. Methods

2.1. Data Sources

In order to find the relationship between rent and factors that might have an influence on it, the data need to be stable for a period of time. For making a comparison, the data are collected from www.heywhale.com. On this website, there are house rent data in Shanghai from 2020.03.03 to 2020.03.09, collected from shell net. All data contain the same variables, including zone, address, type (Entire lease or Share lease), price, area, house orientation, house type (The number of bedrooms, halls and bathrooms) etc. Since the prices are nearly the same and all the data are constant during this period of time, the data of 2020.03.03 are selected to represent this period of time. In this day's data, there are a total of 2900 samples, representing 2900 houses/apartments in Shanghai.

2.2. Data Preprocessing

There are 2900 samples containing data of the house on the website in total. Among these data, 17 samples house types are missing. Due to the small number of these samples, deleting them will not have a greater impact on the model, so these samples are deleted. Other samples' data are completed. In order to understand the distribution of the data, the average rent and house size in the remaining data was calculated. The maximum rent is ¥27000 and the minimum is ¥1000. The average rent of houses in the remaining sample is 3644.36, and the standard deviation is 2441.06. For area, the maximum area is 360 square meters and the minimum is 5 square meters. The average area of houses

in the remaining sample is 46.47 and the standard deviation is 30.99. The level of bias in the data is acceptable, so all the remaining data are used in making the model.

In addition to these numerical data, there are a lot of non-numeric data in the original data. These data will be coded as 0 or 1 as dummy variables and then included in the multiple linear regression model.

2.3. Model

Regression is a method widely used in statistics and data analysis. It obtains the correlation between variables by processing a large amount of data obtained by observation or experiment. For the correlation between variables, their relational expressions in the mean sense can be determined through regression analysis. In addition, regression analysis can determine which independent variables have a more significant impact on the dependent variable in a series of independent variables that affect the dependent variable. Multiple linear regression is especially suitable for dealing with many independent variables and a single dependent variable, and the results obtained can help people predict changes in dependent variables [9].

3. Making Model

The specific process of building the model is divided into three parts. In the first part, the variables are explained in detail. Some of the variables are transformed so that they can be used in the multiple linear regression model. Some of the variables from the original data are deleted. In the second part, these variables are contained in a multiple linear regression model to show how these variables relate to rent. Some variables have been adjusted to make the results more accurate. In the third part, using the data from the model, the relationship between all independent variables and rent is analyzed

3.1. Variable Selection

The independent variable in the model is Rent, in RMB. The dependent variables are as follows:

Area: The size of the house/apartment in square meter.

Bedroom: The number of bedrooms in the house/apartment.

Hall: The number of halls in the house/apartment.

Bathroom: The number of bathrooms in the house/apartment.

Type: There are two types of rental styles. The entire lease means the house/apartment is rented out to tenants as a whole without sharing with others. The shared lease means the house/apartment is rented out to generally two tenants. Based on experience, the rent for the entire lease type will be higher than that of the shared type. So dummy number 1 is used here to represent entire lease and dummy number 0 is used here to represent shared lease.

Zone: The area where the house is located, involving 16 districts in Shanghai. To simplify this variable, the sixteen regions are divided into central and non-central regions. According to the population and economic development level of each district, the central district is defined as Huangpu District, Jing'an District, Xuhui District, Changning District, Hongkou District, Yangpu District, Zhabei District, Putuo District and Pudong New District. The rest of the area is a non-central area [10]. Based on experience, the rent for house/apartment in central regions will be higher than that in non-central regions. So dummy number 1 is used here to represent central regions and dummy number 0 is used here to represent non-central regions.

House orientation: The orientation of the windows in the house. The house orientations in the sample have eight orientations: east, south, west, north, southeast, northeast, southwest, and northwest. Some of the samples have more than 1 orientation. Based on experience, south-facing houses have many advantages and are welcomed by most of the people. So dummy number 1 is used here to represent south-facing houses and dummy number 0 is used here to represent others.

The variable address in the original data is not included in the model because the addresses are too detailed and are different for each sample, making comparisons impossible. After sifting, there remain 7 independent variables.

Table 1 below shows some detailed information about these variables:

Table 1 Data details

Variables	Minimum	Maximum	Mean
Bedroom	0	5	3.42
Hall	0	4	0.97
Bathroom	0	5	1.22
Rent of central area located houses/apartments	1000	27000	3873.30
Rent of non-central area located houses/apartments	1000	25000	2902.87
Rent of entire rental houses/apartments	1000	27000	4171.90
Rent of shared rental houses/apartments	1000	4950	1915.17

From the data, it can be roughly seen that area and lease type have an impact on rent, but the specific relationship between them is not clear enough. These variables need to be included in a linear regression model for a full analysis.

3.2. Research Results and Improvement

There are a huge number of variables in each sample. Changes in some of these variables may have a large effect on rents, while changes in other variables may have so small effects on rents that these variables can be ignored. To filter out these valid variables, all variables were included in the model for significance level testing. Variables that can't pass the significance level test will be deleted and some of the variables will be adjusted to get a more accurate model.

3.2.1. Establish Linear Regression Model

Under the condition that all 7 independent variables are contained in the model, it is expected to get formula (1) in this type:

$$Rent = a_1 * Area + a_2 * Bedroom + a_3 * Hall + a_4 * Bathroom + a_5 * Type + a_6 * Zone + a_7 * House orientation + b \quad (1)$$

By doing multiple linear regression, Table 2 below can be made.

Table 2 Linear regression model

Variables	coefficient	p-value	t Stat
Intercept	1345.252	0.0000	4.284315
Area	35.67827	0.0000	14.94457
Bedroom	-278.613	0.0003	-3.60809
Hall	72.79484	0.4267	0.794917
Bathroom	225.9806	0.0109	2.546307
Type	535.3898	0.0228	2.277755
Zone	1118.379	0.0000	12.03019
Orientation	-495.287	0.0244	-.31982
Adjusted R-square:0.29300			

According to this model, there is a correlation between rent and most of the variables and this model does show some of the coefficients in an acceptable way. However, there are also lots of disadvantages to this multiple linear regression model. The coefficient before the variable bedroom and orientation in this model is negative, which is not equal to the experience in daily life. The P-value of variable hall, bathroom, type and orientation are all more than 0.0005, meaning that they are not significant under the level of 95%. The adjusted R-square is only 0.29300, meaning this model is not very persuasive. These problems conduct a message that the model needs to be improved.

3.2.2. Improved Linear Regression Model

To solve these problems, the original data has been corrected. First, the number of bedrooms and bathrooms under the shared type can't represent the true situation because most the house/apartments are shared by two people. When checking in, only half of the total number of rooms and bathrooms can be used by each person. In order to get a more accurate model, the number of bedrooms and bathrooms under the shared type is divided by two and the number of halls remains the same. Second, to reduce errors, the natural logarithm of variable rent and area has been taken, so that the range of the variables can be narrowed [11]. Third, variable orientation is deleted because it has a low P-value and the coefficient before it doesn't reflect the real condition.

After correction, it is expected to get formula (2) in this type:

$$\ln Rent = a_1 * \ln Area + a_2 * Bedroom + a_3 * Hall + a_4 * Bathroom + a_5 * Type + a_6 * Zone + b \quad (2)$$

By doing multiple linear regression, Table 3 below can be made.

Table 3 [Improved linear regression model]

Variables	Coefficient	p-value	t Stat
Intercept	6.75615	0.0000	107.7202
Area	0.108696	0.0000	4.546939
Bedroom	0.089533	0.0000	7.447088
Hall	0.007392	0.7237	0.353563
Bathroom	0.101158	0.0000	5.244925
Type	0.479264	0.0000	16.04428
Zone	0.308342	0.0000	15.6715
Adjusted R-square:0.36326			

In the improved model, except for variable hall, all variables have very low P-value and all coefficients are positive. After model improvement, the adjusted R-square is 0.36326, which is more acceptable. All variables have passed the significance level test except for the number of bathrooms, meaning they are significantly under the level of 95%. So, this model can be used to analyze the impact of these variables on rent.

3.3. Variable Analysis

Using the sheet above, all variables have a positive effect on rent. The independent variables in the model can explain 36.326% of the variation in the dependent variable, which is not very high but acceptable. Specifically, for each variable, the analysis is as follows:

Area: According to the model, under the condition that other variables remain unchanged, for every 10% increase in area, rent increases by about 1.0% on average. That means the size is not an important factor in the rent change. Area should be considered with other factors.

Bedroom: According to the model, under the condition that other variables remain unchanged, for every additional bedroom, rent increases by about 9.4% on average.

Bathroom: According to the model, under the condition that other variables remain unchanged, for every additional bathroom, rent increases by about 10.6% on average.

Type: According to the model, under the condition that other variables remain unchanged, the rent of the entire rental house/apartment is 61.5% higher than that of shared rental house/apartment.

Zone: According to the model, under the condition that other variables remain unchanged, the rent of house/apartment located in the central area is 36.1% higher than that in the non-central area.

The results of the model show that the number of living rooms has no significant effect on rent. Variable type and area have a greater impact on rent, while the number of bedrooms and bathrooms has a low impact on rent. Without considering other variables, the increase in the size of the house has little effect on the rent, which means that the size of the house itself has a very small effect on the rent.

4. Conclusion

Based on the multiple linear regression model, the relationship between each variable and rent is well shown, from which many useful conclusions can be obtained. The size of the property is no longer the most important factor in determining rent. On the contrary, the number of bedrooms and bathrooms has a greater influence. If a tenant wants to save money, the house/apartment with fewer rooms but a larger area in each room can be chosen. Of all the variables, type and zone have the biggest impact on rent, with entire rental houses/apartments even being 60% more expensive than shared houses/apartments on average under the condition that other variables remain unchanged. Shared housing can be a great option for tenants looking to save money in house renting. And governments, need to develop a better pricing strategy to make the pricing between regions with different economic levels more scientific. The government should also pay more attention to rent control so that all pricing is within a reasonable range.

Although the effectiveness of the model has been improved based on the above improvements, there are still some limitations. Since the house rent varies greatly with time, the variables in the model can only represent the rent in this small period of time, which has a strong limitation in time. Also, there should also be multicollinearity relationships between variables in the model, and simple multiple linear regression models cannot show these correlations. To improve the model, we need to look for more relevant rent influencing factors, such as distance from the city center and accessibility. Finding correlations between these variables and trying a non-linear regression might give more accurate results.

References

- [1] Chunyu Xie, Wenjun Wan. Discrimination and Analysis of Influencing Factors of Stock Prices in Secondary Market——Multiple Linear Regression Based on R. *China Business Review*, 2022(15):99-102. DOI:10.19699/j.cnki.issn2096-0298.2022.15.099
- [2] Wei Zheng, Yao Xu, Jun Yang, Shuhuan Yang. Research on the Impact of Employment on GDP Based on Multiple Linear Regression Model. *Journal of Economics and Management: Chinese and English Editions*, 2022, 11(1):1 -8.
- [3] Yuting Yang. Influencing factors and prediction of rental housing based on multiple regression *Journal of Wenzhou University (Natural Science Edition)*, 2020, 41(4): 10-15. DOI: 10.3875/j.issn.1674 -3563.2020.04.002.
- [4] Xiaojin Zhong. Real estate price assessment based on multiple linear regression analysis. *Business Times*, 2014(36):133-134. DOI:10.3969/j.issn.1002-5863.2014.36.056.
- [5] Yanbo Zhang. Prediction and analysis of housing prices in Xi'an based on multiple linear regression analysis. *Western Leather*, 2020,42(10):71.
- [6] Guangyuan Dang,Tao Yang. Multiple Linear Regression Analysis of Influencing Factors of House Prices in Tangshan City. *Journal of Hebei United University (Social Science Edition)*, 2014(2):21-25.
- [7] Lei Dai, Xueting Li. Analysis of Influencing Factors of Second-hand Housing Prices Based on Multiple Linear Regression Model: Taking a District in Chengdu as an Example. *Henan Building Materials*, 2019(5):80-82. DOI:10.3969/j.issn.1008-9772.2019.05.046.
- [8] Chunlin Li, Zelin Zhuang. Analysis and Excavation of Factors Influencing Housing Rent in Shijiazhuang. *National Circulation Economy*, 2020(31):129-131. DOI:10.3969/j.issn.1009-5292.2020.31.042.
- [9] Junfei Wang. Prediction Model of Commodity Housing Price Based on Multiple Linear Regression. *Science and Technology Vision*, 2013(10):210,203.
- [10] Yirong Yu. Research on the Basic Influencing Factors of Housing Rental Prices Based on Multiple Linear Regression: Taking Shanghai Housing Rental Market as an Example. *Industrial Innovation Research*, 2021(20):79-81.
- [11] Shicheng Yu, Qiqi Wang, Xiaojuan Long, et al. Multiple linear regression with natural log transformation of variables. *Chinese Journal of Preventive Medicine*, 2020, 54(4): 451-456. DOI: 10.3760/cma.j.cn112150-20191030-00824.