Academic Journal of Engineering and Technology Science
ISSN 2616-5767 Vol.3, Issue 7: 11-15, DOI: 10.25236/AJETS.2020.030702

Prediction of California House Price Based on Multiple Linear Regression

Zixu Wu*

Guanghua Cambridge International School, 2788 Chuanzhou Road, Kangqiao Town, Pudong New Area, Shanghai, China
**Corresponding author e-mail: 442586308@qq.com*

ABSTRACT. *In recent years, the price of real estate continues to increase, which is related to the interests of the people and the society and has become a hotspot issue today. Therefore, it is very important to reasonably predict the price of real estate. This paper uses California house price data to solve the problem of how to predict the average annual sales price of California houses through multiple variables. The main distribution of housing prices is obtained through the data, and the influencing factors are analyzed by linear and lasso regression, including the floor area, the number of rooms, the proportion of low-income people, and educational resources in nearby areas. The research on the influencing factors of housing price is very important and basic for solving a series of problems in the current real estate market. Only when we have a full understanding of the micro mechanism of housing market price formation and the influencing factors of housing price, can we effectively and moderately regulate the price of commercial housing.*

KEYWORDS: *California House Price***,** *Multiple Linear Regression***,** *commercial housing*

1. Introduction

The price of real estate is one of the hot issues discussed in modern society. The price of real estate fluctuates greatly in domestic and abroad [1-2]. Ordinary people are very concerned about the price of real estate. The level of house price will affect the interests of many aspects. The prediction of real estate price can not only provide reference for investment decision and consumption decision-making, but also provide reference for administrative decision-making of relevant government departments. In recent years, even under the control of policies, housing prices in the whole country are still rising, and the real estate bubble problem is very serious. The research on the influencing factors of housing price is very important and basic for solving a series of problems in the current real estate market [3]. Only when we fully understand the micro mechanism of housing market price formation and the influencing factors of housing price, can we effectively and moderately regulate the price of commercial housing, which is the purpose of this essay.

In the past, scholars used grey Markov forecasting model to predict the real estate price, but there was no or only one verification of the model, which was lack of scientific, and rarely compared with other models [4-5]. This paper takes the annual average selling price of houses in California as an example, uses the multiple linear regression model to predict the house price, and compares the accuracy of the prediction.

2. Methodology

$Y = ax + B$ is a one variable linear equation, and $y = ax_1 + bx_2 + C$ is a linear equation in two unknowns. Among them, "times" refers to the maximum idempotent of the unknown number, and "binary" refers to the number of unknowns in the expression. "Multivariate" means that there are many independent variables in the expression. When $B = 0$, the equations $y = ax$, y and X always match $Y / x = a$. the coordinates of any point on the image, y value is a times of x value. This kind of relationship is called "linear". The image of linear function is a straight line, so the image of multiple linear regression function is also a straight line.

Multiple linear regression has the characteristics of multiple and linear. Regression means to use a straight line to summarize the distribution law of all points. Multiple linear regression is to describe the common characteristics of some hash points by the relationship between multiple x and result y. These images of the relationship between x and a y do not completely satisfy the relationship between any two points, but this line is the most suitable to describe the common characteristics of all the points, because the sum of the distances from all points is the smallest, that is to say, the total error is the smallest. Therefore, the expression of multiple linear regression can be written as follows:

$$y= w_0x_0 + w_1x_1 + w_2x_2 + ... + w_nx_n$$

In the linear function $y = Ax + B$, B is the intercept. It is unknown whether the regression function image predicted in multiple linear regression function passes through the origin, so it is necessary to keep a constant as the intercept. That is to say, $x_0 = 1$ in the above formula can obtain $y = w_0 + w_1x_1 + w_2x_2 +... + w_nx_n$, and w_0 is the intercept. If there is no W_0 term, the equation is a linear function of the image formed by $(N + 1)$ independent variables passing through the origin. It makes a straight line passing through the origin to describe the distribution law of some hash points, which increases the limitation and leads to the decrease of the accuracy of the result function.

The content of maximum likelihood estimation is: if two events a and B are independent, then the probability of a and B occurring simultaneously satisfies the formula

$$P(A,B) = P(A) * P(B)$$

The purpose of using multiple linear regression is to summarize the laws of some unrelated elements, and it can also be considered to summarize the probability of all events occurring at the same time [6]. The greater the probability of all things happening, the more accurate the predicted laws are. The greater the probability of an event's observed dimension, the smaller the probability of the corresponding unobserved dimension, and the more accurate the summarized rules are.

The main advantages of this method are as follows: a phenomenon is often associated with multiple factors. The optimal combination of multiple independent variables to predict or estimate dependent variables is more effective and more practical than using only one independent variable to predict or estimate. Therefore, multiple linear regression is more practical than simple linear regression. Multiple linear regression analysis is the most basic and simple one in multiple regression analysis. Using regression model, as long as the model and data are the same, the only result can be calculated by standard statistical method.

3. Data research

Each record in the database describes a Boston suburb or town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA)in 1970. The attributes are defined as follows (taken from the UCI Machine Learning Repository1):

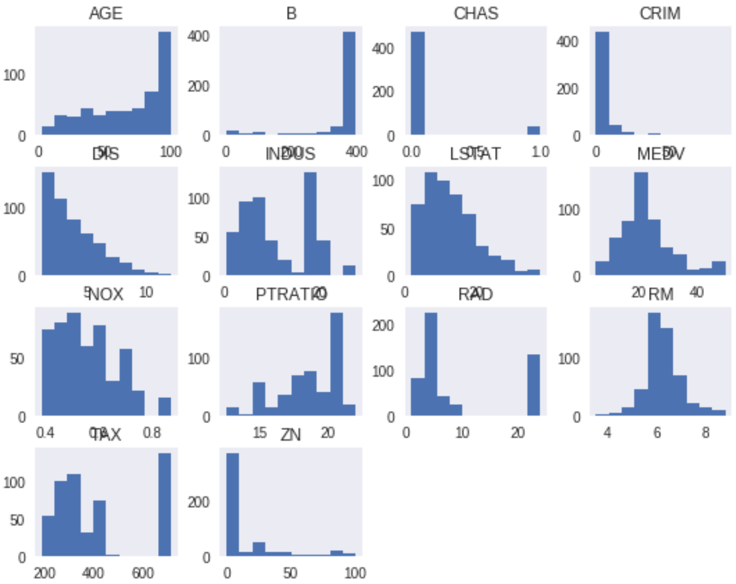

Figure. 1 Histogram of different features

CRIM is defined as per capita crime rate by town. ZN is defined as proportion of residential land zoned for lots over 25,000. INDUS is defined as proportion of nonretail business acres per town. CHAS is defined as Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). NOX is defined as nitric oxides concentration (parts per 10 million). RM is defined as average number of rooms per dwelling. AGE is defined as proportion of owner-occupied units built prior to 1940. DIS is defined as weighted distances to five Boston employment centers. RAD is defined as index of accessibility to radial highways. TAX is defined as full-value property-tax rate per $10,000. PTRATIO is defined as pupil-teacher ratio by town. B is defined as 1000(Bk−0.63)2 where Bk is the proportion of blacks by town 13. LSTAT is defined as percentage of lower status of the population. MEDV is defined as Median value of owner-occupied homes in $1000s.

4. Results

It can be found from Figure 1 that the age of houses in this area is generally older, most of which are over 100 years old. Black people account for a large proportion. There are few houses by the river and the crime rate is basically zero. As the distance to the employment center increases, the number of houses gradually decreases. The farther away from the highway, the fewer houses there are. The higher the proportion of people with low status, the less houses. Housing prices are mainly distributed around $20000, with less than $40000. The number of rooms is mainly 6, more than or less than are less. The proportion of residential land above 25000 feet is almost zero. We can see that some variables have exponential distribution, such as CRIM, ZN, AGE and B; we can see that other variables have two-peak distribution, such as RAD and TAX. I think the functions of "RM", "LSTAT", "PTRATIO" and "MEDV" are indispensable. Other unrelated functions have been excluded. From the above data analysis: the increase of RM value increases the MEDV value, that is, the housing price. The lower LSTAT value, the higher MEDV value, and the smaller PTRATIO value, the higher MEDV value. The variables with strong correlation with MEDV were RM (. 7), LSTAT (- 0.74) and PTRATIO (- 0.51). The results of LR is -23.768361 and that of LASSO is -27.493923 (10.882721).

To sum up, according to linear regression and lasso regression, the variables RM, LSTAT, PTRATIO have great influence on the target variables. The main reasons include: the more rooms, the larger the occupied area, the higher the price; the larger the proportion of low status population, the uneven distribution of income in the region, the slow economic development, and the difficulty of rapid growth of housing prices; the higher the teacher-student ratio, the more education resources in the region Rich, higher prices.

5. Conclusions

This essay uses California house price data and multiple linear regression model to solve the problem of how to forecast the annual average sales price of local houses. Through the data, it is found that the house prices are mainly distributed around 20000 US dollars, and the more or less are gradually reduced. According to linear and lasso regression, the main influencing factors include the number of rooms, income distribution, teacher-student ratio, and when the number of rooms is more, the house price is higher; when the income distribution is more uniform, the house price is higher; when the teacher-student ratio is higher, the house price is higher.

References

[1] Mishkin, F. S. (2001). The transmission mechanism and the role of asset prices in monetary policy (No. w8617). National bureau of economic research.

[2] Liu, J. G., Zhang, X. L., & Wu, W. P. (2006, May). Application of fuzzy neural network for real estate prediction. In International Symposium on Neural Networks (pp. 1187-1191). Springer, Berlin, Heidelberg.

[3] Cain, M., & Janssen, C. (1995). Real estate price prediction under asymmetric loss. Annals of the Institute of Statistical Mathematics, 47(3), 401-414.

[4] Li, D. Y., Xu, W., Zhao, H., & Chen, R. Q. (2009, July). A SVR based forecasting approach for real estate price prediction. In 2009 International Conference on Machine Learning and Cybernetics (Vol. 2, pp. 970-974). IEEE.

[5] Sarip, A. G., Hafez, M. B., & Daud, M. N. (2016). Application of fuzzy regression model for real estate price prediction. Malaysian Journal of Computer Science, 29(1), 15-27.

[6] Xiaolong, H., & Ming, Z. (2010, July). Applied research on real estate price prediction by the neural network. In 2010 The 2nd Conference on Environmental Science and Information Application Technology (Vol. 2, pp. 384-386). IEEE.