House prices and properties in New York

lab_06_Re_late_2

This manuscript was compiled on November 14, 2021

Housing price is an issue that many people are concerned about now. This investigation aims to determine which properties influence New York house price by building a multiple linear regression model for prediction. We use data from random sample of 1734 houses taken from full Saratoga Housing Data (De Veaux). For our research, we divided the research question into some different parts to solve it. We finally decided to use the multiple linear regression model with the selected variables. We found that waterfront is main factor will affect New York house price. And as the result of our research, we can find that adjusted R-squared is 0.6554 and RMSE is 57972.24.

Introduction

Background. There are many factors can affect house price. People will be more likely to buy a house with waterfront, so the price of a house with waterfront will increase. The land value is another important factor that affect house price, the bigger the house, the higher the price. People also want to live in a house with more bedroom and bathroom, so it will influence the house price. The more bedroom and bathroom there are, the more expensive house price will be. Air conditioner is also necessary, people need a livable space warm in winter and cool in summer to live in. Therefore, house with air condition have a higher price. The new construct houses and the old houses are not popular with home buyers. These houses usually dons have a high price.

Our research question is to find which properties influence New York house price and we build a multiple linear regression model to solve it. We also created an additional model for comparison.

Data Set. The original data contains 14 numerical variables and 3 categorical variables. They are Price, Lot.Size, Waterfront, Age, Land.Value, New.Construct, Central.Air, Fuel.Type, Heat.Type, Sewer.Type, Living.Area, Pct.College, Bedrooms, Fireplaces, Bathrooms, Rooms and Test. For the analysis, we rename the column names of data, and the type of some variables are changed.

For the variables 'waterfront,' 'new_construct,' 'central air' and 'test,' the original samples are "0" and "1," we change these to the format "TRUE" and "FALSE." Moreover, for the variables "fireplaces," "fuel_type," "heat_type" and "sewer_type," we change these from type "chr" to "fct." Furthermore, we remove all the 'na' values from the data set.

Hence, by the initial data cleaning, now this data set is consisting of 10 numerical variables and 7 categorical variables.

For our report, the response variable is New York house price. The data range of this variable is from 5000 dollars to $775000. The median is 189700 dollars and the mean is 211545 dollars. From the histogram, we could see that the majority of the house price is around 200000 dollars.

Analysis

Assumptions. Assumption checking was carried out before and after performing variable selection to guarantee validity. Residuals are symmetrically distributed above and below zero, thus the linear assumption is reasonable. The New York house price is not affected by the price of the other house, the properties between the two houses should be independence, which follows the independent assumption. There are some outliers in the Residuals vs Fitted graph because the sample size is big, therefore we use the central limit theorem. The residuals do not appear to be fanning out or changing their variability over the range of the fitted values so the homoskedasticity assumption is met. Residuals QQ plot shows that most points are close to the straight line, also relying on the central limit theorem, the normality assumption is satisfying.

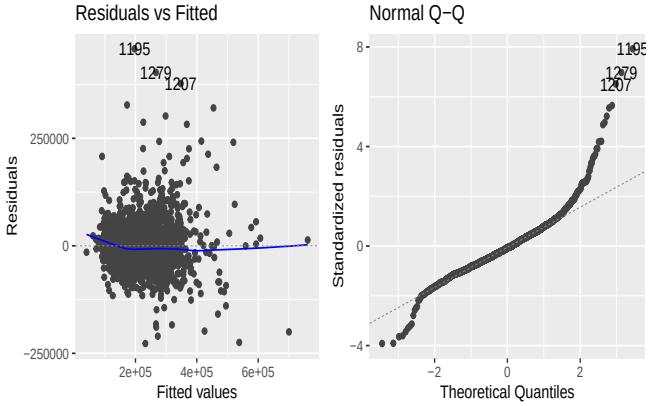

Fig. 1. Residuals plots for the final regression model

Model Selection. Log transformations were performed in attempt to fit the house price variable into a linear relationship with the dependent variable. Firstly, we fit the full model, there are some variables that have high p values and so are not significant, so we need to drop them and find a more accurate one. We then use backward and forward search using AIC to find out 2 regression models and then compare them and choose the one with smaller AIC value which will be the more appropriate one. The models found by backward

Fact

Because there are many variables that affect New York housing prices, including continuous variables to see the relationship of linearity through ggplot, and some categorical variables to observe normality through boxplot and qqplot, the forecasting factors are a bit complicated. The final model tells us the relationship between the living area and the house prices is very close.

and forward search are quite similar. They have the same r-squared(0.65544) and RMSE (57972.24) value.

Results

Inferences.

$$\begin{aligned}\widehat{\text{price}} = 5902.21 &+ 7412.19(\text{lot\_size}) - 140.27(\text{age}) \\ &+ 120542.25(\text{waterfront}_{\text{TRUE}}) \\ &+ 0.92(\text{land\_value}) \\ &- 44721.76(\text{new\_construct}_{\text{TRUE}}) \\ &+ 9633.18(\text{central\_air}_{\text{TRUE}}) \\ &+ 9971.3(\text{heat\_type}_{\text{Hot Air}}) \\ &- 379.26(\text{heat\_type}_{\text{Hot Water}}) \\ &- 32502.48(\text{heat\_type}_{\text{None}}) \\ &+ 69.95(\text{living\_area}) \\ &- 7621.33(\text{bedrooms}) \\ &+ 23125.86(\text{bathrooms}) \\ &+ 3032.63(\text{rooms}) + 5825.32(\text{test}_{\text{TRUE}}) + \epsilon\end{aligned}$$

The final model tells us the relationship between the living area and the house prices is very close. Forward and backward AIC search yielded the same results for predicting housing price, indicating a stable regression model. The house price will be higher when house with waterfront; or the lot size, land value, living area, number of bathrooms and number of rooms of the house is larger. However, the house Price will reduce depend on new construct houses; the age history of house, and number of bedrooms will influence the house price decreasing.

Performance. We compared our final model with the full model which contains all variables in the data set being used as predictors. Out of sample performances are tested using "Caret" package at a 10-fold cross-validation. Our final model formula consists of 12 variables. With the root-mean square(RMS) , R-squared and mean Absolute Error we are able to measure the data of performance. Meanwhile, by contrasting with the simple model which possesses a RMSE of 58101.24, the RMSE of the final model(57972.24) is significantly smaller. The difference indicates a smaller prediction error of the final model. In addition, the final model has a larger R-squared value(0.655). We also find that the mean absolute error(MAE) of the final model is 41378.33. It is also much smaller than the MAE of the sample model(41496.82).

```
## [1] 57792.64
```

```
## [1] 57830.78
```

Table 1. Performance results of models

Attributes	Full Model	Final Model
adjusted r-squared	0.6482	0.6554
In Sample RMSE	58101.24	57972.24
In Sample MAE	41496.82	41378.33
Out of Sample RMSE	57792.64	57830.78

Discussion

Limitations. It is not difficult to find that whether it is the R-squared of the full model or the final model, their results are not big. And their RMSE and MAE are both very large numbers. This may lead to inaccurate results.

This formula can not completely represent that the housing prices in New York are determined by these variables. Because the total number in the data set is only 1734. This data set is not large enough. There is also a selection bias. Because this data may be collected through online or offline questionnaire surveys. It can not force everyone to fill in. It is possible that some people are unwilling to fill out the survey. Therefore, this does not represent the impact of the entire New York housing price trend.

Moreover, it may also have non-response bias. Some people may not disclose the content of their house structure completely in order to protect their privacy. This may cause the value in the formula to be affected.

Conclusion. Regarding some suggestions for improvement in the future, I think we should first expand the data capacity of the sample. In this way, we can receive more samples so that we can further improve the accuracy of the data in the formula. In addition, we can provide more variables in the questionnaire, such as the income of the buyer, and the commercial construction and transportation system in the surrounding area. In the future, these are variables that are most likely to affect housing prices.

Although there are some factors that affect the value and are difficult to correct in this survey, this does not mean that the research is meaningless. Through the content of the data set and the research results, we can find that people's requirements for the quality of life are getting higher and higher. For example, the type of heating, the surrounding environment of the house and the central air-conditioning. The size of the house and the number of rooms are no longer decisive variables for house prices, they also hope that they can get a more comfortable place to live. Therefore, I feel that we need to update this research from time to time. Perhaps in the future, the variables affecting house prices may be more interesting and even more different from the conclusions we have studied today.

References

GitHub repository

[1]Strozier, Matthew (23 December 2011). "Introducing the Home-Price Scorecard." Wall Street Journal.https://www.wsj.com/articles/BL-DVB-20633

[2]K1084-2011: Recognizing Yunnan Province and Chongqing Municipality of the People's Republic of China as a "Sister City" with New York City. New York State Senate. [2012-12-16].https://www.nysenate.gov/legislation/resolutions/2011/k1084

[3]Property Prices Index by Country 2019. [2019-02-17]. https://www.numbeo.com/property-investment/rankings_by_country.jsp