The research on factors influencing house value-take California as an example

Tianzhen Li[1], Xuezhou Yang[2,3]

[1]Capital Normal University High School, 10048, China
[2]University of Connecticut, Storrs Campus, 06269, United States of America

[3]xuezhou.yang@uconn.edu

Abstract. Housing price is a popular and important topic in today's society. This article aims to find the factors that have impacts on the housing price. To find the relationships between factors, this article uses Multiple Linear Regression as the method to perform a significant analysis of factors. 1000 samples of California's block groups in 1990 are selected for this research. Based on the assumption, this research chooses 8 explanatory variables for the analysis. Because of the relationships between explanatory variables, the article also adds interaction terms between latitude and longitude, and population and total bedrooms to solve the multicollinearity problem among explanatory variables. To optimize model analysis effectiveness, this research compares the significance, VIF value, and GVIF value of explanatory variables. The analysis result shows that the geographical location (Latitude and longitude), the housing median age, the total bedrooms, the population, and the median income make significant impacts on the housing value. Among these factors, the median income is the main factor.

Keywords: Housing prices, multiple linear regression, interaction terms, California.

1. Introduction

Housing prices, one of the most valued economic indicators in today's society, seriously affect the daily and economic life of the population. Housing has significant meaning for each individual [1]. It is also one of the important factors in people's health and social welfare [2]. At the same time, housing also has significant impacts on economic development [3]. The demand for this behavior has made the house value become an important topic for the society. At the same time, the estate is also a popular financial product, which makes house value more significant for research. For nearly 40 years, U.S. home prices have been growing fast, and sometimes unstable. Understanding factors influencing the house value is important for predicting the housing price and further economic influences. This paper aims to use California's house value as an example to predict the median housing price based on the different possible factors.

In real society, housing prices have complex relationships with many kinds of potential factors. In academia, finding the factors influencing house value is also a popular topic. Mao et al. used the King County Houses Sales data to analyze the geographic feature for housing price prediction. Multiple linear regression methods and 10-fold cross-validation are used in their research, and the influence of the number of bedrooms, latitude, and longitude is examined to be the feature for housing price. Their analysis is predictive and quite understandable [4]. Graha et al. used exploratory data analysis of the changes in population to find the relationship between population and house using data in Lisbon. Their research examined the relationship between different groups of people and different types of housing. Population and housing price is proven not only have a two-sided relationship in their article. This analysis is complicated but accurate [5]. Hao et al. also used California's housing data to examine the possible features of housing prices. They used multiple linear regression to analyze the relationships between housing price and median house value, median income, median housing age, total rooms, total bedrooms, population, and households [6]. Paul-Francois et al. use linear and nonlinear ARDL models to evaluate the effects of the economic, financial, and political risk factors of country risk on the prices of different segments of houses [7]. Na Li et al. have done research on the effect of policies like macroeconomic regulation and control or the two-child policy on Chinese housing prices [8]. Onur Özsoy et al. use CART to approach and the results indicate that sizes, elevators, the existence of security, the existence of central heating units, and the existence of view are the most important variables crucially affecting housing prices in Istanbul [9]. Wei-Shong Lin et al, compared different factors' influence on housing price in The Northeast of America and the West [10]. In summary, this article will use the multiple linear regression model to analyze the influence of longitude, latitude, housing median age, total rooms, total bedrooms, population, households, and median income on California's housing prices.

2. Methods

2.1. Data source

This research uses the dataset from the Kaggle website (California House Price). The dataset was based on the California Census by the US Census Bureau in 1990. The US Census Bureau uses block groups as the smallest geographical units for the samples in this dataset, and this dataset contains 20,640 block groups (samples). This research selected 1,000 of them randomly as samples.

2.2. Data preprocessing

The original dataset has 207 null values for total bedrooms. To fix this, all the null value is filled using the median value for these variables. At the same time, one variable (ocean proximity) is a categorical variable. This research chose to remove this variable. Eventually, 1,000 of 20,640 samples in the original dataset are chosen randomly to use as the dataset for this research. The data contains 8 explanatory variables (longitude, latitude, housing median age, total rooms, total bedrooms, population, households, and median income) and 1 target variable (median house value). The symbols and the meanings of each variable are shown in Table 1.

Table 1. Symbols and Meanings of Variables.

Variable	Symbol	Meaning
Longitude	x_1	Longitude of the block group
Latitude	x_2	Latitude of the block group
Housing Median Age	x_3	The median of the housing age in the block group
Total Rooms	x_4	Number of total rooms in the block group
Total Bedrooms	x_5	Number of total bedrooms in the block group
Population	x_6	Population of the block group
Households	x_7	Number of households in the block group
Median Income	x_8	The median of people's income in the block group
Median House Value	Y	Median of house value in the block group

2.3. Method introduction

This paper uses a multiple linear regression model to analyze the factors influencing housing values. In order to get the optimized model, the research compares the accuracy of several multiple linear regression models using different combinations of variables.

The multiple linear regression model is a model to explain the linear relationship between the target variable and more than one explanatory variable. It uses ordinary least squares to estimate the regression coefficients for each explanatory variable so that the Residual Sum of Squares between the actual target variables and predicted target variables is minimized.

3. Results and discussion

3.1. Multiple linear regression

To increase the accuracy of the prediction, a process of checking factors that have no or weak correlation to the median housing price is needed. The result of this process is shown in Fig. 1.

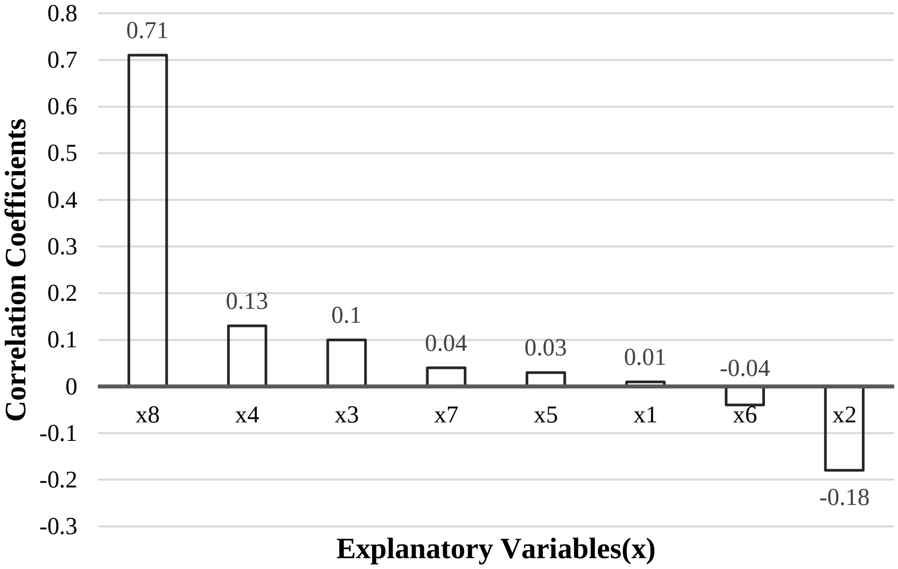

Figure 1. Relevance Analysis Between Dependent Variables and Independent Variable

From Figure 1, the Pearson test shows the correlation coefficient between all the factors and median house value. The data shows that the median income has the strongest positive relationship with the median house value; Total Rooms, Housing Median Age, Households, Total Bedrooms, and Longitude have positive relations from strong to weak; And Latitude and Population shows negative relationship with the median house value.

After the Pearson test, a multiple regression analysis was applied to the dataset. The general mathematical model for multiple linear regression is:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_8 x_8 + e \quad (1)$$

Where β_0 is the intercept t(constant), and e is the residual.

Table 2 shows the regression coefficient of the multiple linear regression equation model. The P-values of the T-test of x_1, x_2, x_3, x_5, x_6, x_8 did not exceed 0.001. However, the P-values of x_4 and x_7 both exceed the significant level. This means that t6 explanatory variables have a significant impact on the target variable Y. The impact of x_4 and x_7 are not significant. Therefore, these two variables were taken out from later analysis for the model's prediction accuracy. Table 3 is the regression coefficient of the multiple linear regression table without the data of total rooms and households.

Table 2. Linear regression coefficient table.

	Estimate	Std. Error	T Value	significance	VIF
Constant	-2,988,000	299,700	-9.967	0.000	
X1	-36,310	3,424	-10.603	0.000	9.014
X2	-38,360	3,246	-11.820	0.000	9.351
X3	1,419	201.5	7.041	0.000	1.254
X4	-7.881	3.754	-2.099	0.004	14.04
X5	111.3	26.49	4.204	0.000	23.85
X6	-15.43	2.787	-5.535	0.000	3.583
X7	-0.3522	29.86	-0.012	0.991	26.94
X8	42,370	1,511	28.039	0.000	1.753

The relevant multiple linear regression equation can now be obtained from Table 3.

$$E(Y) = -3{,}075{,}000 - 39{,}990x_1 - 37{,}590x_2 + 1{,}447x_3 + \cdots + 40{,}420x_8 \tag{2}$$

The equation is used to predict the original data and get several results: The correlation coefficient R of this model is approximately 0.798; the coefficient R-squared for fitting multiple linear regression is 0.637; and the adjusted R-squared is 0.6348. This indicates that this multiple linear regression equation has a certain ability to explain the relationship between the target variable and explanatory variables.

Table 3. Linear regression coefficient table without x_4 and x_7

	Estimate	Std. Error	T Value	significance	VIF
Constant	-3,075,000	289,900	-10.607	0.000	
X1	-39,990	3,054	-13.092	0.000	8.258
X2	-37,590	3,279	-11.464	0.000	8.239
X3	1,447	201.2	7.192	0.000	1.246
X5	-16.84	2.533	-6.648	0.000	2.952
X6	75.47	9.477	7.963	0.000	3.046
X8	40,420	1,202	33.639	0.000	1.105

After linear regression analysis, a Normalized P-P Plot in Figure 2 was constructed using the processed dataset.

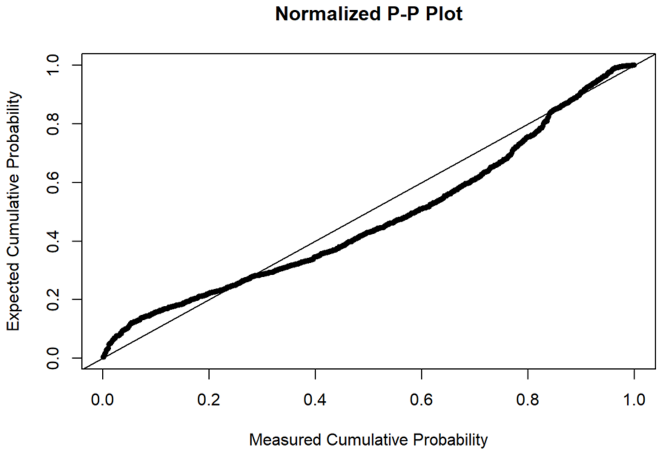

Figure 2. Normalized P-P plot of regression standardized residuals.

As the plot shows, the overall pattern of measured cumulative probability and expected cumulative probability follows approximately a straight line, which means the data fit a normal distribution.

3.2. Interaction terms

Considering the fact that in a model without interaction terms, one explanatory variable's changes depending on the value of another explanatory variable may result in a high bias in the estimated regression coefficient. An analysis of interaction terms among factors that influence median house value is necessary for the model prediction accuracy. The longitude and latitude may combine as position as a single factor, and the population and total bedrooms probably have a strong link between values. To solve this problem the coefficient of the interaction terms must be added to the previous equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_7 x_1 x_2 + \beta_8 x_5 x_6 + \varepsilon \tag{3}$$

x_1x_2 and x_5x_6 are interaction terms of the combination of Latitude-Longitude and Total Bedroom-Population, each has their own correlation coefficients β_7 and β_8. And the result of multiple linear regression model of the equation with interact terms are shown in Table 4.

Table 4. Linear regression coefficient table with interact terms

	Estimate	Std. Error	T Value	significance	GVIF
Constant	-6,299,000	3,440,000	-1.811	0.070	
X1	-64,390	28,770	-2.238	0.025	1.212
X2	45,070	95,960	0.470	0.639	1.212
X3	1,305	204.1	6.395	0.000	1.295
X5	84.33	9.749	8.651	0.000	1.200
X6	-29.70	4.433	-6.700	0.000	1.200
X8	40,090	1,200	33.407	0.000	1.114
X1 X2	719.0	798.9	0.900	0.368	
X5 X6	0.003049	0.0008698	3.505	0.000	

As the influence of the interaction terms, x_1x_2 and x_5x_6 , been considered in the model and shows a positive effect on median house value prediction, some of the other factors became less significant in prediction such as x_1. The model's accuracy is improved after taking x_1 out of the table. Based on several tests on other factors, only x_1 shows negative

Effect on data prediction and other factors contribut to it. As the result, the model been adjust again and this time only x_2, x_3, x_5, x_6, x_8, x_1x_2, and x_5x_6 are kept. The equation of the multiple linear regression:

$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \cdots + \beta_6 x_1 x_2 + \beta_7 x_5 x_6 + \varepsilon \tag{4}$$

The result of this analysis is shown in Table 5.

Table 5. Linear regression coefficient table with interaction terms without x_1

	Estimate	Std. Error	T Value	significance	GVIF
Constant	1,466,000	113,800	12.882	0.000	
X2	-167,500	13,840	-12.101	0.000	1.188
X3	1,335	204.0	6.543	0.000	1.290
X5	82.64	9.739	8.485	0.000	1.189
X6	-28.79	4.423	-6.509	0.000	1.189
X8	40,350	1,197	33.719	0.000	1.104
X1 X2	-1,057	91.15	-11.599	0.000	
X5 X6	0.002903	0.0008691	3.340	0.000	

All the significance are smaller than 0.000 which means all the factors considered in this model are making significant impact on the median house values. And x_2, x_3, x_5, x_6, x_8, x_1x_2 and x_5x_6 explained 63.98% of Y. The model formula is:

$$Y = 1{,}466{,}000 \pm 167{,}500x_2 + 1{,}335x_3 + 82.64x_5 \\ + -28.79x_6 + 40{,}350x_8 + -1{,}057x_1x_2 + 0.002903x_5x_6 \quad (5)$$

4. Conclusion

This research randomly selected 1000 samples from the dataset of the California Census by the US Census Bureau in 1990. The selected samples are preprocessed, which has 1 target variable and 8 explanatory variables. This research uses multiple linear regression analysis as the method to get an accurate and detailed relationship between variables.

During the analysis, 4 multiple linear regression analyses are used for comparison to find the possible relationships between explanatory variables and the target variable. To get further relationships, The research also adds the interaction terms between explanatory in multiple linear regression analysis. In result, the factors that make significant impacts on the house values are the geographical location (Latitude and longitude), the housing median age, the total bedrooms, the population, and the median income. From these factors, the median income is the main factor. Total rooms and households cannot be proved to have impacts on the house values in this research.

With this result, the government and related companies can have a reference to adjust strategies for society. Individuals can have a reference to get an ideal budget for housing from different angles. However, this research cannot fully explain the relationships between factors. The sample size is relatively small, and the factor number is also relatively small. At the same time, the datasets are from 1990. This means some time-sensitive factors will affect the accuracy of the results. To improve this, the newest data and more factors should be considered for further study.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Adams J S 1984 The Meaning of Housing in America. Annals of the Association of American Geographers, 74, 515-526.

[2] Rolfe S, et al. 2020 Housing as a social determinant of health and wellbeing: developing an empirically-informed realist theoretical framework, BMC Public Health 20, 1138.

[3] Maclennan D, Ong R and Wood G 2015 Making connections: housing, productivity and economic development. Australian Housing and Urban Research Institute Limited, Melbourne.

[4] Mao Y and Yao R 2020 A Geographic Feature Integrated Multivariate Linear Regression Method for House Price Prediction. 2020 3rd International Conference on Humanities Education and Social Sciences.

[5] Garha N S and Azevedo A B 2021 Population and Housing (Mis)match in Lisbon, 1981–2018. A Challenge for an Aging Society. Soc. Sci., 102.

[6] Hao Y, Zhuang L, Ying Z and Zhai J 2023 Influencing Factors of California Housing Prices in 1990: a Multiple Linear Regression Analysis. ICEMME, EAI.

[7] Muzindutsi P F, et al. 2021 The effects of political, economic and financial components of country risk on housing prices in South Africa. International Journal of Housing Markets and Analysis.

[8] Li N, Li R Y M and Nuttapong J 2022 Factors affect the housing prices in China: a systematic review of papers indexed in Chinese Science Citation Database. Property Management796.

[9] Özsoy O and Şahin H 2009 Housing price determinants in Istanbul, Turkey: An application of the classification and regression tree model. International Journal of Housing Markets and Analysis, 167-178.

[10] Lin C T, et al. 2014 Effects of socioeconomic factors on regional housing prices in the USA. International Journal of Housing Markets and Analysis.