ISSN:2790-1661 DOI: 10.56028/aemr.2.1.71

Research on Housing Price in Boston

YuBo Zhang, Tan Wang , and Zhi heng Ma

Northwestern Polytechnical University

617409144@qq.com

Abstract. In the United States, due to the characteristics of sparsely populated areas, its housing prices are not quite the same as those in China. At the same time, due to factors such as the freedom of firearms in the United States, many people consider more factors in purchasing houses, resulting in great differences in housing prices in the United States. The housing price problem has also become a major concern.

This paper first analyzes fourteen variables, respectively, thirteen independent variables were analyzed to find the relationship between independent variables and housing prices, establish regression model analysis, and then through the frequency histogram to find the relationship between independent variables and frequency, through the relationship between frequency and housing prices to illustrate the relationship between independent variables and housing prices. By analyzing the correlation to find out the relationship between the variables and housing prices.

Through the statistical description and analysis model of data, we can establish the normal distribution relationship between independent variables and dependent variables, and approximate the normal distribution parameters through the average and variance of independent variables. The relationship between each independent variable and dependent variable is fitted, and the influence of each variable on housing prices is finally obtained.

Keywords: regression analysis, frequency, numerical analysis, orthogonal fitting

1. Enviroment of problems

1.1 Scenario

In today 's China, housing has become an indispensable part of our daily life due to the influence of the traditional idea of land relocation and the rapid economic development of new China. Therefore, in order to better predict and understand the distribution and trend of housing prices in China, we can start with the price distribution of housing in the past and the impression factor of livability, and make a more specific analysis of housing forms in China.

Based on data collection for 14 variables, Harrison and Rubinfield collected data on per capita crime rates, the proportion of large-scale occupied housing areas and the proportion of non-retailed commercial housing areas, which were analysed by belsley et al in 1980

In order to predict and analyze the data, we collate and summarize the data.

1.2 Problem Posing

China 's housing prices are rising rapidly, and the prediction of housing prices and the improvement of livability have become the rigid demand of the people. Housing is not only a place for people to live, but also an important financial product for Chinese investment and financing.

There are 14 groups of data in this analysis, from the per capita crime rate to the proportion of students / teachers. These data have an independent or joint impact on house prices. To further study the mathematical model, the following tasks are proposed :

1. Through exploratory data analysis (histogram, correlation analysis, etc.), the impact of variables on housing prices in Boston is explained in detail.

2.Establish housing price forecasting model in Boston and evaluate the forecasting results.

3.Houses in Boston are classified and analyzed according to their livability.

4.Based on the analysis and results of the above problems, write a report of no more than one page, give different groups of housing suggestions.

2. Problem analysis

Through the analysis of the data provided in the topic, it is not difficult to find that each data can be seen as an independent variable to affect the house price. Finally, the variables are combined with the algorithm to obtain the relationship between house prices and variables.

2.1 Problem One Analysis

Firstly, the relationship between variables and house prices is studied separately, and the house price histogram, frequency histogram, average house price histogram and related data are analyzed. Based on the interpolation curve and normal distribution principle, the mathematical relationship between variables and house prices is preliminarily constructed, and the influence of each variable on house prices is analyzed.

2.2 Problem Two Analysis

Based on the statistical description and analysis model of data, we first make two assumptions :

1.Suppose variables are normal distribution or elementary function combination distribution.

2.Assume that each variable can affect the results independently.

Based on these two assumptions, we interpolate the histogram of the relationship between variables and housing prices. For the normal distribution model, we approximate u with the mathematical expectation of variables, and approximate σ with the variance of variables. For the elementary function combination distribution, we use Newton interpolation, Lagrange interpolation and Hermit interpolation respectively, and choose the multi-step method with higher relative accuracy.

2.3 Problem Three Analysis

By analyzing and comparing the influence of various variables on the livability, we define the livability factor and classify the houses in Boston by comparing the livability factors.

2.4 Problem Four Analysis

Given that different housing purchase groups have different requirements for houses, we classify housing sources according to livability, school district housing, security, luxury and convenience. Each classification has different requirements for variables. Referring to the solution of the third question, we define the requirements factor by analyzing and comparing the influence of each variable on the classification requirements, and conduct a comprehensive evaluation of Boston houses and give suggestions by comparing the requirements factors.

3. Problem solving

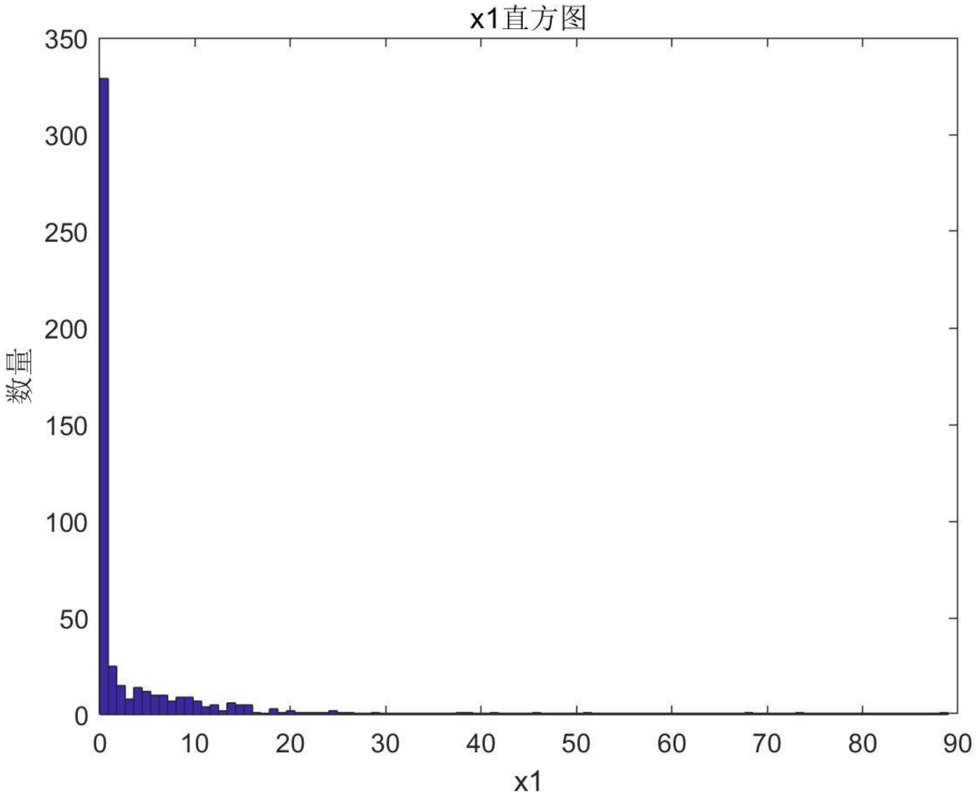
3.1 Problem I

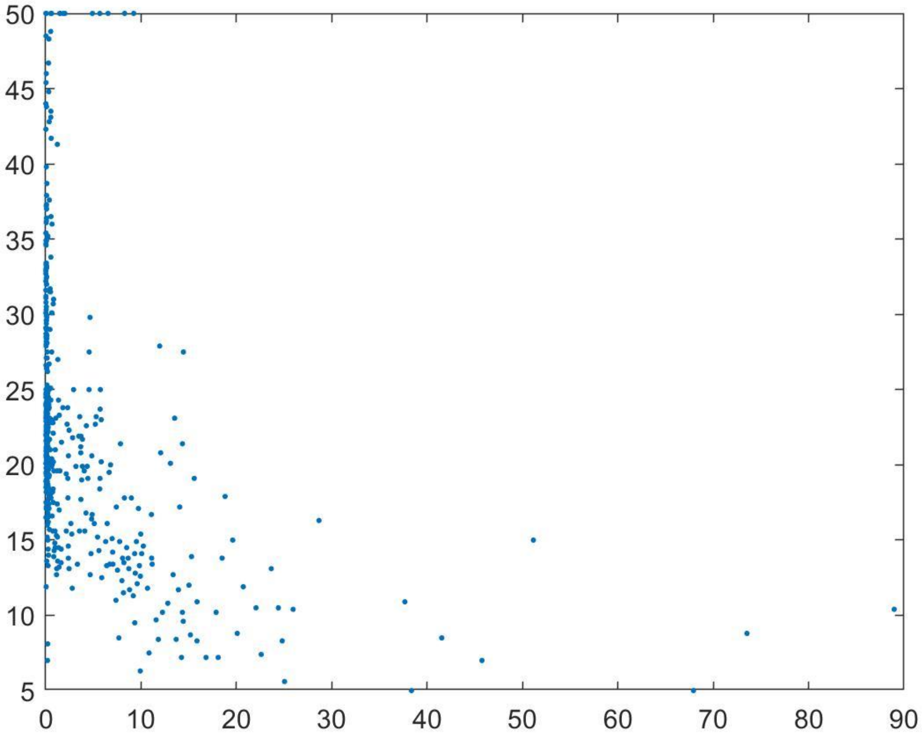
3.1.1 Data Analysis

Firstly, according to the existing data, we draw the house price scatter diagram, frequency histogram and average house price scatter diagram. Compare the three maps and calculate the statistical data of each variable.

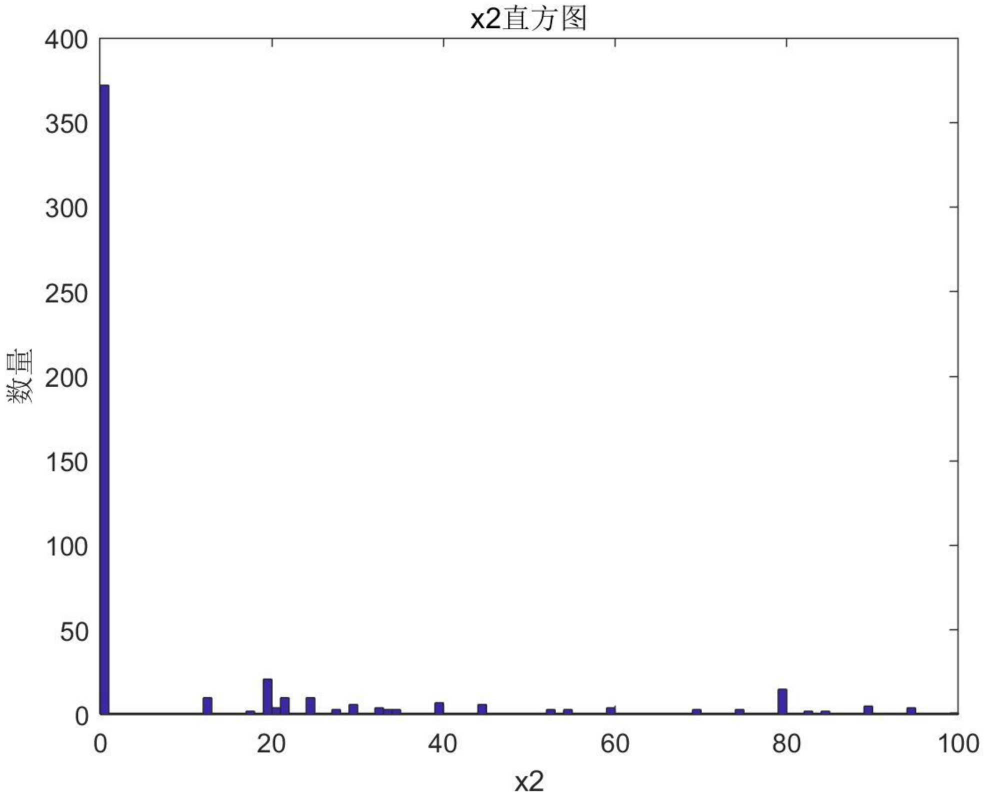
3.1.2 problem and explain

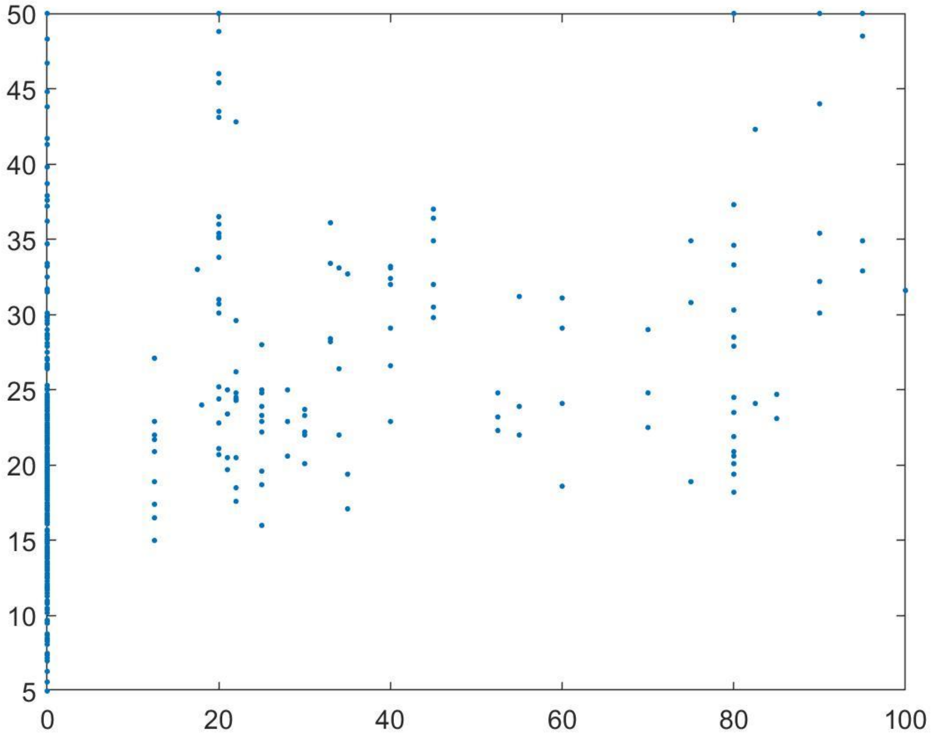
X1 is the per capita crime rate, and it can be seen from the scatter plot of housing prices that housing prices in places with low per capita crime rate are generally high, and the per capita crime rate is inversely proportional to housing prices and changes in an exponential form ; it can be seen from the frequency histogram that people generally choose places with low per capita crime rate, and when the sample size is large enough, it can be considered that people generally choose places with zero per capita crime rate, so the housing price is higher in places with low per capita crime rate.



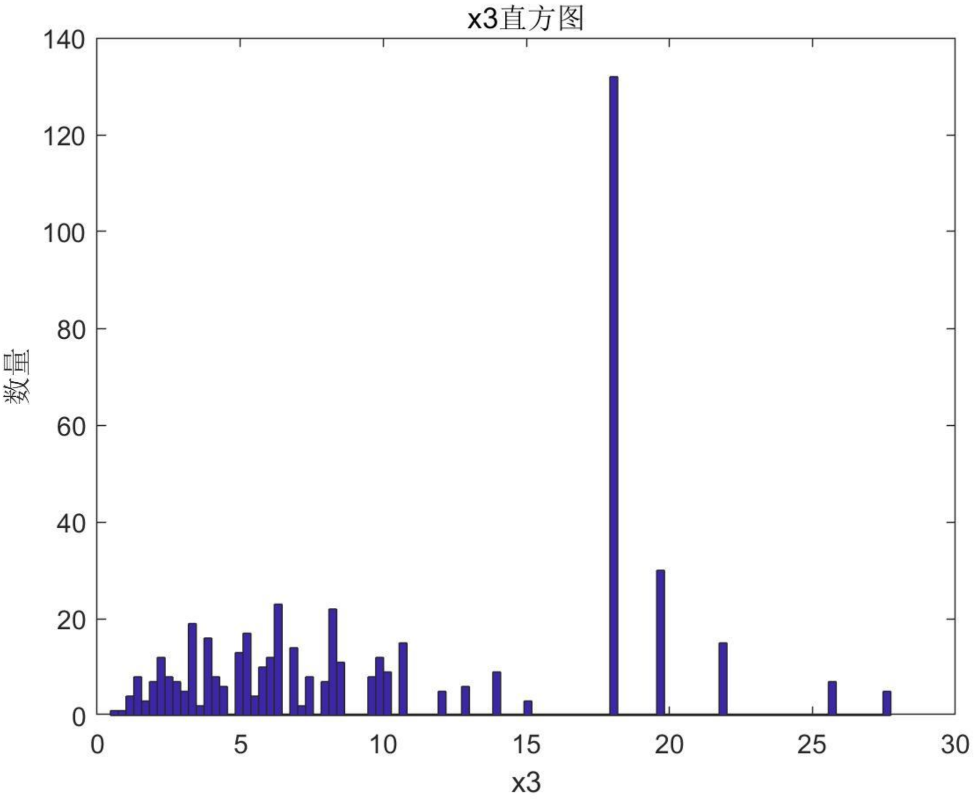


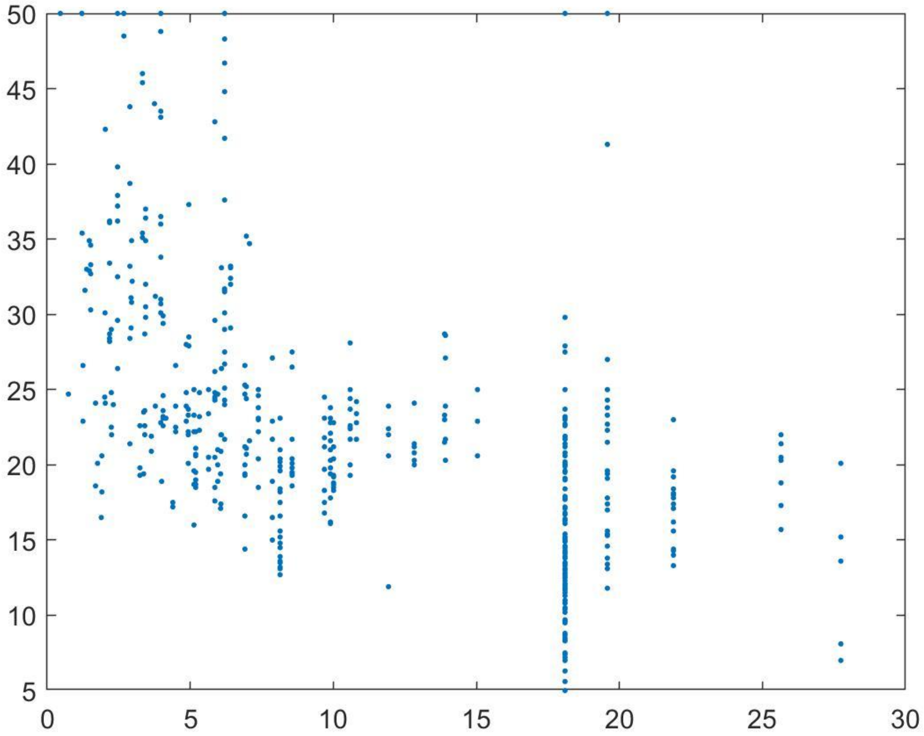
X2 is the proportion of large-scale residential areas. It can be seen from the scatter plot of house price that when the proportion of large-scale residential areas gradually decreases, the house price generally has a downward trend, and the house price is extremely poor. When the proportion of large-scale residential areas is between 40 % and 60 %, the house price remains basically unchanged. On the whole, there is a positive correlation between the proportion of large-scale residential areas of independent variables and the housing price of dependent variables. It can be seen from the frequency histogram that more people like to buy houses in areas with a low proportion of large-scale residential areas, and a few people will buy houses in areas with a high proportion of large-scale residential areas, which is in line with the proportion of the rich and the poor in terms of the number of people. Therefore, it can be approximated that the proportion of large-scale residential areas is proportional to housing prices.





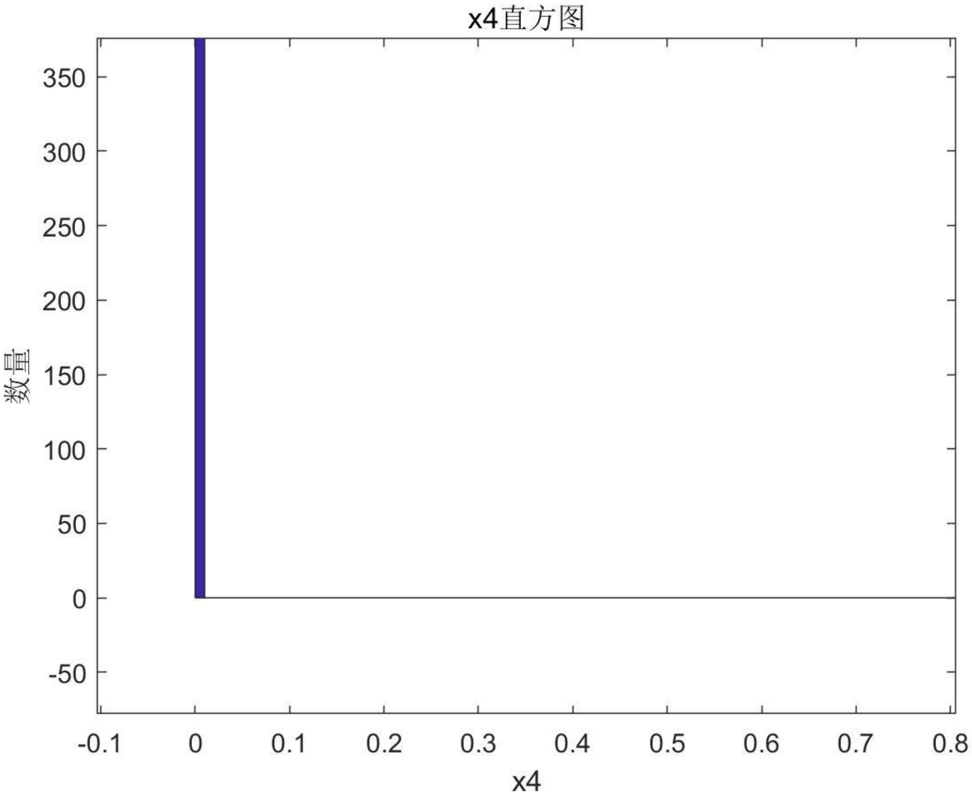
x3 is the proportion of non-retail business land. It can be seen from the scatter plot of house price that the house price is generally high in places where the proportion of non-retail business land is low, while the house price remains stable in the region where the proportion of non-retail business land is between 7 % and 17 %. When the proportion of non-retail business land is higher than 17 %, the house price shows a downward trend. Combined with the frequency histogram, it is easy to find that the number of people selected is the largest in the part where the proportion of non-retail business is about 17 %, while the number of people selected is small in the part where the proportion of non-retail business is low, which is in line with the proportion of the rich and the poor. Therefore, it can be seen that the impact of the proportion of non-retail business on housing prices : the housing price is higher in the place where the proportion of non-retail business is low, and the two are approximately inversely proportional.

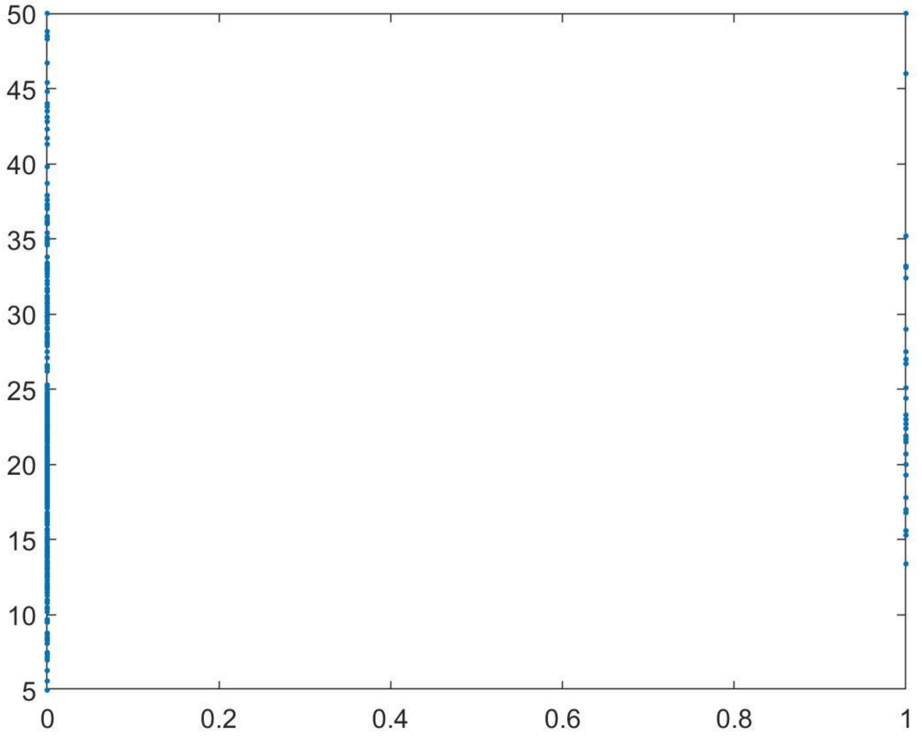




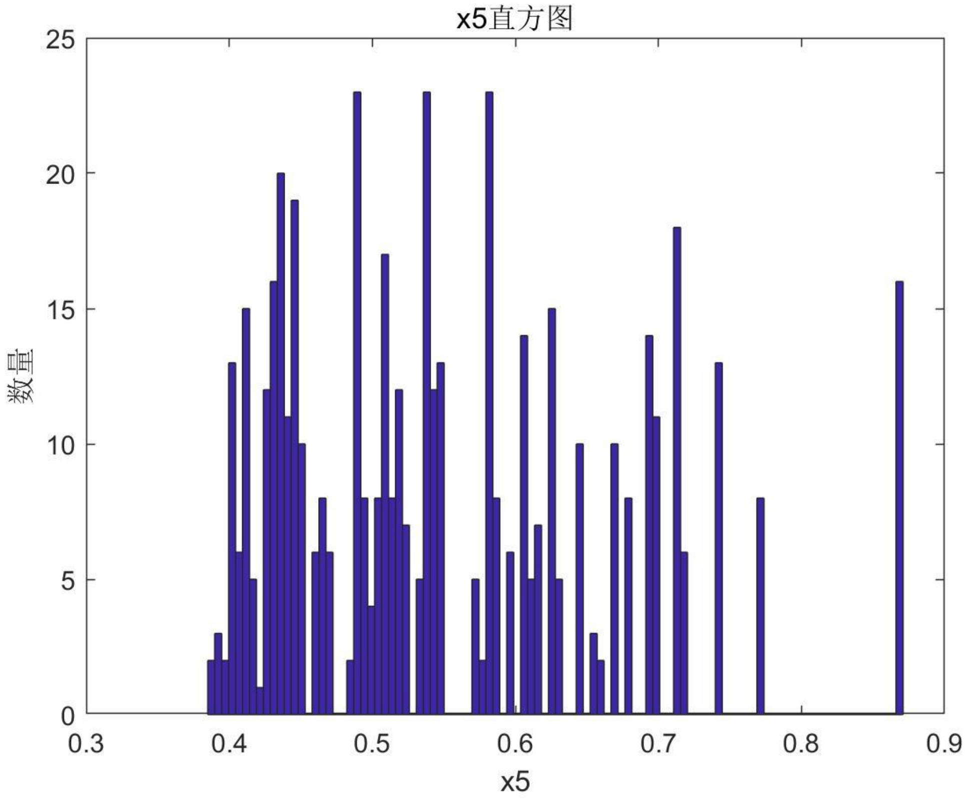
X4 is the virtual variable of the Charles River. Through the scatter diagram of house prices, it can be seen that when the virtual variable of the Charles River is 0 or 1, the house prices are from low to high, but the average house price of the Charles River is about 22.1 when the virtual variable of the Charles River is 0, which is lower than the average house price of the Charles River when the virtual variable of the Charles River is 1, and the range is larger when the virtual variable of the Charles River is 0. It can be seen from the frequency histogram that the frequency of the Charles

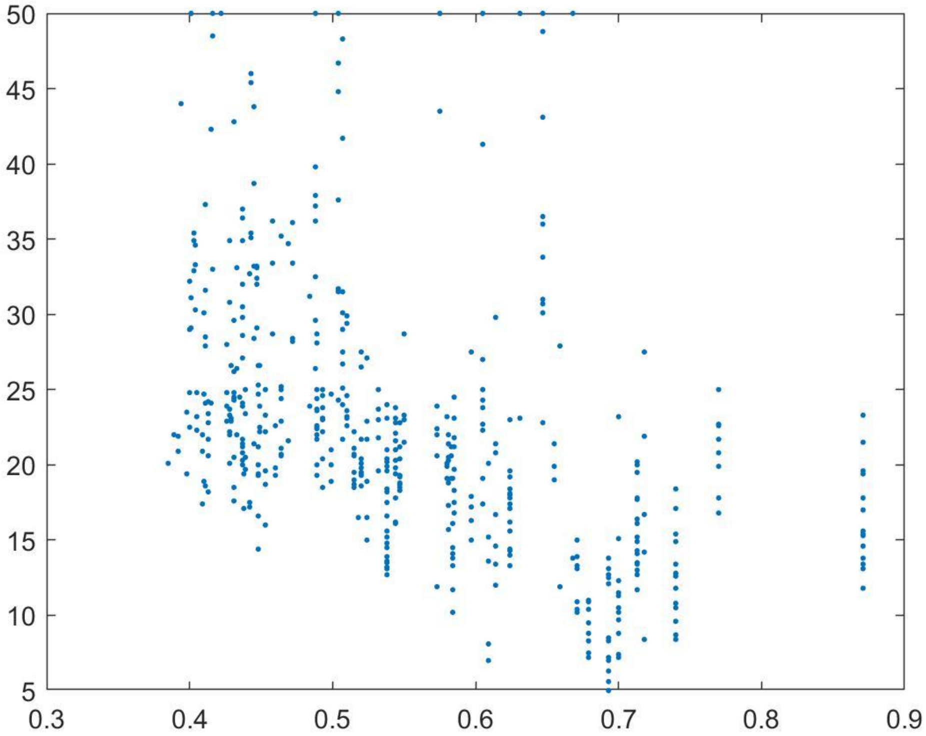
River virtual variable is 0 is larger, and the frequency of 1 is smaller. Therefore, more people choose the Charles River virtual variable with low average housing price as 0.



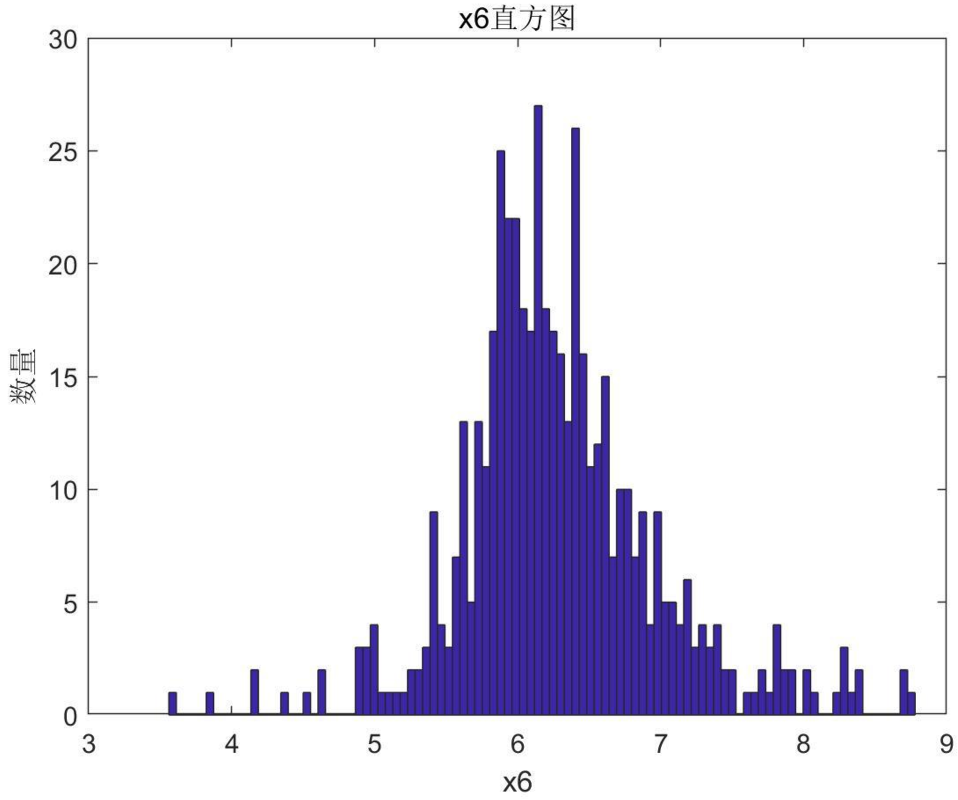


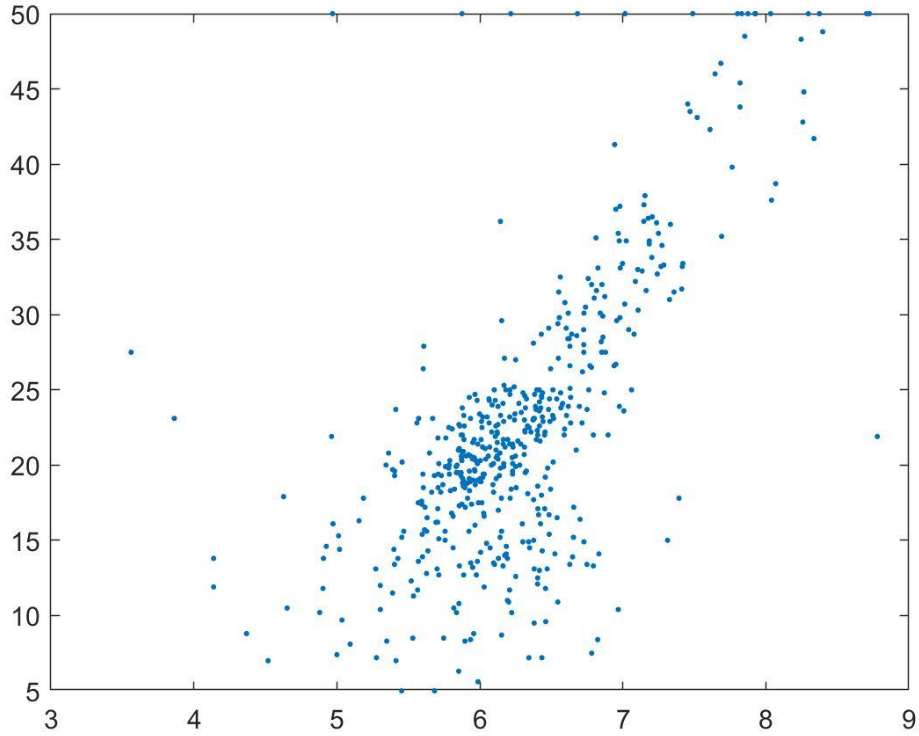
x5 is the concentration of nitrogen oxides. It can be seen from the scatter plot of housing prices that the concentration of nitrogen oxides is inversely proportional to housing prices. The lower the concentration of nitrogen oxides is, the higher the average housing price is. It can be seen from the frequency histogram that the frequency of nitrogen oxide concentration is approximately in line with the normal distribution, and u is about 0.55. The frequency histogram indicates that most people are used to choosing houses with moderate nitrogen oxide concentration, and compared with a few people, they choose houses with high or low nitrogen oxide concentration.



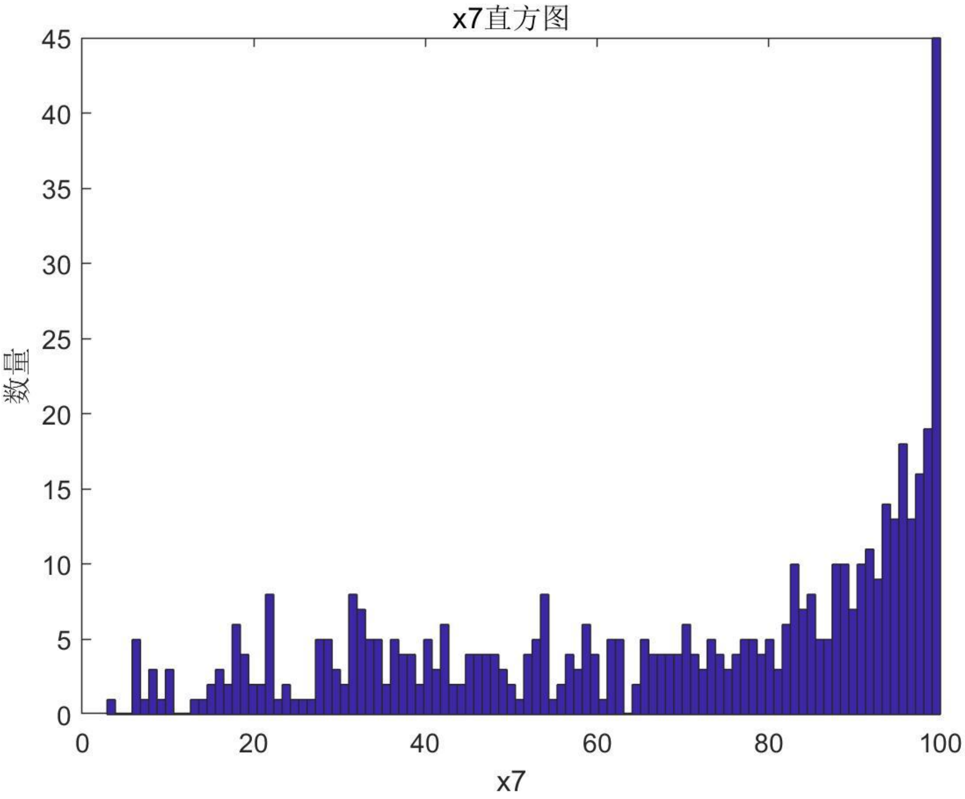


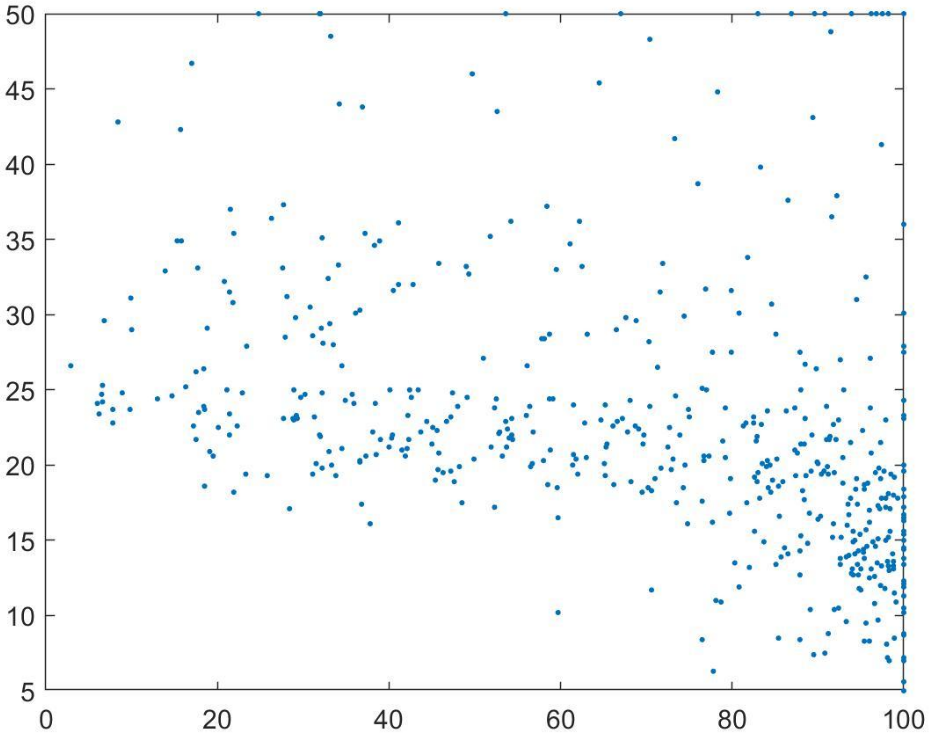
x6 is the average number of rooms per household. It is easy to see through the scatter plot of house price that the more the average number of rooms per household is, the higher the house price is. Through the average fitting, it can be found that the relationship curve is proportional. It can be seen from the frequency histogram that the number of rooms is in line with the normal distribution, u = 6.2. The graph shows that most people choose the house with moderate room number, and a few people choose the house with less room number and more, which is in line with the proportion of the rich and the poor. Therefore, the average room number per household is proportional to the housing price.



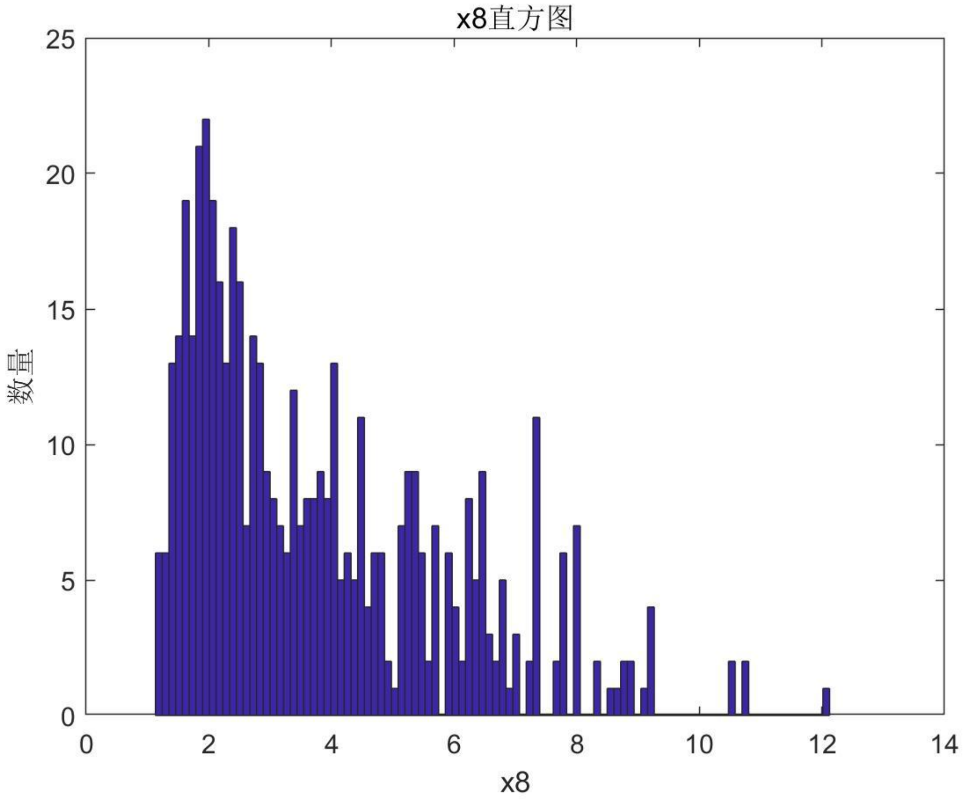


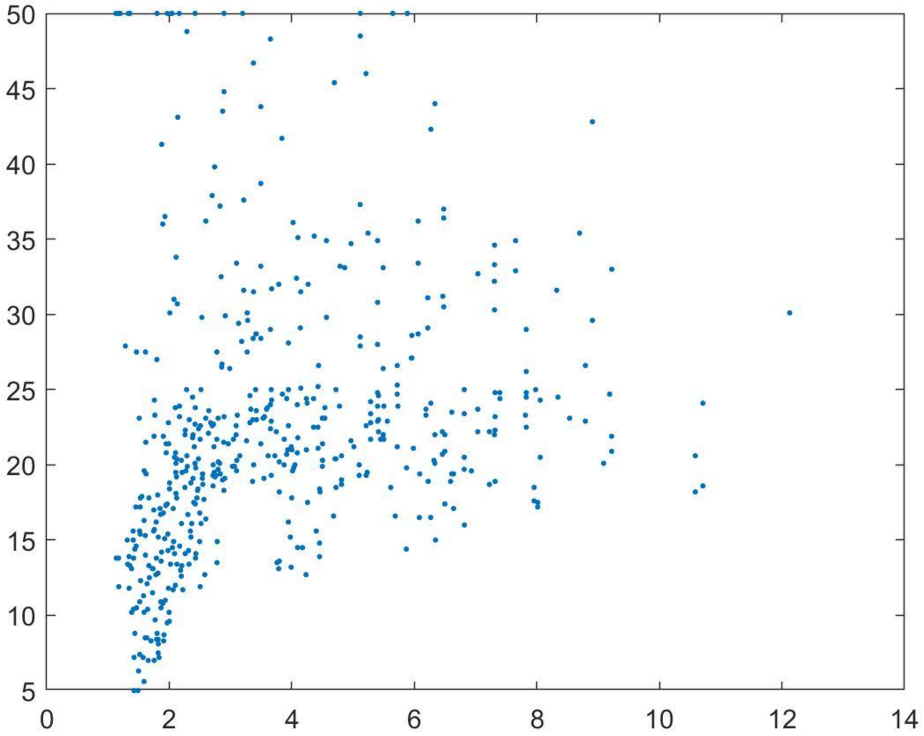
x7 is the ratio of owner-owned housing constructed before 1940. From the scatter plot of house prices, it can be seen that the ratio of owner-owned housing constructed before 1940 is approximately inversely proportional to house prices, and the slope of the overall function is small. When the ratio of owner-owned housing constructed before 1940 is 0, the house price approximate value is about 30, and when the ratio of owner-owned housing constructed before 1940 is 100, the house price approximate value is about 17. It can be seen from the histogram of frequency that the higher the proportion of house owners built before 1940 is, the higher the frequency is. According to the relationship between supply and demand and price, the lower the house price is. Therefore, it can be proved that the higher the proportion of house owners built before 1940 is, the lower the approximate value of house price is.



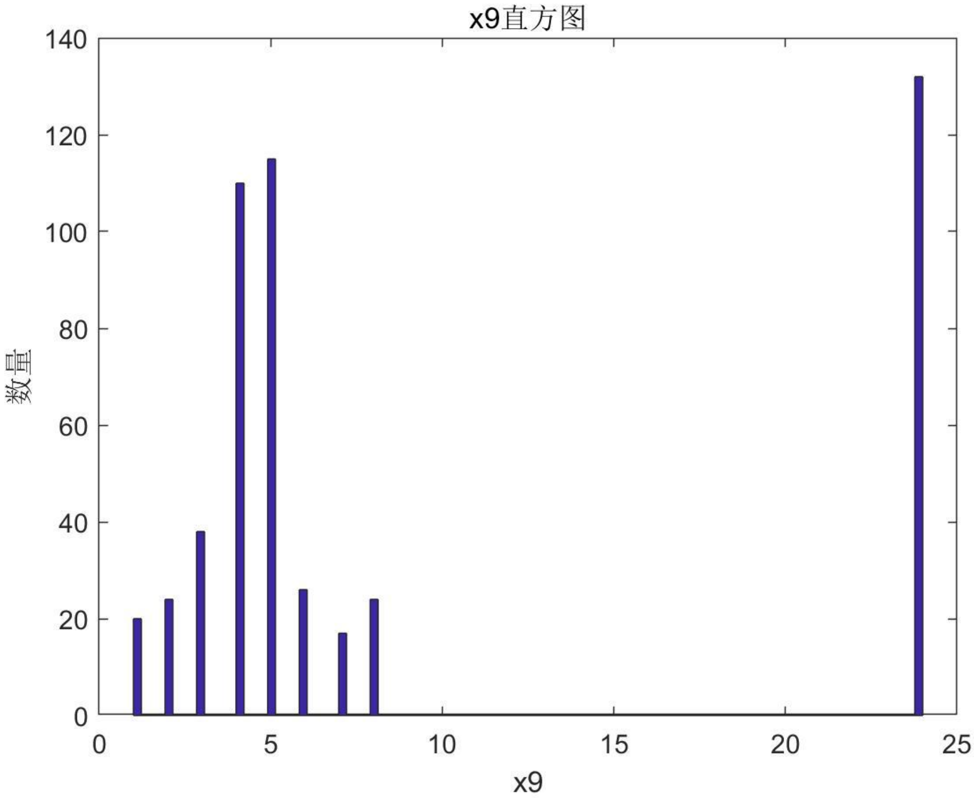


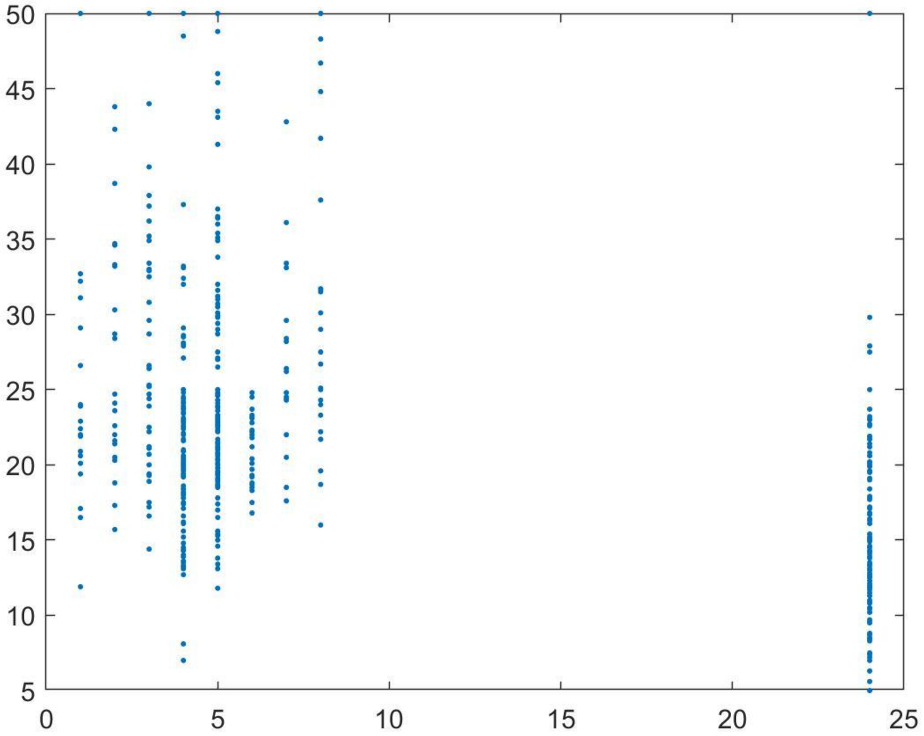
x8 is the weighted distance with five Boston labor-intensive areas. The house price scatter diagram shows that when the weighted distance is less than five, the house price is proportional to the distance. When the weighted distance with five Boston labor-intensive areas is about 5 to 9, the house price remains unchanged. When the weighted distance with five Boston labor-intensive areas is greater than 9, the house price is proportional to the weighted distance again. From the frequency histogram, it can be seen that most people choose the places with smaller weighted distances from the five Boston labor agglomerations, and a few people choose the places with larger weighted distances from the five Boston labor agglomerations. Considering the rich and poor population, most people need to work near the labor agglomeration, and a few rich people choose to stay away from the agglomeration, while the number of people selected is relatively average in the part with moderate distance from the labor agglomeration, so the house price remains unchanged.



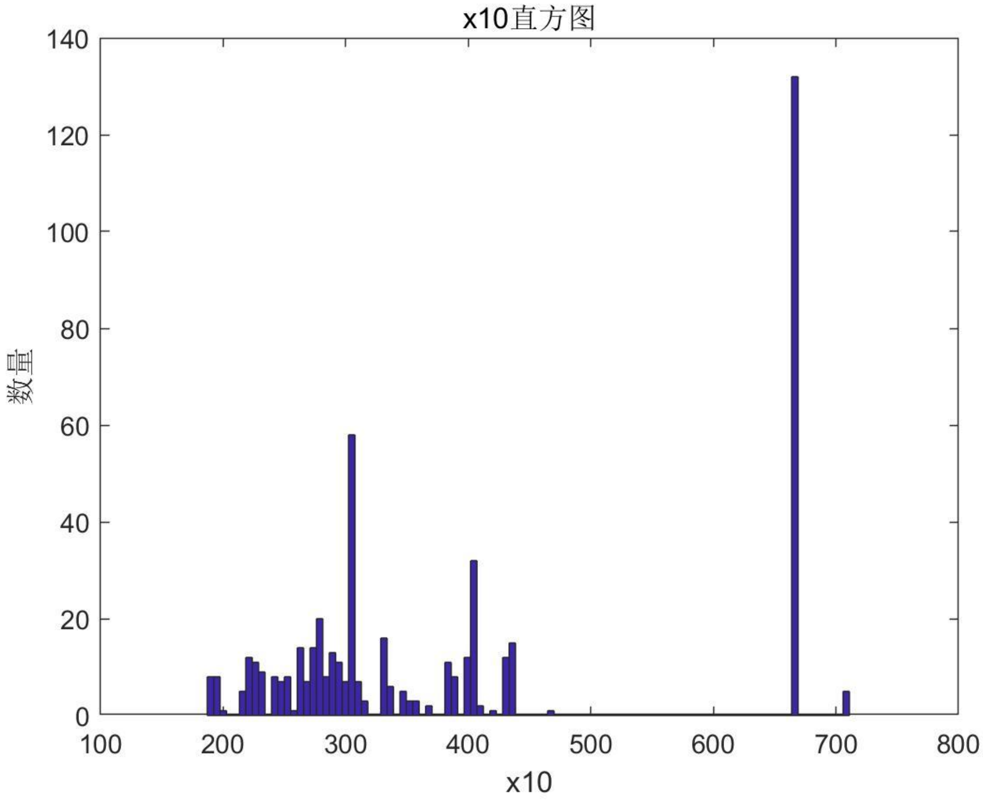


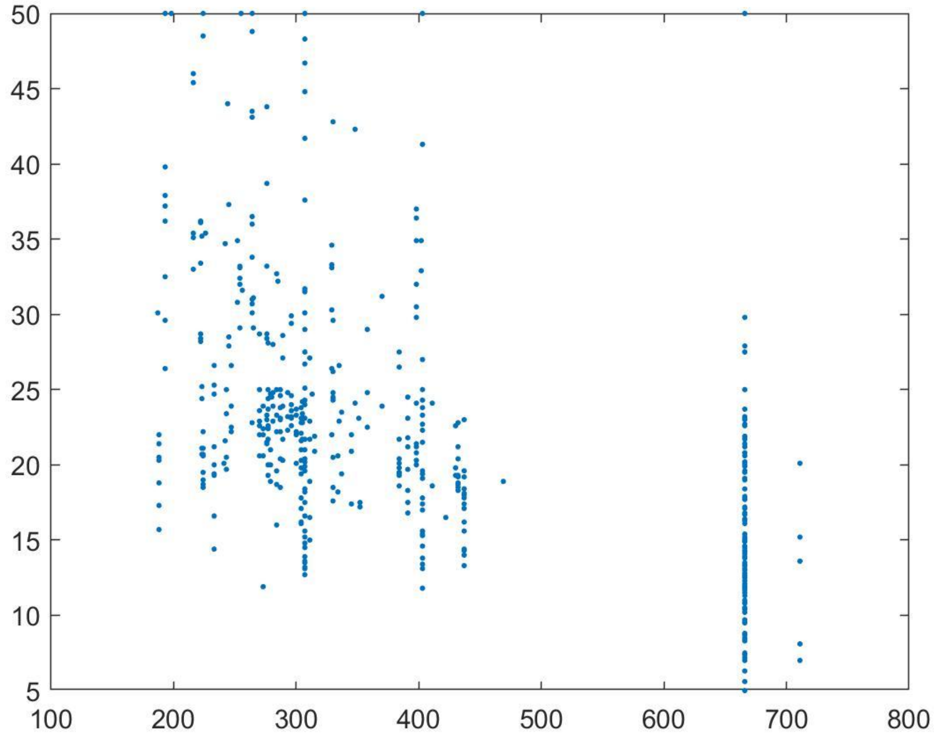
x9 is the proximity index to the radial highway. From the scatter diagram of housing prices, it can be seen that the housing prices are generally higher in places with smaller proximity index to the radial highway, and are generally lower in places with larger proximity index to the radial highway. When the proximity index to the radial highway is small, the housing prices are very poor, about 40 ; close to the radial highway index is small, housing prices range is small, about 30.



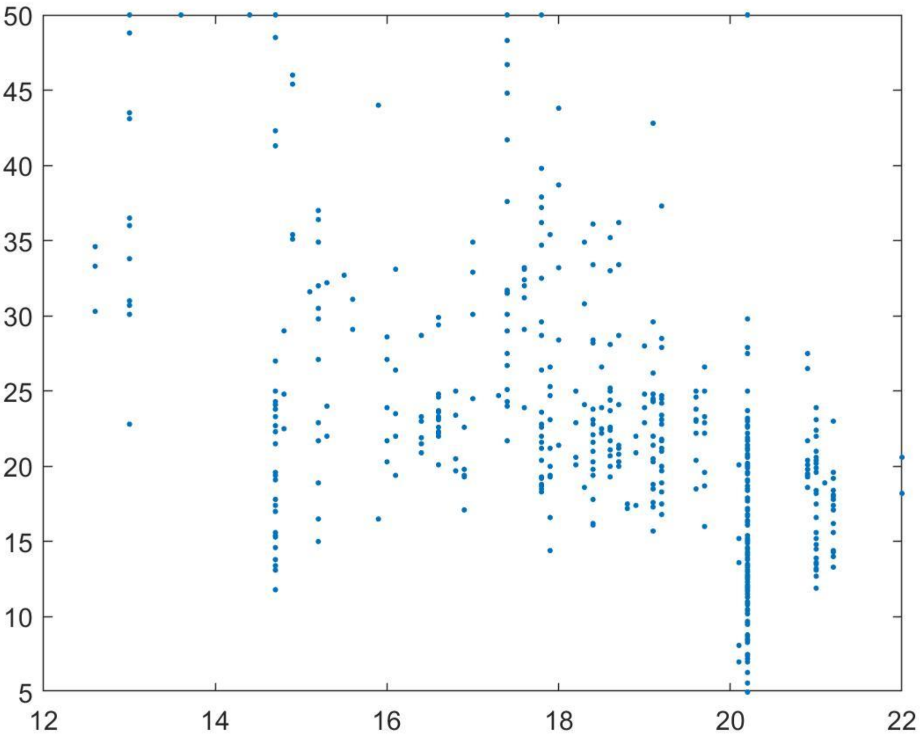


x10 is the total property tax of USD 10,000. It can be seen from the scatter plot of house price that the lower the total property tax of USD 10,000 is, the higher the house price is, and the house price is extremely poor, about 40. It can be seen from the frequency histogram that most people choose houses with low total property tax per 10,000 dollars, and a few people choose houses with high total property tax per 10,000 dollars.

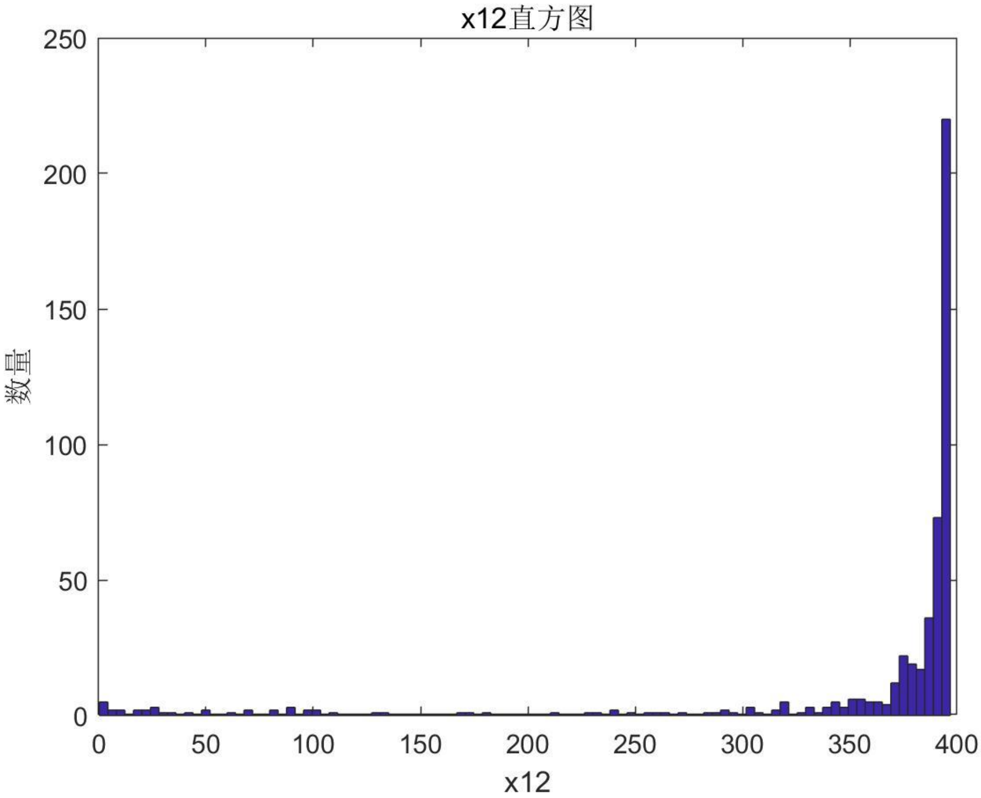


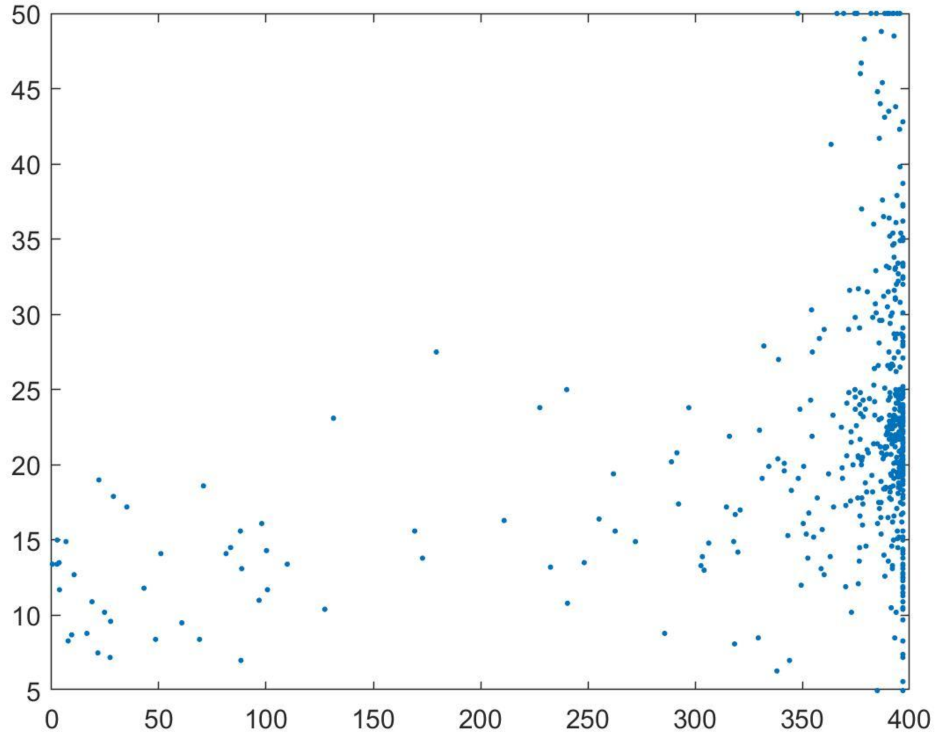


x11 is the proportion of students / teachers. It can be seen from the scatter plot of housing price that the higher the proportion of students / teachers is, the lower the housing price is, and the range of housing price is about 25.

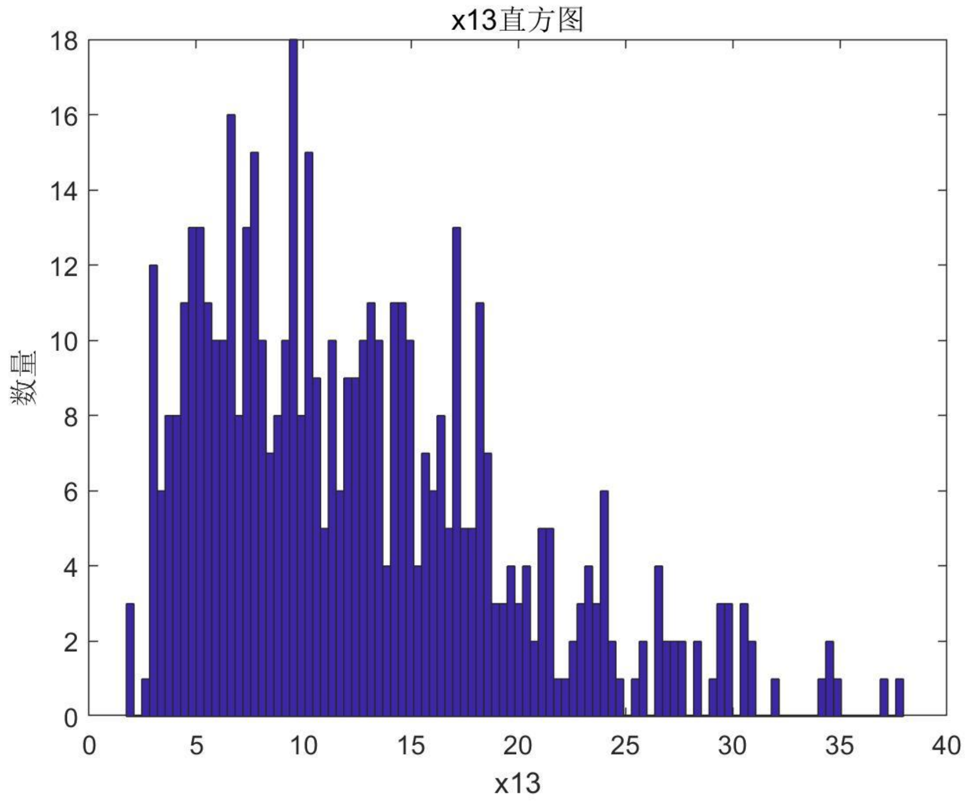


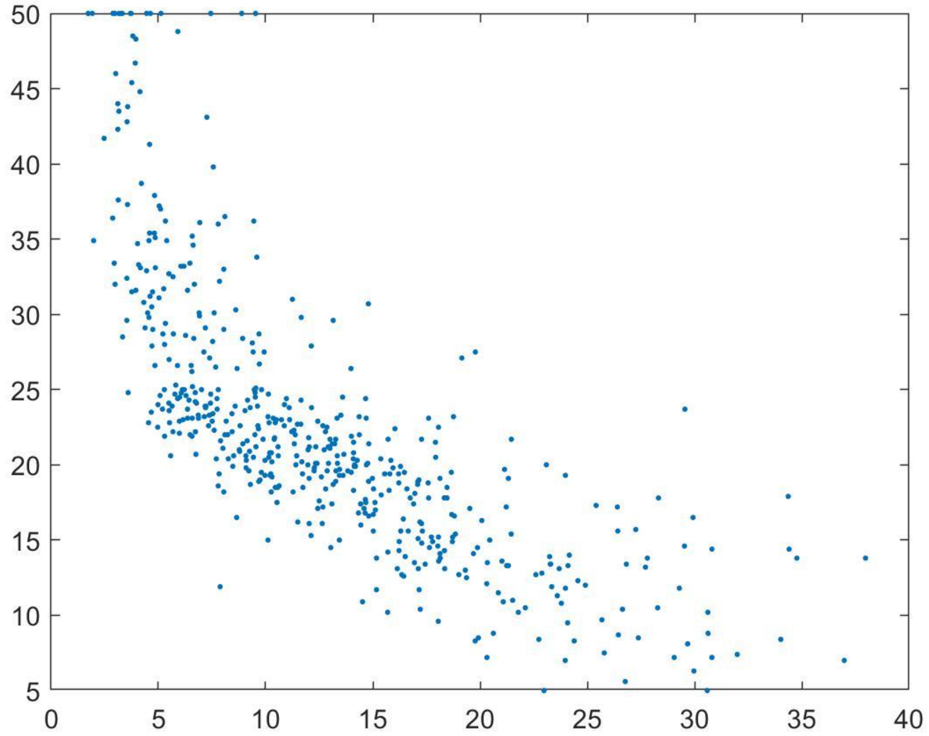
x12 For the proportion of African-Americans, as can be seen from the housing price scatter diagram, the smaller the B is, the smaller the proportion of African-Americans is, the greater the value is, and the higher the housing price is. It can be seen from the frequency histogram that most people choose places where the proportion of Afro-Americans is small. Considering the wealth gap between Afro-Americans and non-Afro-Americans, it is not difficult to find that B is inversely proportional to housing prices, and the independent variable is proportional to housing prices.





x13 is the proportion of the population with low social status. It can be seen from the scatter plot of house price that the larger the proportion of the population with low social status is, the lower the average house price is. And when the proportion of low social status population is less than 17 %, the housing price decline rate is faster, when the proportion of low social status population is about 17 % to 30 %, the housing price decline rate is slow, when the proportion of low social status population is higher than 30 %, the housing price decline rate is accelerated again.





3.2 Problem

3.2.1 Model Construction

Based on the statistical description and analysis model of data, we construct the model of housing price and independent variables. First we made two assumptions :

1.Suppose variables are normal distribution or elementary function combination distribution.

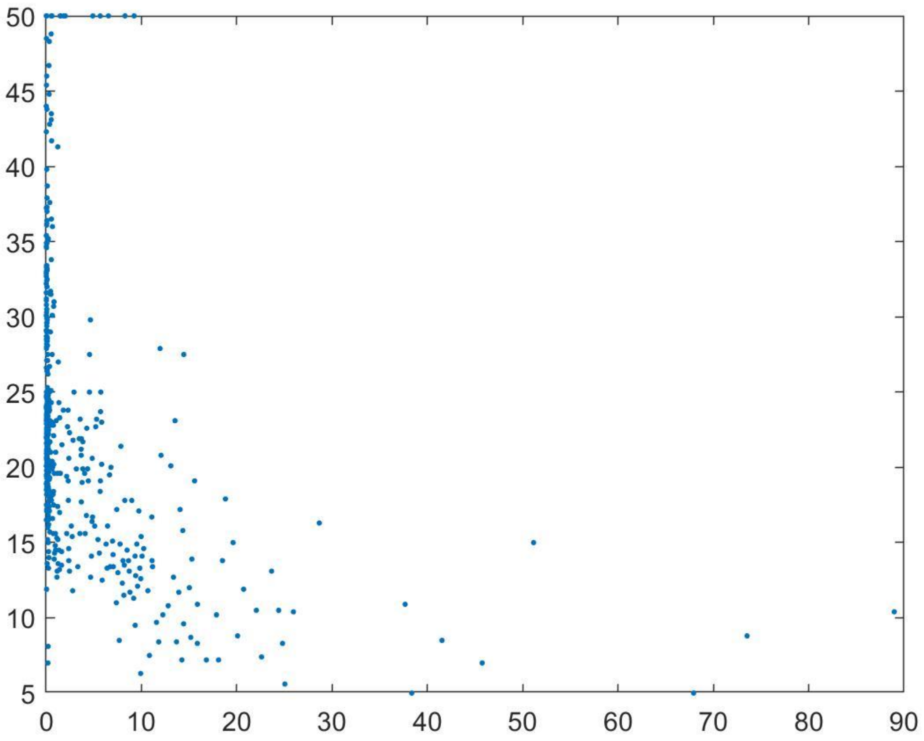
2.Assume that each variable can affect the results independently.

Based on these two assumptions, we build the model.

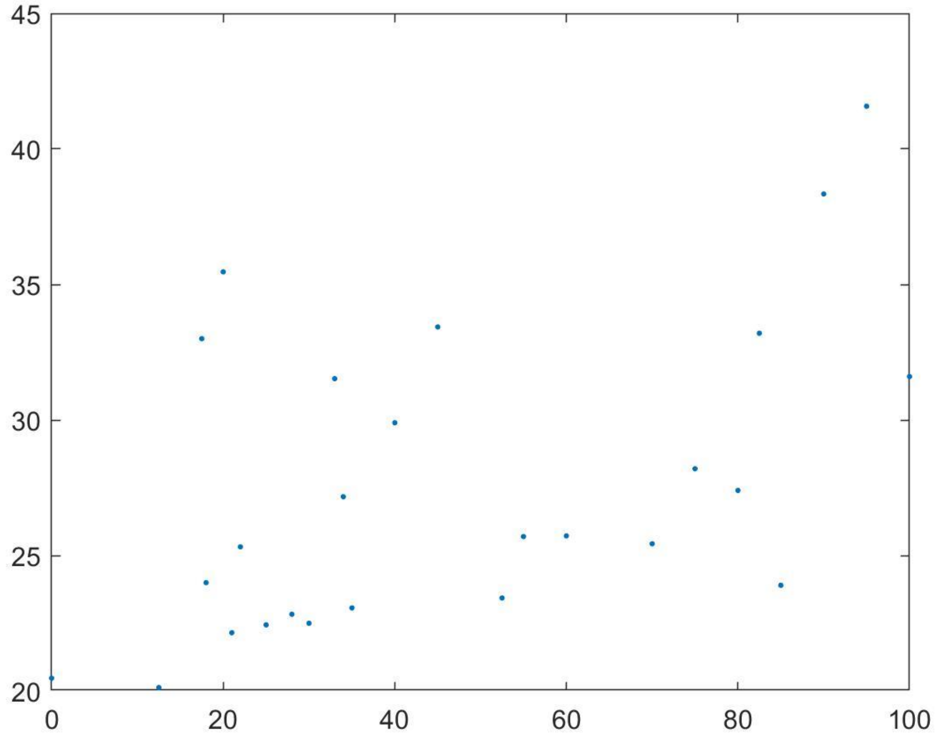
3.2.2 Data Processing

First of all, we classify different independent variables and housing prices, respectively take the average value of housing prices when the independent variable is a certain value, and draw the scatter plot of the average value of housing prices. According to the fitting principle, we fit the function curve.

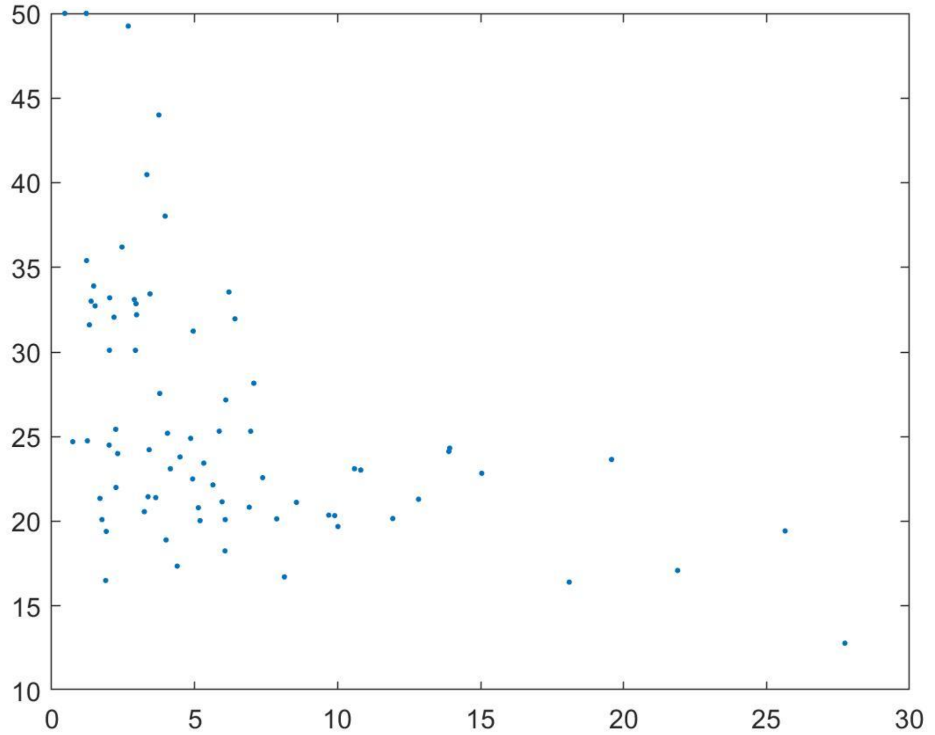
As shown in the figure, it is the scatter plot of the average housing price ;



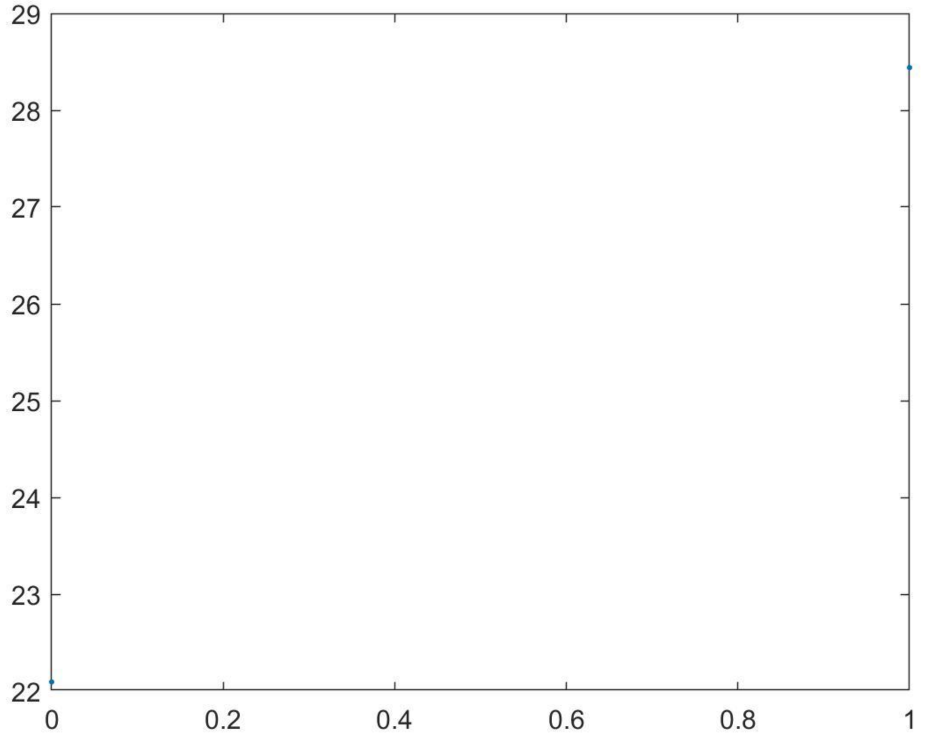
X1： Per capita crime rate



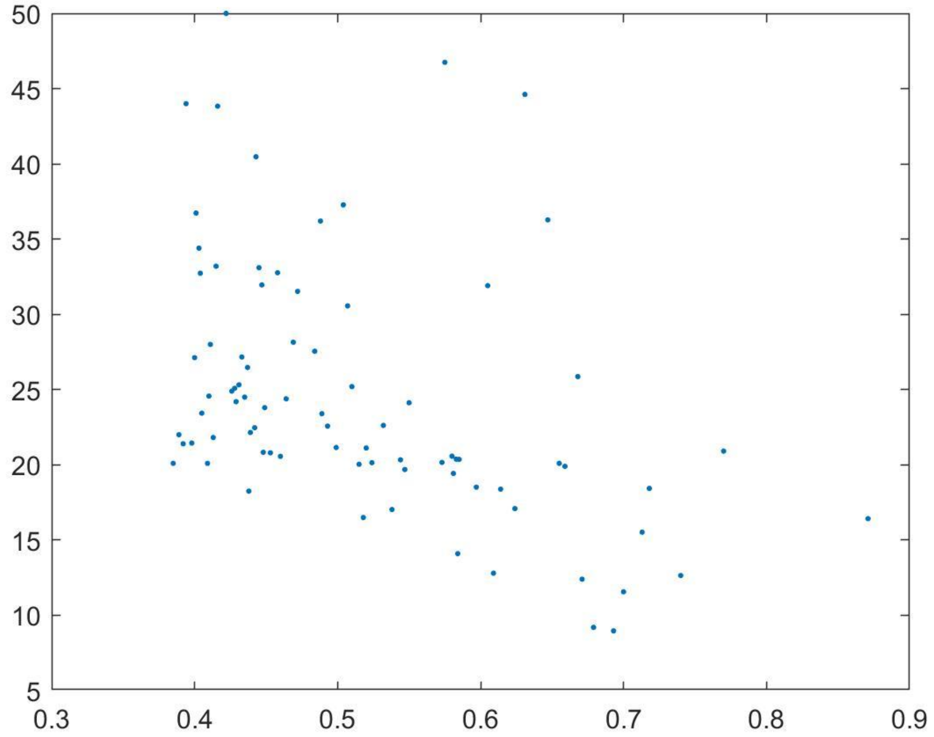
X2： Proportion of large residential areas



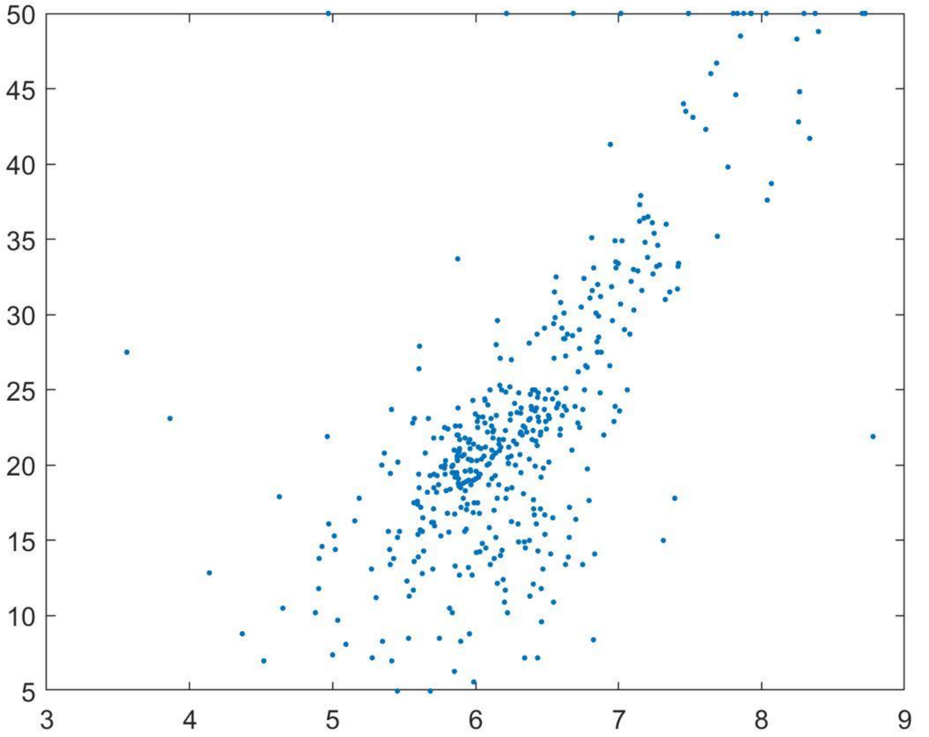
X3： Proportion of non-retail business land (acres)



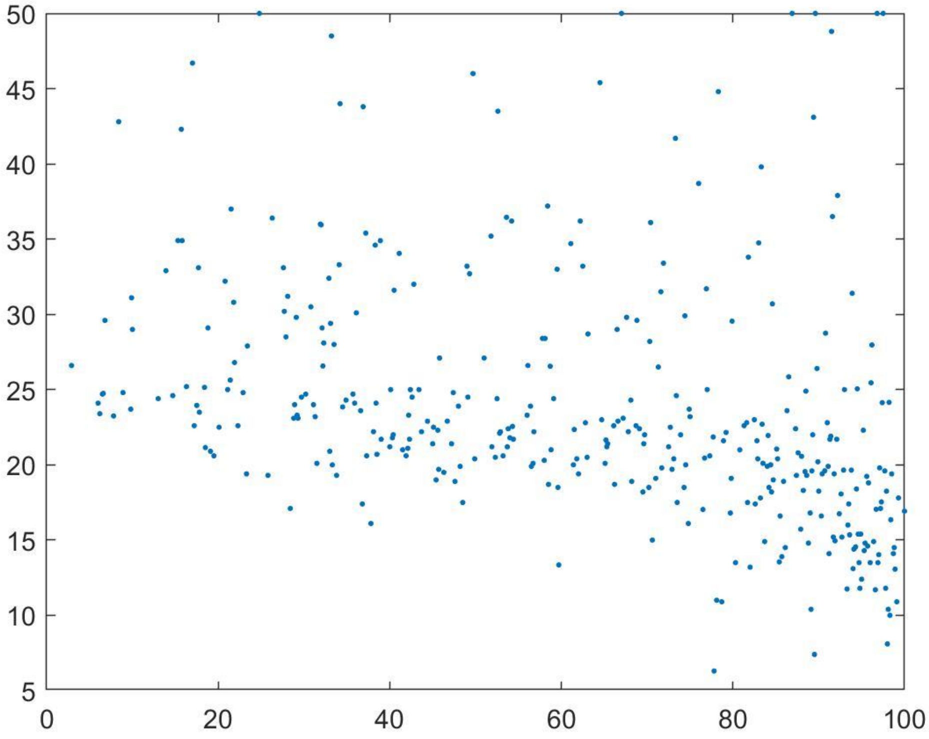
X4： Charles River virtual variables (if close to the river with 1 ; otherwise expressed as 0)



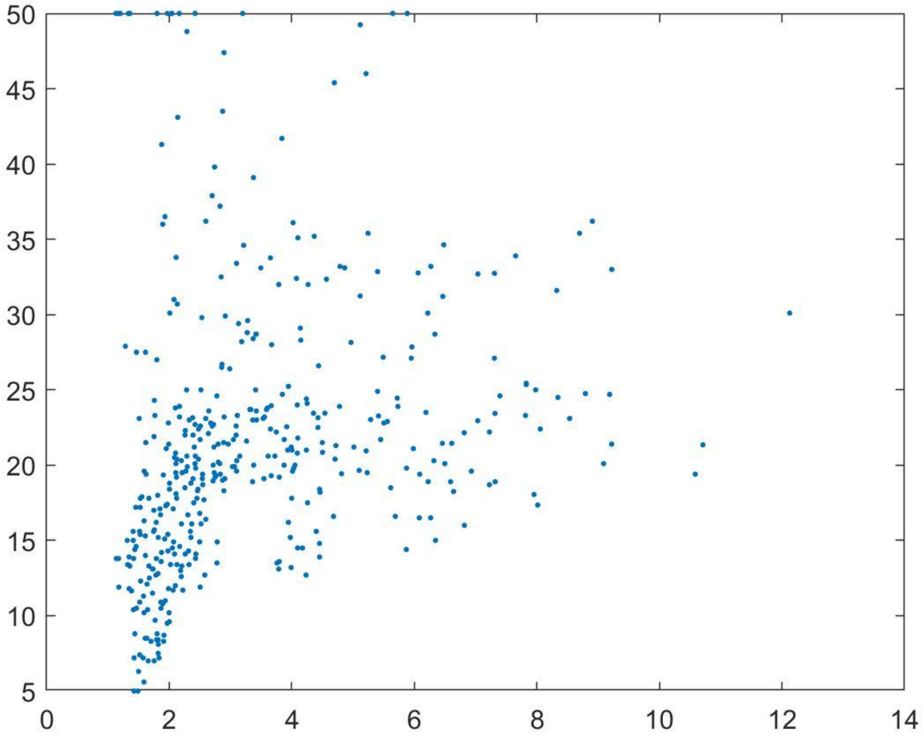
X5： concentration of nitrogen oxides



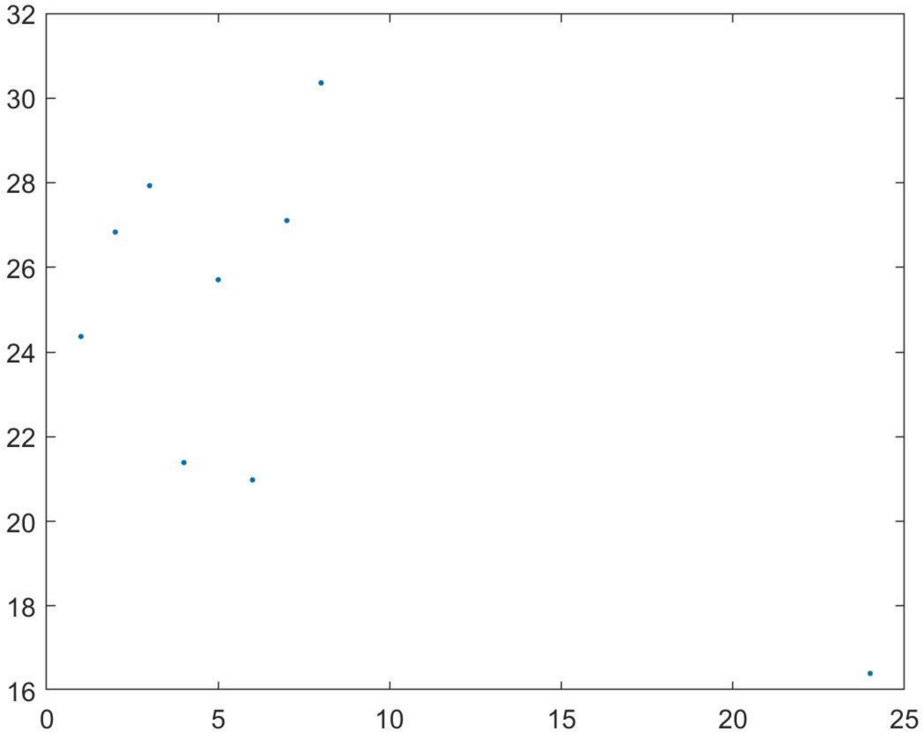
X6： Average number of rooms per household



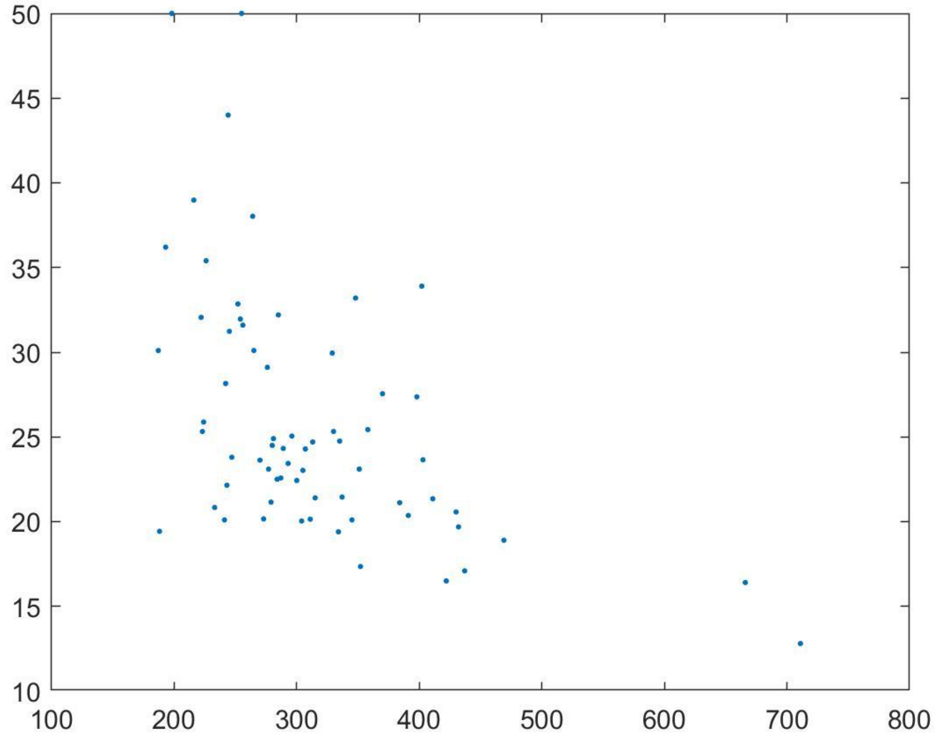
X7： Proportion of all homeowners built before 1940



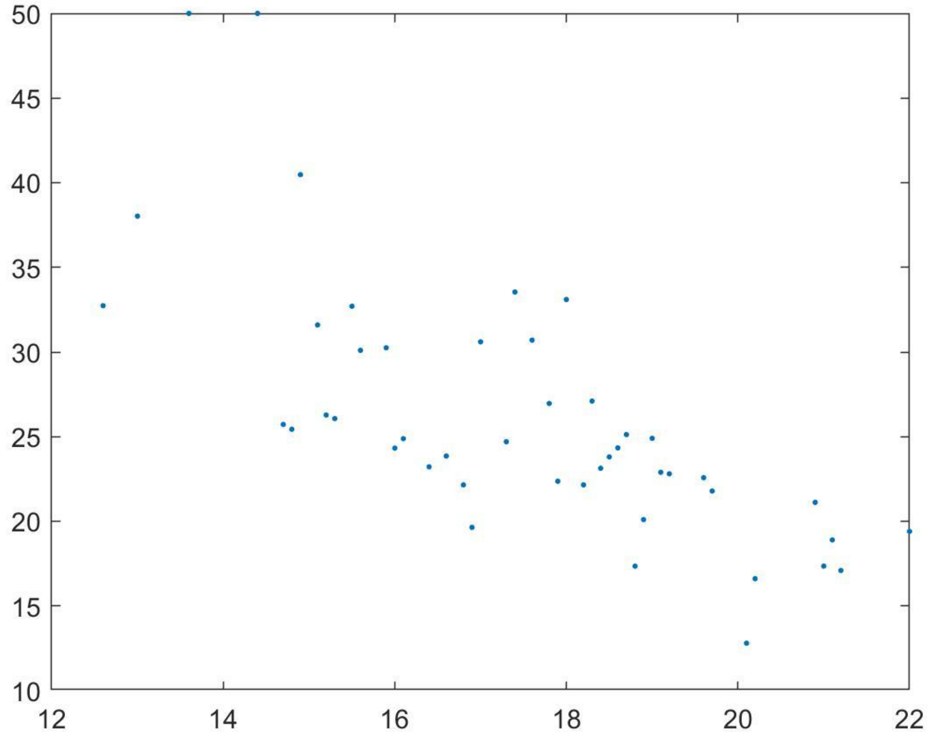
X8： Weighted distance from five Boston labour concentration areas



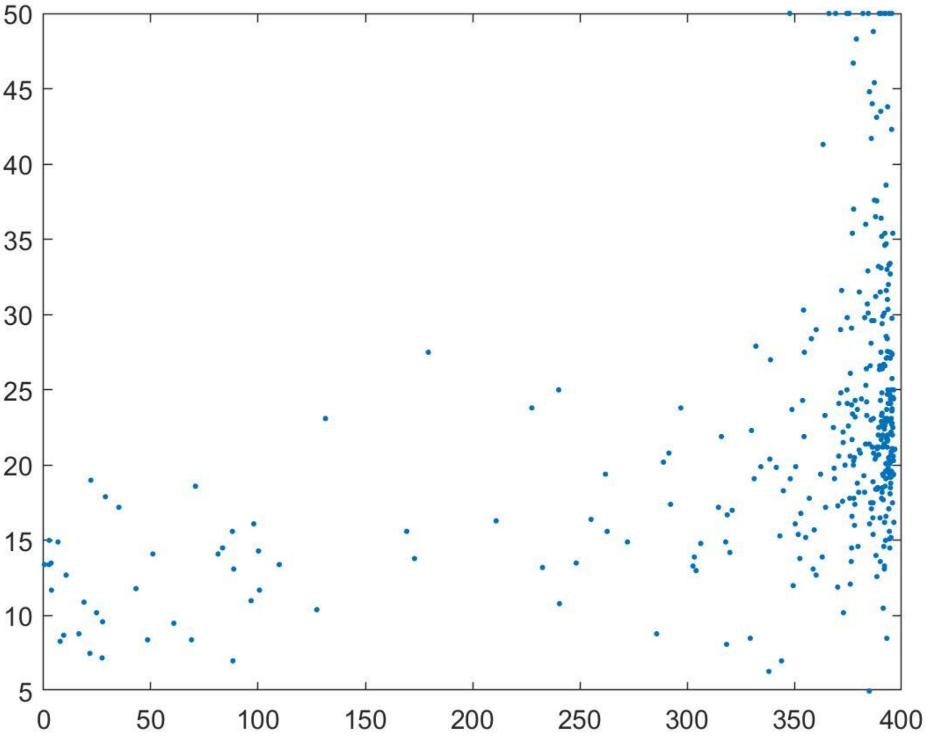
X9： Closeness index to radial highway



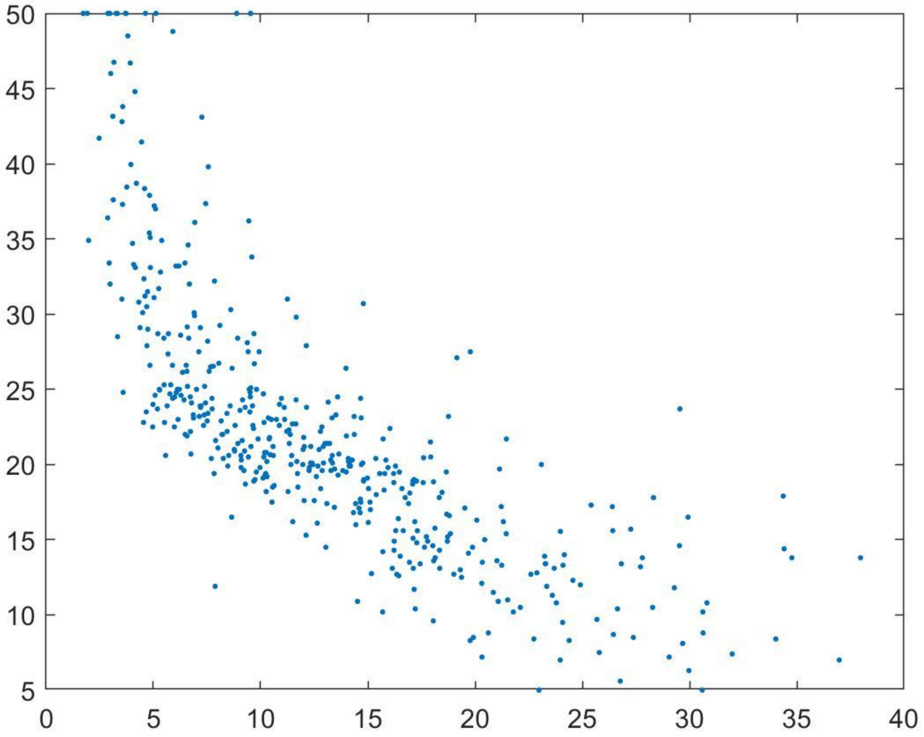
X10：Full value property tax per 10,000 dollars



X11：Student / teacher ratio



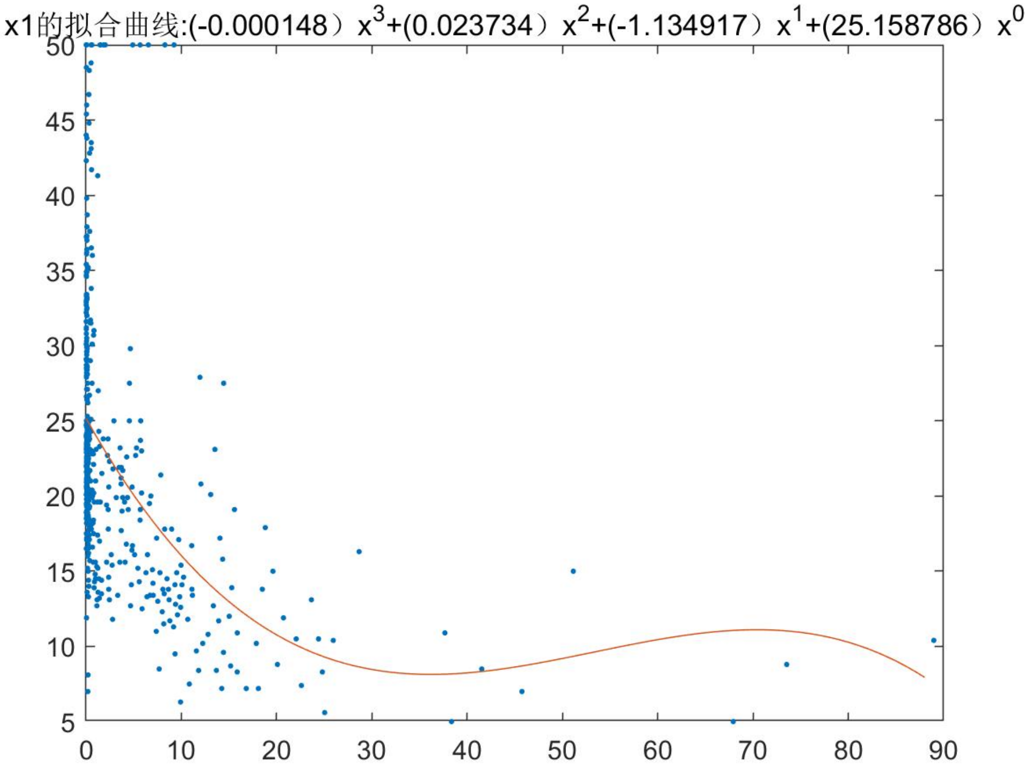
X12：$1000(B-0.63)^2 I(B<0.63)$，Proportion of African Americans

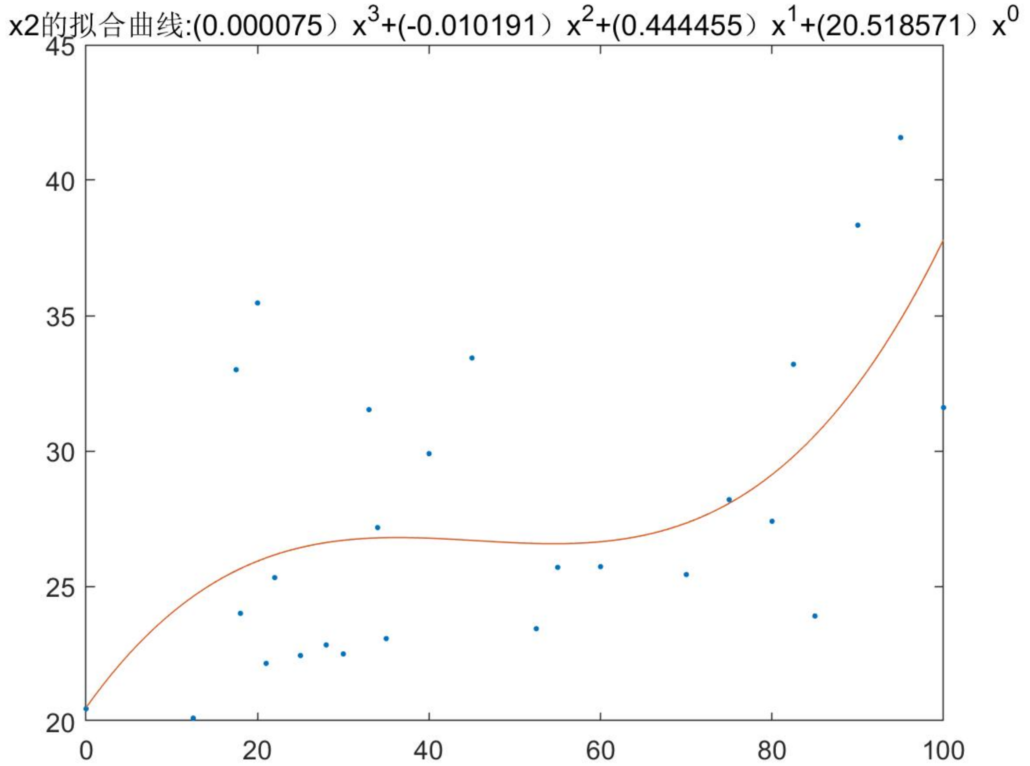


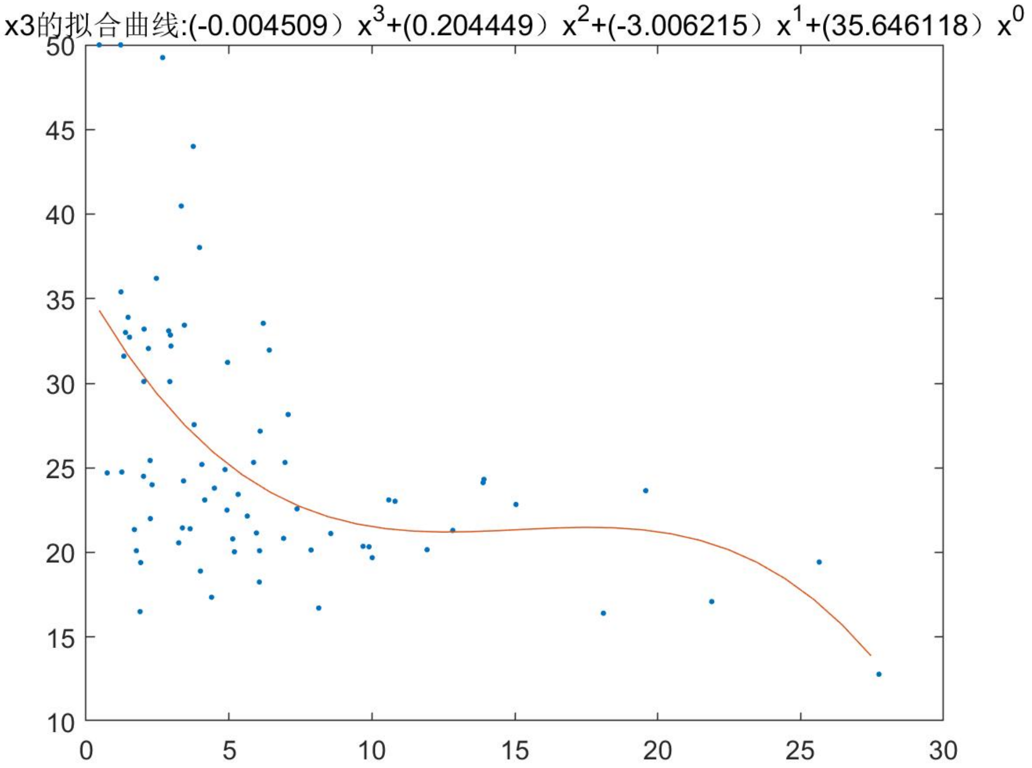
X13：Proportion of low social status population（%）

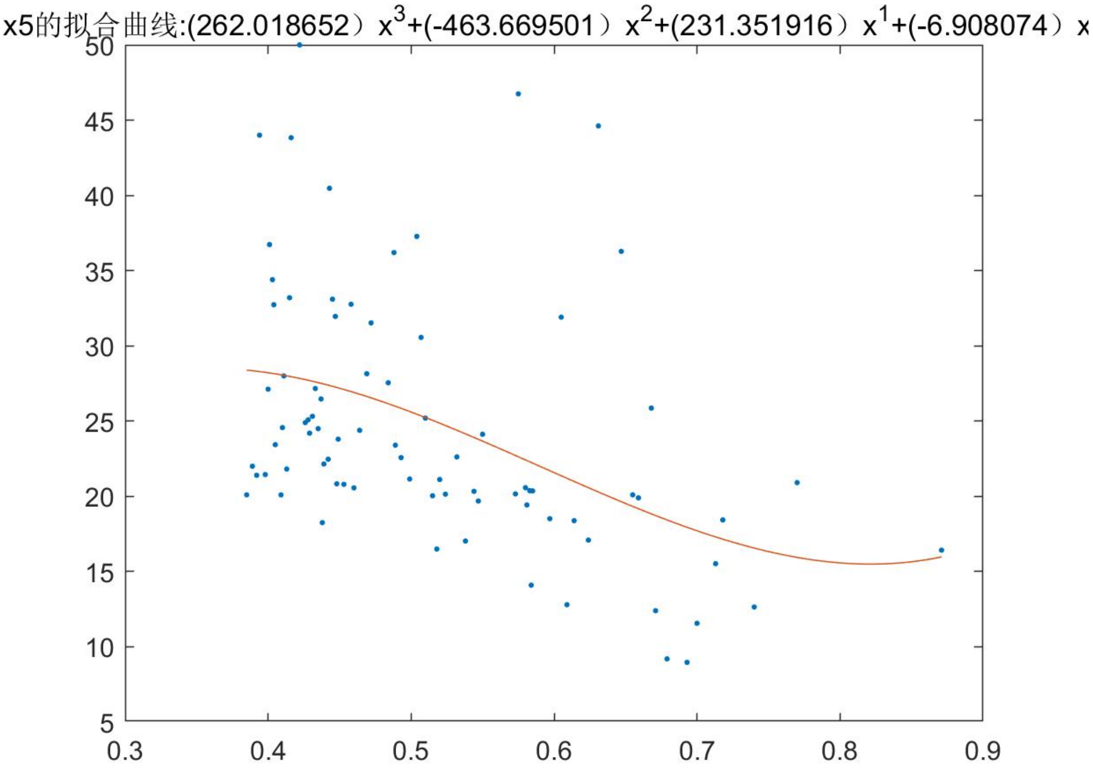
3.2.3 Model Prediction

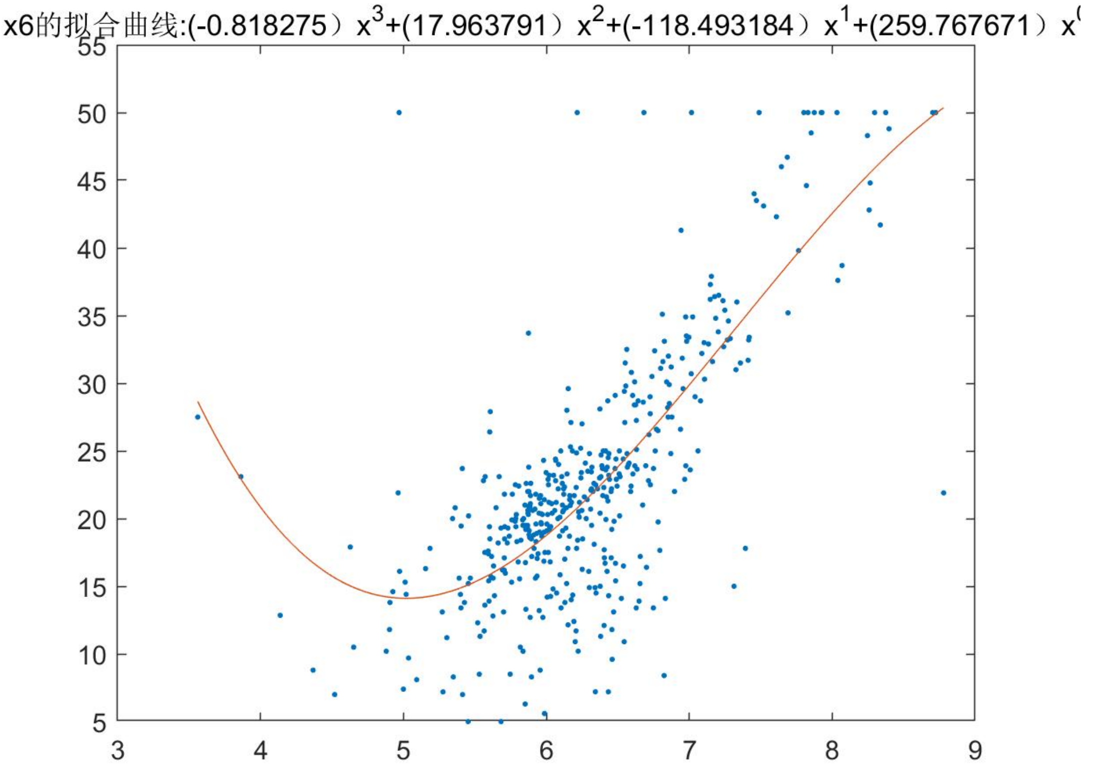
Through the scatter plot of average house price shown in the above figure, we construct the relationship curve between house price and independent variables, thus fitting the prediction model of house price, as shown in the following figure:



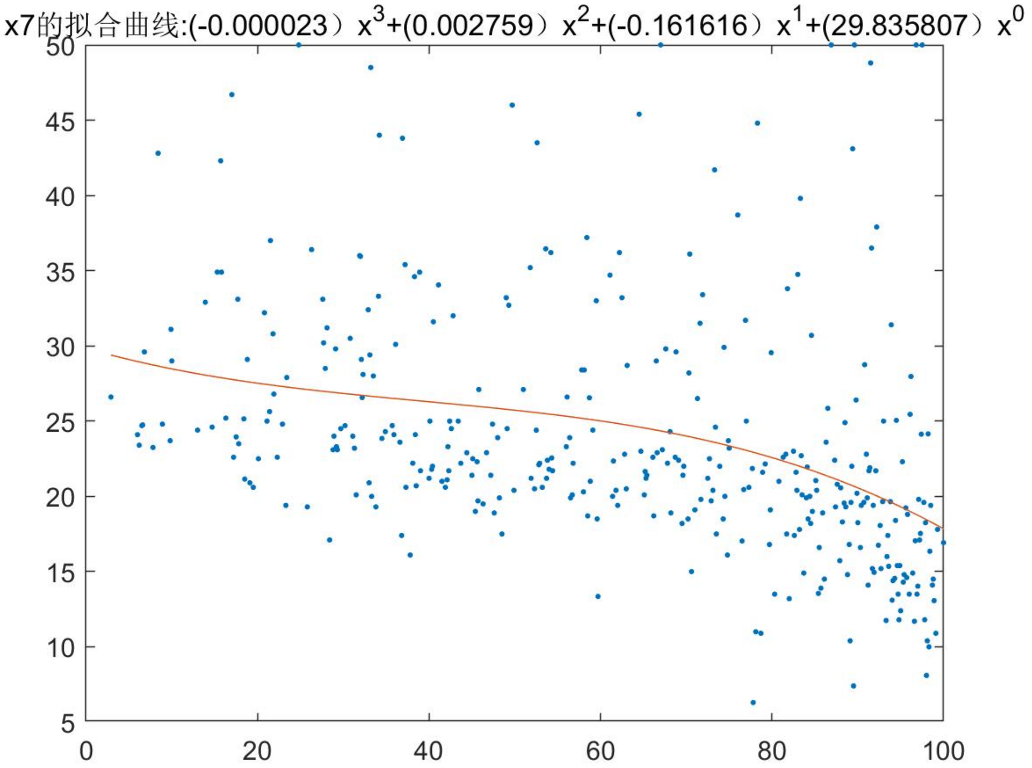



x3的拟合曲线:(-0.004509）x^3+(0.204449）x^2+(-3.006215）x^1+(35.646118）x^0
50
45
40
35
30
25
20
15
10
0
5
10
15
20
25
30

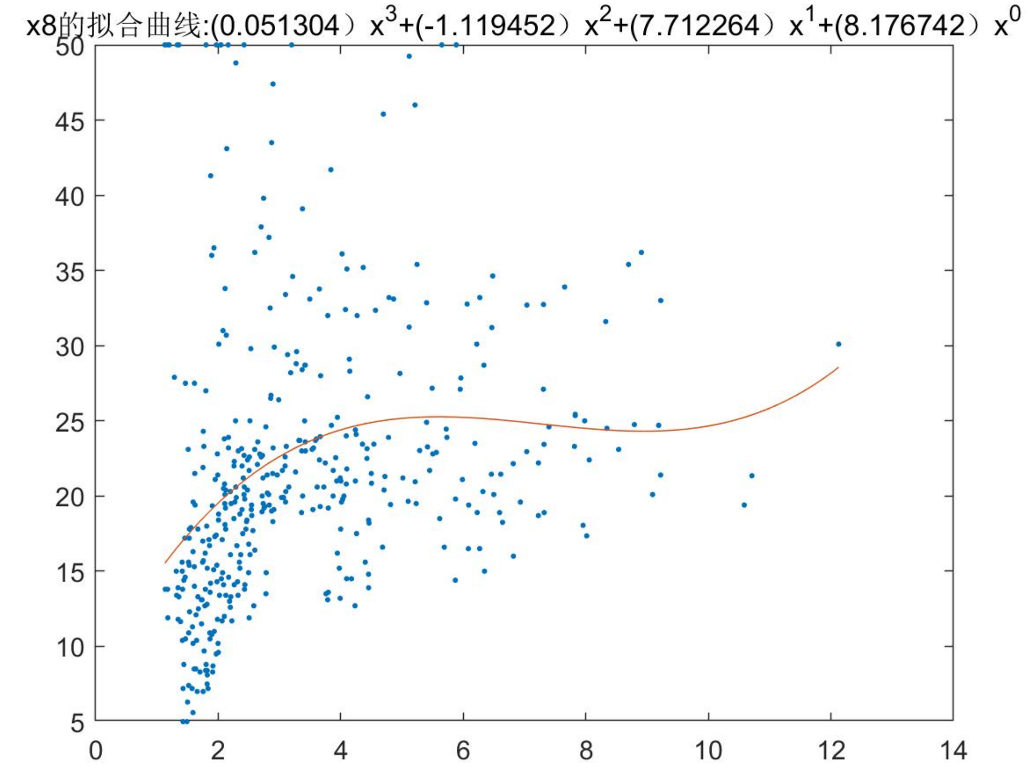

x5的拟合曲线:(262.018652）x^3+(-463.669501）x^2+(231.351916）x^1+(-6.908074）x
50
45
40
35
30
25
20
15
10
5
0.3
0.4
0.5
0.6
0.7
0.8
0.9

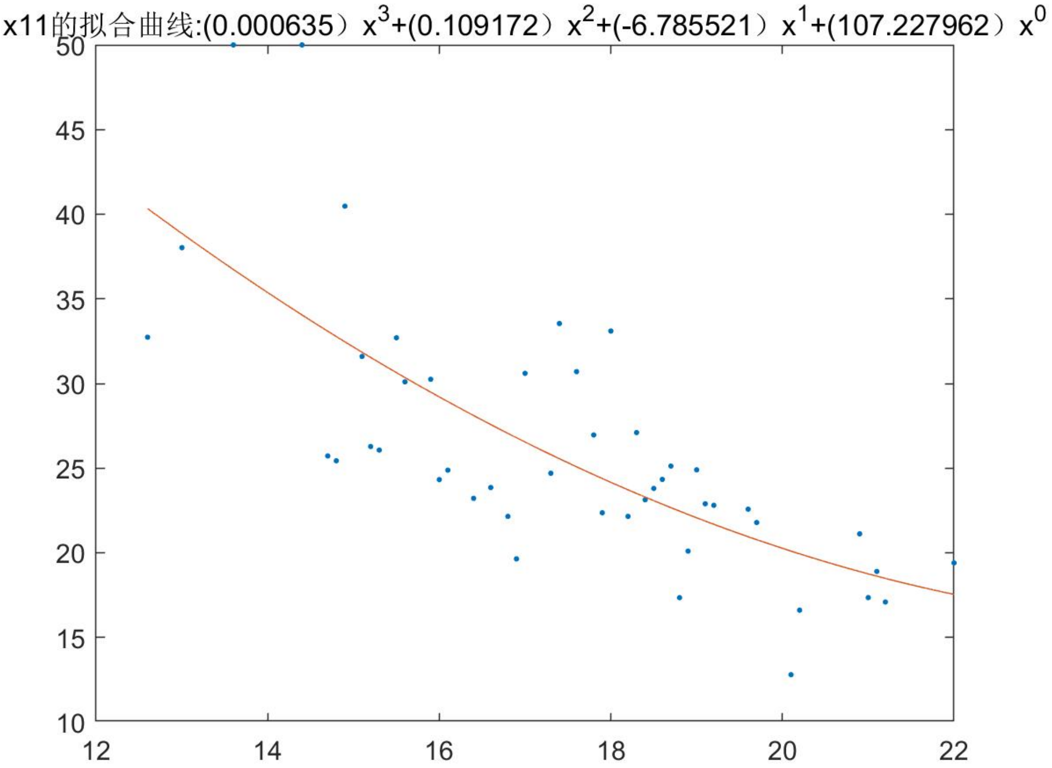


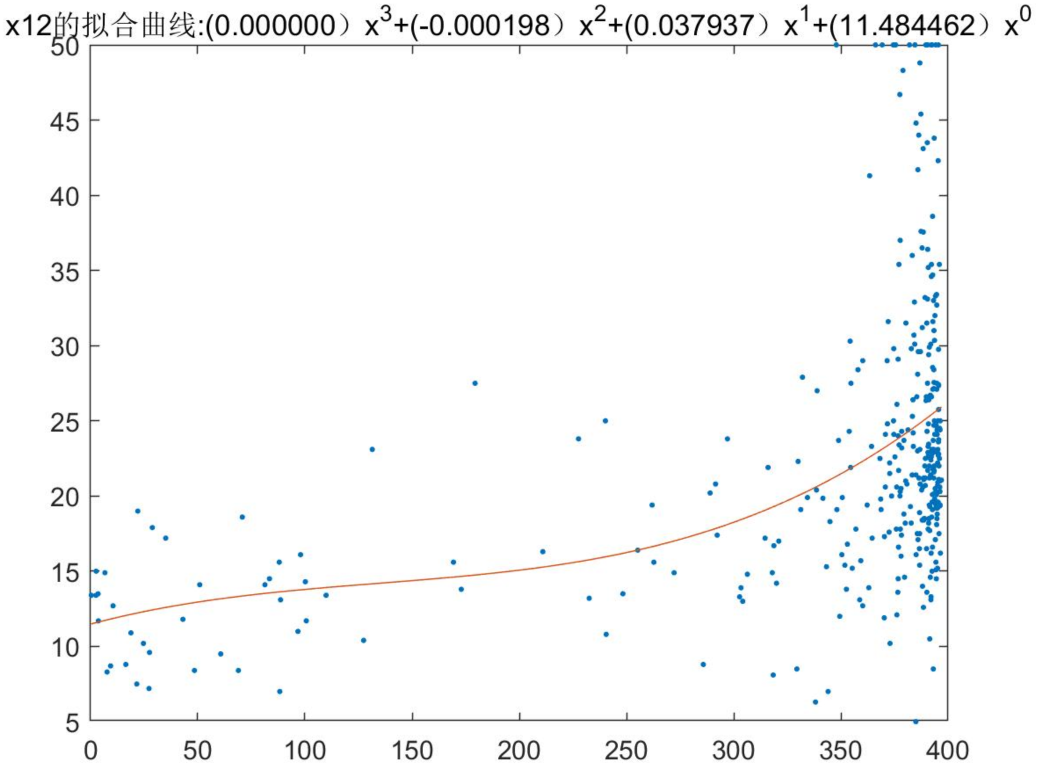
x6的拟合曲线:$(-0.818275）x^{3}+(17.963791）x^{2}+(-118.493184）x^{1}+(259.767671）x$

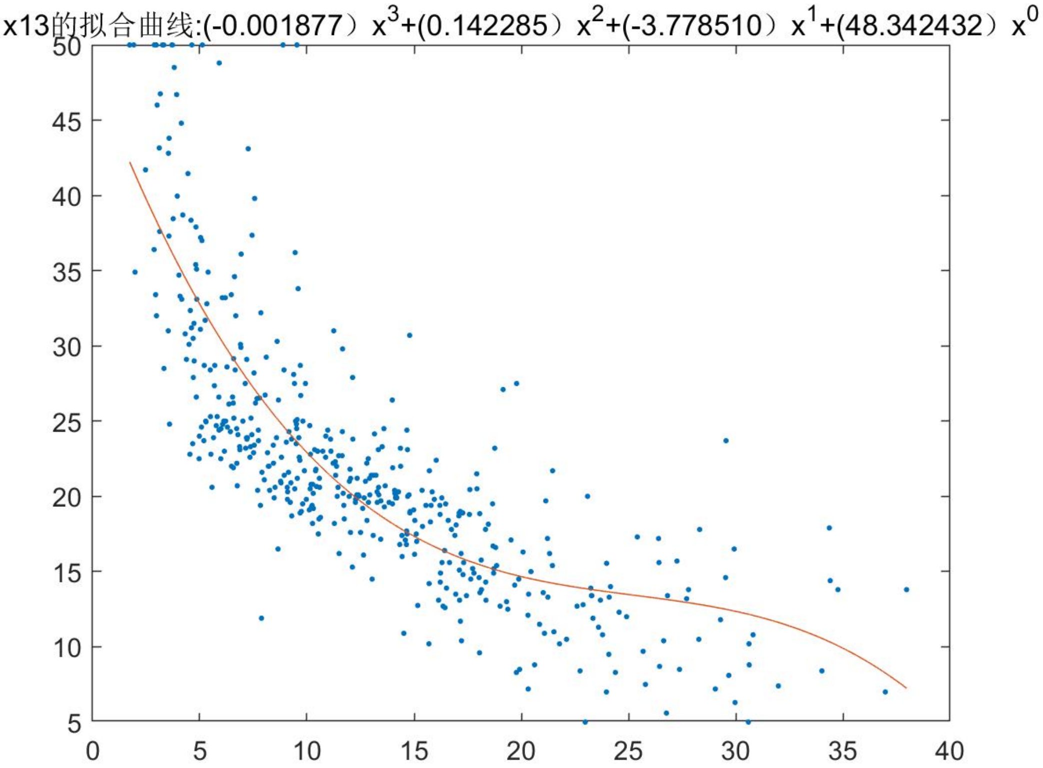


x7的拟合曲线:$(-0.000023）x^{3}+(0.002759）x^{2}+(-0.161616）x^{1}+(29.835807）x^{0}$


x8的拟合曲线:(0.051304）x^3+(-1.119452）x^2+(7.712264）x^1+(8.176742）x^0
50
45
40
35
30
25
20
15
10
5
0
2
4
6
8
10
12
14


x11的拟合曲线:(0.000635）x^3+(0.109172）x^2+(-6.785521）x^1+(107.227962）x^0
50
45
40
35
30
25
20
15
10
12
14
16
18
20
22





3.3 Problem

3.3.1 Conditions Analysis

In view of the requirements of livability, we believe that the house in Boston can be divided into thirteen grades, each grade corresponds to meet the livability requirements of an independent variable. We define the average value of each variable as the standard of whether to reach the level of livability, and analyze the probability of each variable reaching the average value through statistical data, so as to approximate the number of houses in Boston.

3.3.2 Problem Solving

For x1, the mean value is 3.6135, and the probability of meeting the livability is 25.30 %. For x2, the mean value is 11.3636, and the probability of meeting the livability is 26.48 %. For x3, the average value is 11.1368, the probability of meeting the livability is 26.48 % ; for x4, the mean value is 0.0692, and the probability to meet the livability is 6.92 %. For x5, the mean value is 0.5547, and the probability of meeting the livability is 41.30 %. For x6, the mean value is 6.2846, and the probability of meeting the livability is 45.06 %. For x7, the average value is 68.5749, the probability of meeting the livability is 57.91 % ; for x8, the mean value is 3.7950, and the probability of meeting the livability is 41.11 %. For x9, the mean value is 9.5494, and the probability of meeting the livability is 26.09 %. For x10, the mean value is 408.2372, and the probability of meeting the livability is 33.20 %. For x11, the average value is 18.4555, the probability of meeting the livability is 57.71 % ; for x12, the average value is 356.6740, the probability of meeting the livability is 81.82 % ; for x13, the mean value is 12.6531, and the probability of meeting the livability is 44.07 %.

3.4 Problem

In this report, we classify people by income level, which defines the high-income middle-income and low-income groups.

3.4.1 Suggestions on Housing Purchase for High-income People

For the people who can accept high-price housing, the purchase choice of residential housing is very extensive. From the results of this data collection, the characteristics of high-price residential areas are generally low per capita crime rate, that is, the security and stability of the community are better. Therefore, this point does not need to be considered too much. The proportion of large-scale residential areas is small, so the per capita housing area is more intensified, and the non-retail commercial land is larger. The biggest problem is that the concentration of nitrogen oxides is larger, which is common with the problem of any residential area. So the suggestion for high-income people is to buy more environmentally friendly products to reduce the concentration of nitrogen oxides.

3.4.2 Recommendations for Middle-income Groups

For general-income people, the data suggest that middle-income people should pay more attention to housing comfort when purchasing property, and should appropriately increase the proximity index to radial roads and the Charles River dummy variable. In addition, the nitrogen oxides content should be reduced as much as possible to increase the health status of residents. The proportion of students / teachers should be appropriately increased in families with children. Make children in the family easier to receive education.

3.4.3 Suggestions for Low-income People

For low-income people, the data suggests that low-income people should pay more attention to the per capita crime rate in the community when buying property, after all, safety is the most important. And as far as possible to ensure the proportion of children enrolled. That is, to increase the proportion of students / teachers as much as possible. After all, the more schools, the lower the rate of community crime

Reference

[1] Estimation and application of Lin Jianrui ' s spatial adaptive variable coefficient model '

[2] Estimation and application of Huang Suzhen ' s partial variable coefficient single exponential space regression model '

[3] Wang Shuxia ’ s simulation research and application of partial variable coefficient spatial autoregressive model ’

[4] Wang Yuanyuan 's Review of Boston House Price Research Based on Classical and Robust Methods '

[5] Nie Yufeng, Wang Zhenhai ' numerical analysis concise tutorial '