Analysis of the Influencing Factors and Predictions of Poston house Prices based on a Multiple Linear Regression Model

Liangyu Jiang *

ZiBo International Academy at Hi-tech Zone, Zibo, China

* Corresponding Author Email: lujinfeng@ziaedu.cn

Abstract. In this study, the classic Boston house price data set is selected for the analysis of house price correlation. According to the variables in the Boston housing price data set, the linear regression model of Boston housing prices is established by using Python software. The regression equation and regression coefficient were tested for significance, excluding the variable of p >=0.5, multiple linear regression was carried out, and the regression equation with good fitting was obtained. It is found that there are too many variables after multiple linear regression, which is difficult to analyze and predict, so the correlation analysis of variables is carried out. This paper gets the percentage of lower status population, pupil-teacher ratio and average number of rooms per dwelling with medv (The median quoted price for an owner-occupied home, $1,000 per unit) has a significant relationship. Finally, a linear regression equation is established for the independent variable whose correlation coefficient is greater than 0.5, and the housing price is predicted.

Keywords: Regression analysis, social problem, mathematical analysis.

1. Introduction

On a global scale, real estate is an important industry. The real estate industry has an important impact on economic growth, employment, and local fiscal revenue [1]. For example, the real estate industry is one of the pillar industries in the United States and has an important impact on the development of the American economy [2,3]. In Europe, the real estate industry is also an important industry, which has an important impact on the development of the European economy. As one of the pillar industries in China, the real estate industry has an important impact on economic growth, employment, local fiscal revenue, and other aspects. In 2020, the real estate industry and its industrial chain will account for 17% of China's GDP. The real estate industry has a strong driving effect on upstream and downstream related industries, a great impact on investment and consumption, and a systemic impact on economic and financial stability and risk prevention and control [4]. The change in housing prices has an important impact on residents' wealth, consumption, investment, and other aspects. This paper takes the Boston housing price data set as an example to analyze the relationship between the variables that may affect the housing price and the housing price, to help people make a better choice of housing. This paper uses multiple linear regression, variable correlation analysis, and other methods to analyze.

2. Analysis of factors influencing Poston house prices based on a multiple linear regression model

2.1. Introduction to Datasets

The Boston housing price dataset represents the median housing prices in suburban Boston during the mid-1970s. It encompasses 13 indicators, including crime rate and property tax rates within the parish at that time, to establish a correlation between these indicators and housing prices. This study is based on a linear regression model, which is a form of regression analysis that models the relationship between one or more independent variables and dependent variables using a least square function known as a linear regression equation. By employing visualization techniques and preliminary model training, researchers extract crucial features to obtain quantitative insights into the

relationship between housing prices and various factors. In future home purchases, individuals can primarily consider RM (average number of rooms per dwelling), LSTAT (percentage of lower status population), and PTRATIO (pupil-teacher ratio) as key reference points.

2.2. Model Introduction and Variable Setting

2.2.1. Introduction to the variable name

According to the analysis of the Boston House Price Dataset, there may be 13 factors affecting the response variable MEDV (The median quoted price for an owner-occupied home, $1,000 per unit). The following is the introduction of each attribute, as shown in Table 1.

Table 1. Introduction of related variables

Variable abbreviation	Variable meaning
RIM	Urban per capita crime rate
ZN	Proportion of residential land with a floor area of more than 25,000 square feet
INDUS	Proportion of non-retail business in each town
CHAS	Charles River dummy variable
NOX	Nitric oxide concentration (per 10 million parts)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built before 1940
DIS	Weighted distance to five job centers in Boston
RAD	Accessibility index of radial expressway
TAX	Full property tax rate per $10,000
PTRATIO	Student to teacher ratio in town
B	1000 (Bk-0.63) ^2 where Bk is the percentage of blacks in the town
LSTAT	Population status decreased by %
MEDV	The median quoted price for an owner-occupied home, $1,000 per unit

2.2.2. Introduction of the multiple linear regression models

The multiple linear regression model refers to a linear regression model containing multiple explanatory variables, which is used to explain the linear relationship between the explained variable and several other explanatory variables. Its mathematical model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \varepsilon \quad (1)$$

The formula above represents a p-element linear regression model, and it is evident that there is a total of p explanatory variables. The representation of the change in the explained variable y encompasses two components: firstly, the linear alteration in y induced by the variation in p explanatory variables x.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p \quad (2)$$

The second part is to explain why random variables cause y. Change the part that can be used ε partial substitution, which can be called random error, is the parameter in the formula. ε, $\beta_0, \beta_1, \ldots \beta_p$ are all unknowns in the equation, which can be expressed as partial regression constant and regression constant, then the regression equation of the multiple linear regression model is as follows:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \varepsilon \quad (3)$$

3. Problem solution method and knowledge support

In this example, this paper uses Python software to solve the problem. Please refer to Reference for some relevant code [5].

3.1. Regression Analysis

Due to the multivariate nature of the data, it is not feasible to depict the scatter plot using monadic regression analysis. To investigate the relationship between each attribute and response variable, this paper initially established a linear regression model and evaluate its validity. The obtained results including regression coefficients and P-values are presented in Table 2.

Table 2. Coefficients of the regression equation and their p values

	Coefficients	t Stat	P-value
CRIM	-0.101	−3.287	0.001087
ZN	0.118	3.382	0.000778
INDUS	0.0153	0.334	0.738288
CHAS	0.0742	3.118	0.001925
NOX	-0.224	-4.651	4.25E−06
RM	0.291	9.116	<2e−16
AGE	0.00212	0.052	0.958229
DIS	-0.338	-7.398	6.01E−13
RAD	0.290	4.613	5.07E−06
TAX	-0.226	-3.28	0.001112
PTRATIO	-0.224	-7.283	1.31E−12
B	0.0924	3.467	0.000573
LSTAT	-0.407	-10.347	<2e−16

The statistic F=108.1, p<2.2e-16, and the significance levelα= 0.05 was given. If p <α therefore, the null hypothesis is rejected and the regression equation is considered significant. However, according to the above table, some regression coefficients are not significant. In this case, the standard deviation of the residual is 0.516, R^2=0.7406, the adjusted R^2 is 0.7338. The fitting effect is average.

3.1.1. Further analysis of the regression equation

Because some regression coefficients in the regression equation are not significant, the insignificant variables are eliminated. First of all, the variable with the largest P-value is eliminated for regression analysis, and then the remaining variable is eliminated for analysis, which is carried out successively until all the regression coefficients in the regression equation are significant. The results of the regression equation and stepwise regression selection variables are the same.

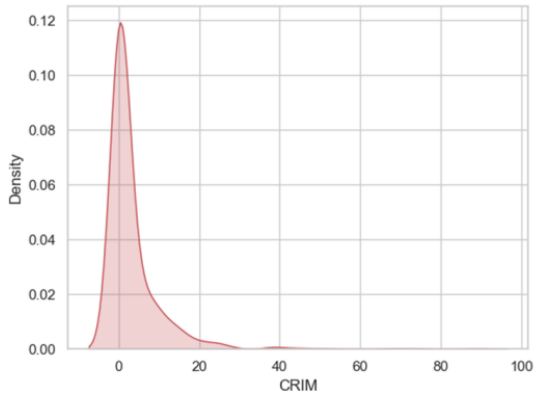
The regression equation established is as follows:

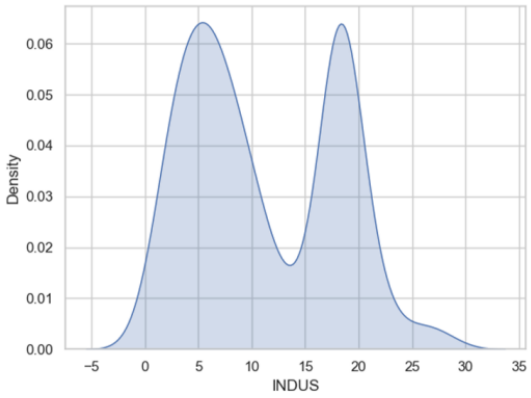
$$\hat{M} = -0.101\hat{CR} + 0.116\hat{Z} + 0.075\hat{CH} - 0.219\hat{N} + 0.290RM - 0.342\hat{D} + 0.284RA - 0.216\hat{T} - 0.223\hat{P} + 0.092\hat{B} - 0.406L$$

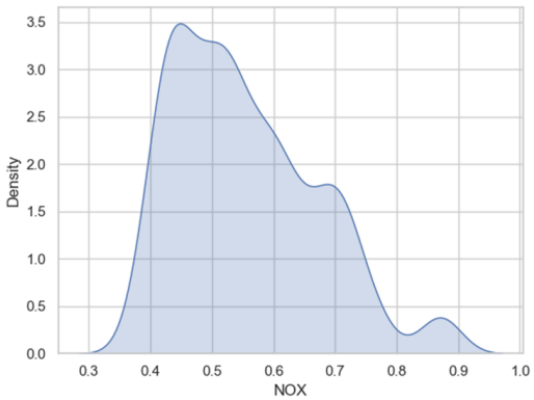
so that, the statistic F = 128.5, p < 2.2e – 16, the regression coefficients in the regression equation are significant, the adjusted R^2 = 0.7348 and each regression coefficient is significant. However, although the model obtained above passes the test, there are a large number of independent variables, and the purpose of dimensionality reduction is achieved by trying to use it. Therefore, this paper tries to carry out data correlation analysis to select more significantly correlated variables.

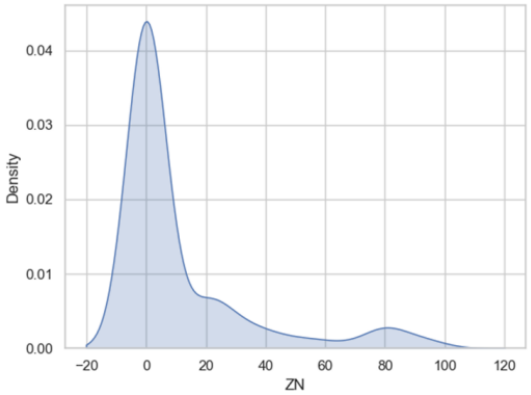
3.2. Data correlation analysis

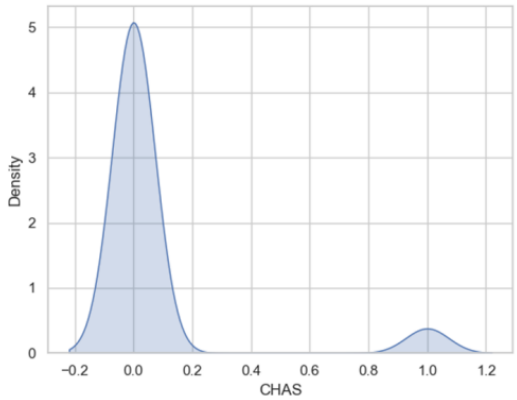
3.2.1. Multivariate study

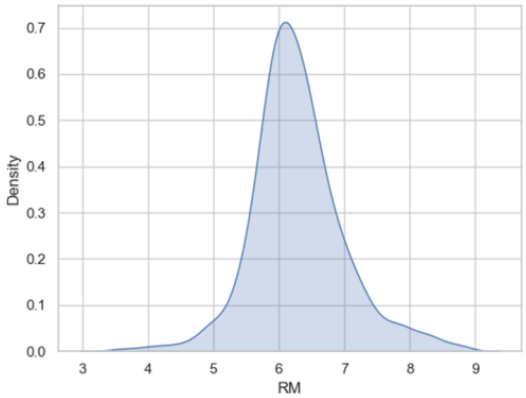


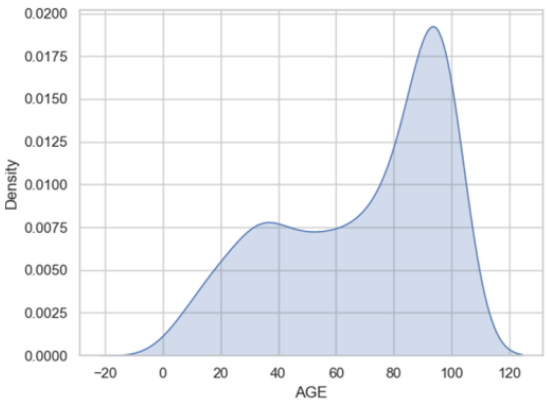


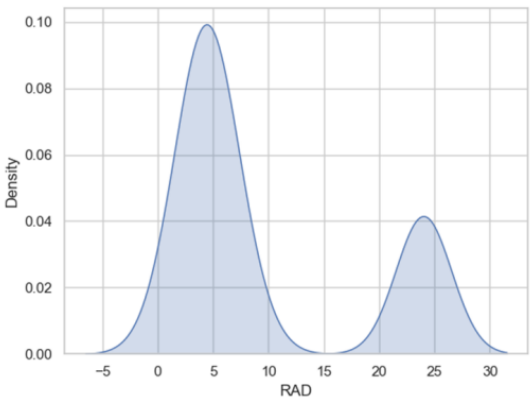












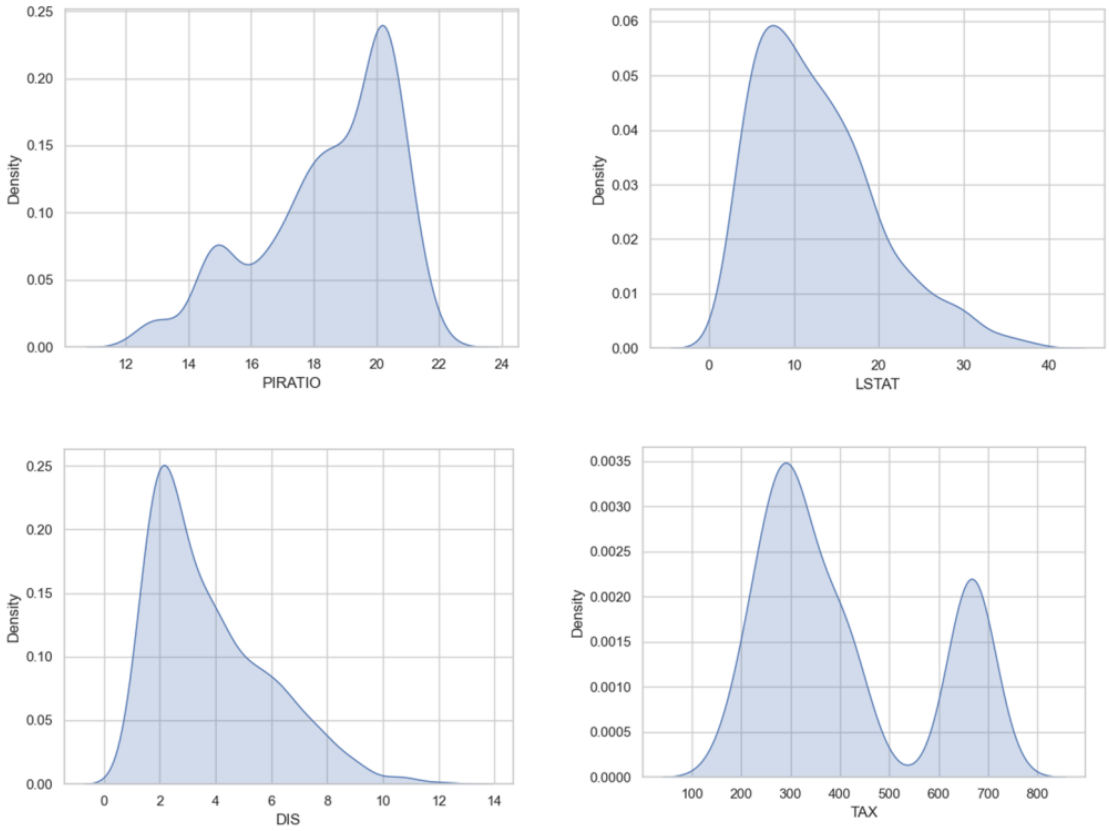


Figure 1. Statistical distribution of all features

As shown in figure 1, it is easy to kown that

NOX (concentration of nitric oxide), RM (average number of rooms in each residence), PTRATIO (ratio of teachers and students in each town), LSTAT (percentage of bottom population) are close to the distribution diagram of price

No outliers were detected from the feature distribution map, so no outliers were processed.

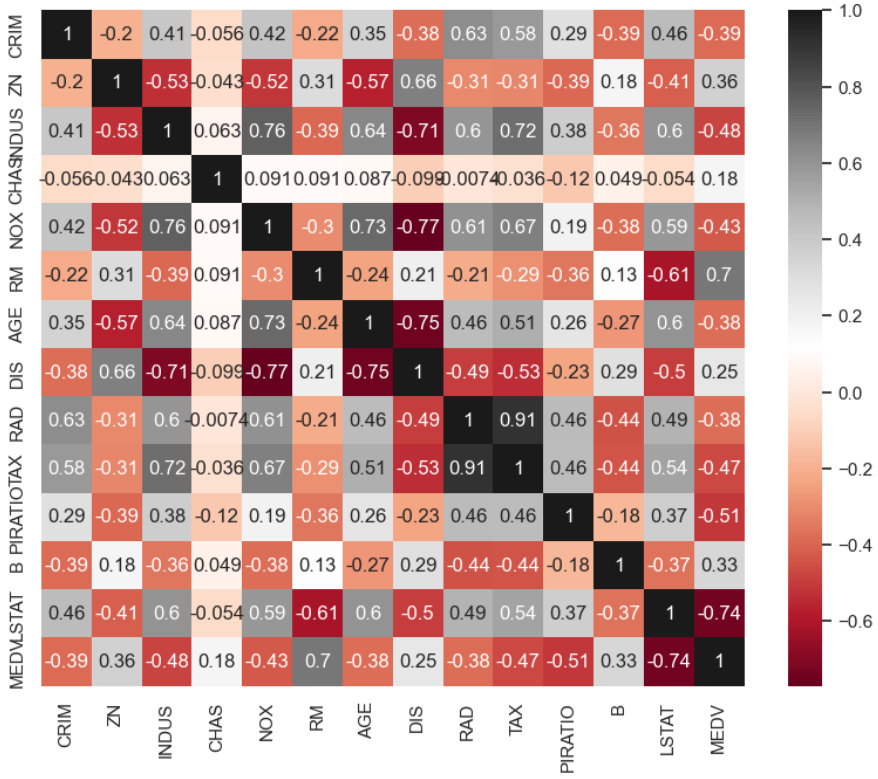


Figure 2. Thermal map of correlation coefficient between all features

Then this paper draws the correlation coefficient heat map between all the features. As shown in figure 2, Among these variables, LSTAT had the highest correlation with MEDV (r=-0.74), followed by RM, PTRAIO(|r|>; =0.5), TAX, NOX(|r|> =0.4).

Thus, house prices may have a certain but not strong correlation with LSTAT (percentage of bottom population), RM (average number of rooms per residence), PTRATIO (teacher-student ratio per ton), TAX (tax rate per $10,000 of full value property), NOX (concentration of nitric oxide)

3.3. Simplify the variable multiple regression analysis

By calculating the correlation coefficient between the independent variable and the response variable, it can be found that RM, PTRATIO and LSTAT are significantly correlated with the response variable, so a linear regression model is established for them, as shown in table3,

1) Regression equation:

$$y = 4.515\text{RM} - 0.930\text{PIRATIO} - 0.571\text{LSTAT}+18.567$$

Table 3. Suggests that the mean of the median house price is

	coefficient	t Stat	P-value
RM	4.515	10.60278	7.73e-24
PIRATIO	-0.930	-7.91069	1.64e-14
LSTAT	-0.571	-13.5402	7.9e-36

2) Interpretation of regression coefficient

As RM increases, so does MEDV. Because as the number of houses increases, the relative price of houses should decrease.

As LSTAT increases, MEDV decreases. Because where there are a lot of low-income people, housing prices are lower in the areas where they live.

As PTRATIO increases, MEDV decreases. Because the teacher-student ratio indicates the education development status of a local area, the larger the ratio, the lack of teachers in the area, the poorer education status, so the housing price in the area will be low.

3.4. Result

Suppose Tom is a home broker in the Boston area and uses this model to assess the median offer for the home they want to sell. What approximate price would Tom recommend for each client's home sale, as shown in table 4 and table 5.

Table 4. Information collected by three customers

Independent variable	Customer 1	Customer 2	Customer 3
Average number of rooms per dwelling	5	2	9
Student to teacher ratio in town	22:1	44:1	5:1
Population status decreased by %	18%	25%	2%

Table 5. Suggests that the mean of the median house price

Customer 1	Customer 2	Customer 3
17.83637	6.365289	43.99365

4. Conclusion

Although the error standard deviation of the model is small, the goodness of fit of the model is average. It may be that the linear regression model is not suitable, or the collected data may not fully explain the value of the response variable. It is necessary to try to build a nonlinear model or other models to improve the goodness of fit. Outliers can be seen in the data, which can neither be blindly deleted nor should be ignored, and should be analyzed and discussed in detail. Data collected in 1978, taking into account inflation, are not applicable today because of changes in relevant policies. The variables collected by the above data can not completely describe a house, and the house value is also

affected by the appearance of the house, the degree of old and new. For a large city like Boston, the regression model applies only to it, not to other towns. House prices can vary widely within a neighborhood (many slums in the United States are close to wealthy neighborhoods), and our forecasting model has the disadvantage of homogenizing house prices within the same neighborhood.

References

[1] He, Xiaonian. Duan, Fenghua. Linear regression case analysis based on the Python. Microcomputer applications, 2022, 38 (11): 35 - 37.

[2] Li, Kuochen. Heteroscedastic difference test and estimation method study in the linear regression model. Shanxi Finance and Economics University, 2023.

[3] Li, Meiqi. Jin, Baisuo. Dong, Cuiling. Estimation of multiple structural change points of dependent data in a linear regression model. Chinese Science: Mathematics, 2023, 53 (07): 1007 - 1024.

[4] Ouyang, Xinyu. Lv Jin. Linear regression model analysis of the influencing factors of real estate price in Nanning city. Residential and Real Estate, 2023, (27): 110 - 112.

[5] Wang, Zhaojuan. Research on commercial housing price forecast in Shandong Province. Cooperative economy and science and technology, 2023, (17): 63 - 65.