Study of Influential Factors on Housing Prices by Using MLR

Chang Liu *

Facalty of Arts and Sciences, University of Toronto, Ontario, Canada

* Corresponding author: Changnn.liu@mail.utoronto.ca

Abstract. With the emergence of the 2019 pandemic, the global economy experienced a downturn, but the housing market quickly rebounded post-pandemic, heightening the anxiety surrounding home buying. Previous research predominantly employed regression models to forecast housing prices, but these studies largely focused on specific factors. Hence, this article aims to offer a more comprehensive viewpoint by exploring the impact of multiple independent variables on housing prices, particularly emphasizing the effects of the age of buildings, the surrounding environment, and architectural factors. The methodology used in this study is the Multiple Linear Regression (MLR) model. It analyzes the real estate market data in Taipei, meticulously constructing and validating multiple models. The findings reveal that the age of buildings and the distance to the nearest subway station negatively influence housing prices, whereas the number of convenience stores positively impacts them. Among these factors, the quantity of convenience stores exerts the most significant effect on housing prices. Overall, this article provides a novel perspective and tools for understanding and predicting housing prices, assisting real estate developers and buyers in making more informed decisions in the complex real estate market. This study highlights the multidimensional nature of real estate value and contributes to the sustainable development of the real estate market.

Keywords: Meachine learning, House price, Prediction.

1. Introduction

With the emergence of the black swan event of the 2019 epidemic, the global economy started into recession. [1] However, the recovery from the epidemic awakened the real estate market at an accelerated step, also intensifying people's anxiety about buying houses. According to CBC News, "In the wake of the short but steep COVID-19 recession, average Canadian house prices hits $816,720— up 20% in the past year [2]. For a long time, buying a house has been an important livelihood issue, thus it's necessary to provide the public with clearer options for choosing properties within a limited budget.

At present, using regression models to predict and estimate the impact of multiple independent variables on a dependent variable based on existing data is a commonly utilized tool in statistics. Furthermore, due to the simplicity and comprehensibility of linear models, regression models have been employed as the primary method for predicting apartment values in this study. As the study paper of authors Dorantes, L. M., et.al, three modeling methods, and multiple variables were used to comprehensively evaluate the impact of Metro Line 12 in Madrid, Spain on the surrounding housing prices. The research result is very significant, the distance between the subway station and the housing price is inversely proportional [3]. Yiu, C. studied the depreciation rate of buildings and the significance of reducing the depreciation rate for Hong Kong's sustainable development plan [4]. Through a cross-sectional hedonic pricing model, studied one of the residential units built by MFSC between 40 years obtained the true depreciation rate of the housing price, The results show that 40-year-old residential units have depreciated by about 45% compared to newly constructed residential units. Ying et.al studied the impact of convenience stores on housing prices [5]. The experiment uses quantile regression to demonstrate the positive and negative relationship between factors such as "availability" and "Density" of convenience stores in Taipei and low-priced and high-priced houses. The results show that the number of convenience stores is related to housing prices, and the higher the density of convenience stores, the more low-priced properties.

Based on the foregoing discussion, current research provides a theoretical foundation for this study in predicting housing prices using linear regression models. Additionally, it offers background

support for subsequent research on the impact of the number of convenience stores on housing prices. However, existing studies predominantly employ standard least squares and spatial error modeling methods. This paper aims to undertake dispersed research, focusing on predictions around multiple subway lines, with a particular emphasis on exploring the impact of a house's age and its surrounding environment and architecture on housing prices.

2. Methods

2.1. Description of Data

The data used in this research comes from https://archive.ics.uci.edu/ml/datasets/Real +estate+valuation+data+set, which collects the housing market transactions in Taipei Taiwan. There are a total of 414 observations and 8 attributes, as the rough data cleaning process mentioned in the method section finally retained 412 observations after removing missing values and obvious extreme values. The variables used in this article are shown in Table 1. House. age represents the age of the house, in years. Mrt. station. distance represents the distance from the house to the nearest MRT station, in meters. Num. CVS represents the number of convenience stores in the residential circle of the house. House. price represents the selling price per unit, in ping (1 ping = 3.3-meter squares).

Table 1. Summary of Important Variable

Variable	min	mean	max
house. age	0	17.7364078	43.8
mrt. station.distance	23.38284	1087.5796903	6488.021
num.CVS	0	4.0970874	10
house. price	11.2	37.8609223	78.3

2.2. Model selection

To ensure the accuracy of the experiment, the data set has cleared the missing variables and observations. Afterward, “house. price” is used as a response variable according to the literature and my research question. Correspondingly, all variables related to housing prices as predictors are used to build a model. The purpose is to select several variables ultimately suitable for this study as predictors.

Model1 is shown in follows

$$\begin{aligned}\text{House.price} = \beta 0 + \beta 1 \text{xTrans.day} + \beta 2 \text{xHouse.age} + \beta 3 \text{xMrt.station.distance} + \\ \beta 4 \text{xNum.cvs} + \beta 5 \text{xHouse.latitude} + \beta 6 \text{xHouse.longtitude}\end{aligned} \quad (1)$$

In the first step, a multiple linear regression model is built with 6 predictors. By observing the summary table of the model, only the variable with a p-value less than 0.05 was kept as the reduced model namely model 2.

After filtering, suppose that the p-value of the predictor variables in model 2 should all be less than 0.05, then the VIF testing was chosen to detect whether there is multi-collinearity in model 2. By observing testing results, if the VIF for all the predictors is greater than 5, it indicates that there is multi-collinearity in the model. To ensure the accuracy of the experiment, reducing the predictors into model 3 should be considered.

Finally, to further optimize the model, the study created a smaller model than model 3 as model 4 to do a partial F test to see if a smaller model could be better if the p-value in the ANOVA table is less than 0.05, we should reject H0, which means the smaller model is better, so the model 3 can be taken as the final model, otherwise, chose the model 4 as the final model, the predictor in that model will be the variable the study chose.

2.3. Model checking

Before the experiment started, the data was divided into a train data set and a test data set at a ratio of 30% and 70%, to test the final effectiveness of the model. The training data is used to build the model at first, while the test data is used to test and compare the model built by the training data. If the model built by the two data sets is different, then there is overfitting exists. If they are similar, it means that the model is correctly bullied to interpret the result. Moreover, the goodness to fit would be applied to test the quality of the model, for example, the better model comes with a larger adjusted r square and smaller AIC and BIC.

2.4. Model validation

The assumption of the model is a crucial premise for the establishment of the model, so the residual plot for the response variable and predictors, as well as the Normal Q-Q plot, will be created to test the four assumptions of the linear regression model. First, the errors should be independent of each other. If there is no cluster in the residual plot, then the assumption is satisfied; otherwise, having dependent error terms means that the predictive ability of the model will be worse in some areas than in others. Second, the error should follow the normal distribution, it can be proved by presenting a straight line on the normal Q-Q plot. If not, a straight line is presented, it means that we cannot use all of the handy properties of Normal distributions, such as linear combinations of Normal random variables. Third, the error term needs to have constant variance, which means the residual plot cannot have increasing or decreasing patterns, otherwise, the assumption fails to hold. Fourth, there is a linear relationship between x and y. If the residual plot does not present a curve pattern, then the assumption holds, if the relationship is not linear, fitting a linear model will produce wrong relationships. And then, to find the outlier in the model, The study will test the model by finding out that the observation where the standard residual is not between 4 and -4. Finally, the cook distance is used to check whether there is an influential point in the model. If the return value is 0, it means that there is no influential point in the model, otherwise, it means that there is.

3. Results

3.1. Validation of variable

To study the relationship between predictors and response variables, the scatter plot is made for three predictors (house. age, num. CVS, mrt. station) with the response variable (house. price) respectively. It can be also seen from Fig. 1 (a) that house. price dose follows the normal distribution. As Fig. 1 (b, c, d) showed, there is a linear relationship exists.

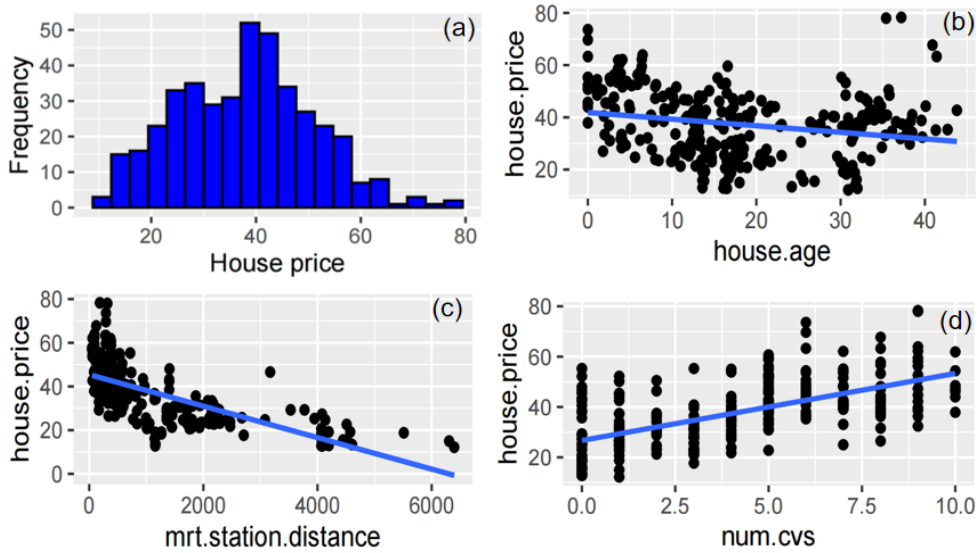

Figure 1. The scatter plot (a)The distribution of the response; (b) The scatter plot for a variable 'house. age' variable 'House Price'; (c) The scatter plot for variable; (d) The scatter plot for variable num. cvs'mrt. station. distance'

3.2. Presenting the Analysis Process and the Result

As the first step, all the variables related to the research question were kept and used in the training data to form the initial model, namely model 1.

$$\begin{aligned}\text{House. price} = -11135.957 + 3.632\text{xT rans. day} - 0.309\text{xHouse. age} - \\ 0.004\text{xMrt. station. distance} + 1.359\text{xNum. cvs} + 201.51\text{xHouse. latitude} - \\ 9.592\text{xHouse. longtitude}\end{aligned} \tag{2}$$

Then, only the variables with a p-value less than 0.05 from the summary table obtained by model are kept, they are house. age, mrt. station. distance, num. cvs and house. latitude, also considering my study focus, the model will be re-fitting to obtain the reduced model 2,

$$\begin{aligned}\text{House. price} = 42.36 - 0.309\text{xHouse. age} - 0.005\text{xMrt. station. distance} + \\ 1.51\text{xNum. cvs}\end{aligned} \tag{3}$$

Next, ensuring that all the variables in model 2 are significant toy, VIF testing is applied to check whether these predictors in model 2 have multi-collinearity. From the results, the VIF for each predictor is 1.017481, 1.555932, 1.574659, all of them less than 5, indicating that the model2 does not have multi-collinearity.

In addition, by further optimizing the selection of the model, the study chose to reduce model 2 to form model 3, which only contains two variables it,

$$\text{House. price} = 38.041 - 0.\ 005\text{xMrt.station. distance} + 1.324\text{xNum.cvs} \tag{4}$$

Compare these two models with a partial F test. From the ANOVA table, know that the p-value is 0, representing that model 2 is better than model 3. so as resulting in the final Model

$$\text{House. price} = 42.36 - 0.309\text{xHouse.age} - 0.\ 005\text{xMrt.station. distance} + 1.51\text{xNum.cvs} \tag{5}$$

3.3. Goodness of the Final Model

3.3.1. Assumption Checking

From Fig.2 and Fig.3, there is no obvious cluster shown in the residual plot, so the first assumption that the errors are independent of each other is satisfied. Observe the Normal Q-Q plot in the Fig. 2, except for a little skew in the tail, it is basically in a straight line, so the second assumption is also satisfied, the error does follow a normal distribution. In addition, the residual plot does not appear the pattern of increasing or decreasing or curve, so the third and fourth assumptions are also satisfied, that is, the error term has constant variance and there is a linear relationship between the predictor and the response variable. In summary, the four conditions of the linear regression model are All satisfied.

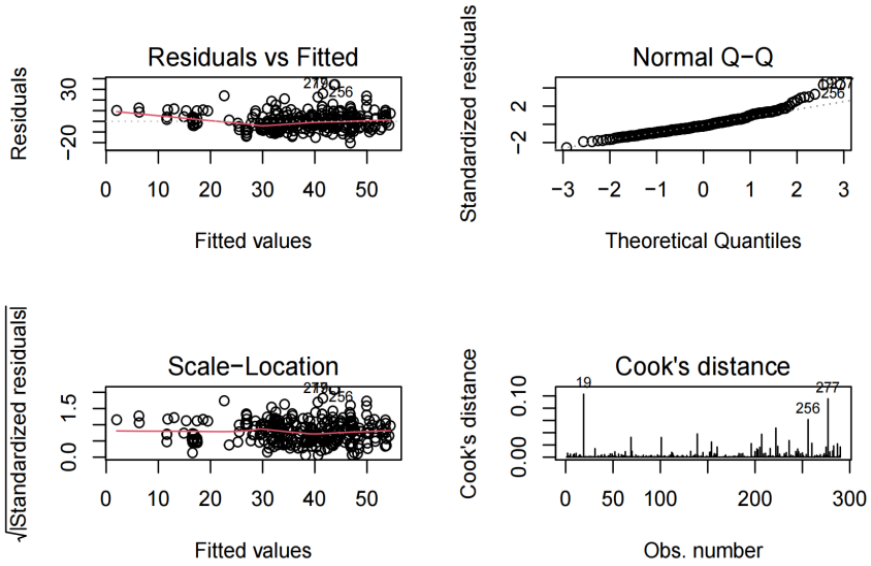

Figure 2. The Residual plot and Normal Q-Q plot for the variable ‘house. Age’

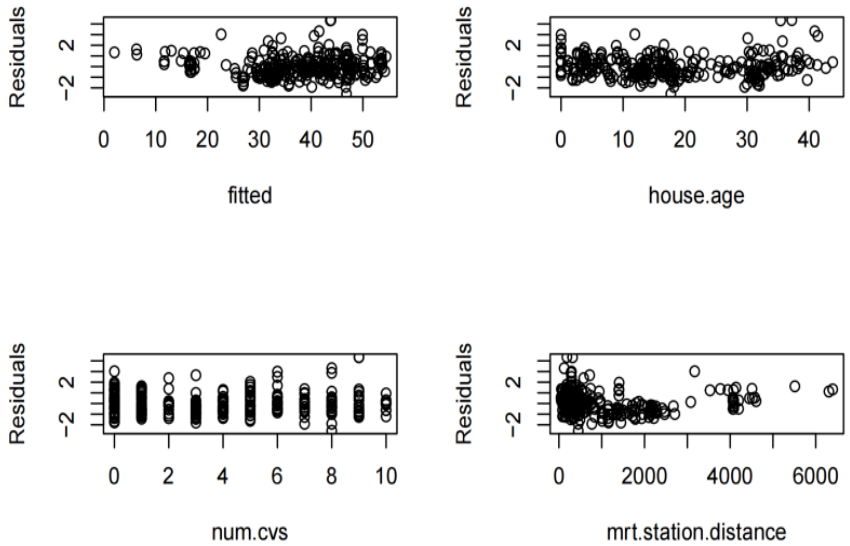

Figure 3. The Residual Plot for 3 Predictor Variables (house. age, num. cvs, mrt. station. distance)

3.3.2. Additional Checking

To find the outliers, it could be checked whether the data has a standardized residual between 4 and -4, and then, there two observations are found as outliers, which are the 19th and 177th observations.

Then, according to check whether the fitted value is greater than this, it can be concluded that there are 17 leverage points in the model that showed in Fig. 2, they are the 124th, 177th, 9th, 250th, 346th, 181st, 49th, 387th, 304th, 117th, 147th, 302th, 332th, 59th, 341th, 165th, 395th observation. Finally, cook distance is used to check whether there is an influential point in the model, and the result of the return value is 0, indicating that there is no influential point in the model.

Table 2. Summary of goodness measures for models fit to log of Prize Money

Model	Adjusted R^2	AIC	BIC
Test Full Model	0.54	531.2546222	549.2747274
Test Final Model	0.5950242	518.2487078	541.8768552
Train Full Model	0.6217126	1206.9067475	1229.2561521
Train Final Model	0.6512366	1186.2832155	1219.6422629

3.4. Comparison Checking

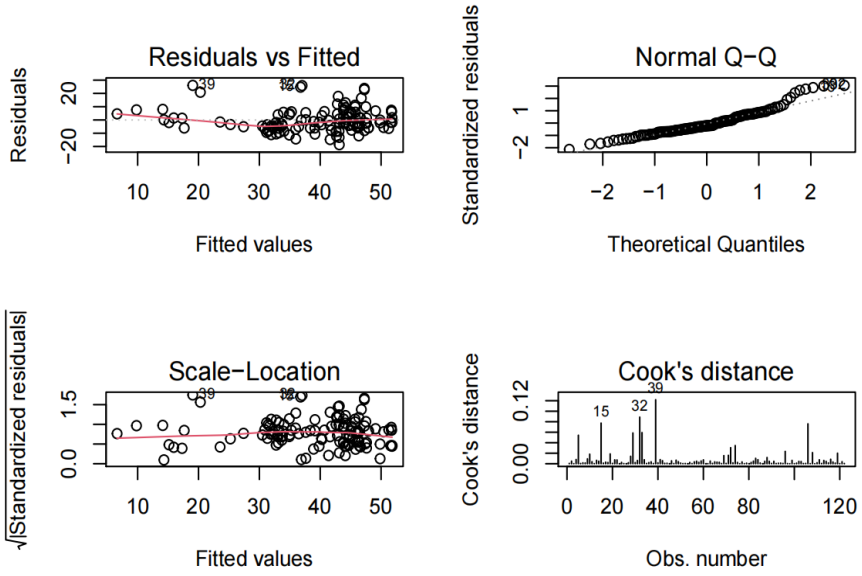

Figure 4. Model Checking for testing data

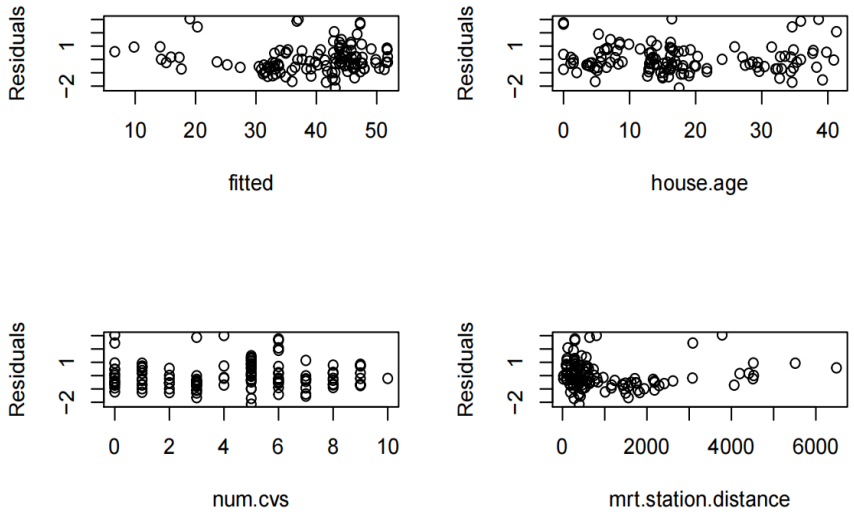

Figure 5. Residual plot for predictors in testing data

For the comparing check, the data is divided into training data and testing data, the modeling selection and checking processes are similar to the training data. From Fig. 4 and Fig.5, there is only little cluster shown in the residual plot, so the first assumption that the errors are independent of each

other is satisfied. Observe the Normal Q-Q plot in Fig. 4, except for a little skew in the tail, it is basically in a straight line, so the second assumption is also satisfied, the error does follow a normal distribution. In addition, the residual plot does not appear the pattern of increasing or decreasing or curve, so the third and fourth assumptions are also satisfied, that is, the error term has constant variance and there is a linear relationship between the predictor and the response variable. In summary, the four conditions of the linear regression model are All satisfied. So, a new model is constructed by using the testing data, namely model 4.

$$\text{House. price} = 40.256 - 0.125\text{xHouse. age} - 0.005\text{xMrt. station. distance} + 1.413\text{xNum. cvs} \quad (6)$$

Table 3. Residual plot for predictors in testing data

Model	house.age	mrt. station.distance	num.CVS	house. price
Train Model	17.7793103	1120.3595918	3.9965517	37.3527586
Test Model	17.6344262	1009.6602523	4.3360656	39.0688525

Comparing the data summary of training data and testing data, in Table 3, the house.age and mrt. station. distance in the training data are larger than the testing data on average, while the num.cvs and house. prices are smaller on average testing data. In addition, the most important thing is to compare the performance of the two models. Overall, the values of intercept and beta1, beta2, and beta3 obtained from the models established based on different two sets of data are similar, indicating that there is no overfitting.

Furthermore, from Table 2, compared with the model established by train data and test data, the model built by training data has a larger adjusted R square, indicating that the final model for training data has a larger proportion of y being explained by x. The training data has a larger AIC and BIC than the testing data, so it shows that the testing data model has a better predictive ability. Specifically, comparing the two models respectively inside the testing and training data, the reduced model is better than the full model, because the two reduced models both have larger adjusted R square, and smaller AIC and BIC.

4. Conclusion

To sum up, as a result, the study gets a model that can predict the housing price and the factors that affect it through MLR. Through the final model, it can be known that the age of the house and the distance from the house to the subway station hurt the price of the house as the value increases, which follows the result in the research literature. However, the number of convenience stores around the house has a positive impact on the house price, which is contrary to the conclusions in the literature. Among the three factors, the one that has the greatest impact on the housing price is the number of convenience stores around the house. Keeping other variables constant, one unit increase in num. CVS will give a 1.51 unit increase in the house. price. Generally speaking, this model reasonably explained the influences of houses. age, mrt. station. distance, and num. cvs on housing prices, because aging houses and scarce facilities around houses will devalue real estate, which is consistent with common sense. Therefore, from the perspective of the developer, they can increase their attractiveness by enhancing the infrastructure around the house or choosing a location with convenient public transportation to build properties, to obtain more profits. On the other hand, general housing buyers, whether it is for investment or self-occupancy, should try their best to choose properties with complete infrastructure and public transportation. As time is an uncontrollable factor, all the properties will slowly depreciate over time regardless of whether they are new or old nowadays. Therefore, on the premise of considering the first two important factors, one can maximize the self-interests by choosing newer houses.

Although this model has explained my research question very well, there are indeed some limitations. Technically speaking, the response variable (house. price) does not fully conform to the

normal distribution, and there are certain outliers and leverage points in the data, which may lead to certain deviations from the actual results. In terms of data, this data set only collects the data of the real estate market in a specific region within a short period, which does not have strong representative significance. Overall, as almost the most expensive economic activity in people's daily lives, buying a property needs to be thought through. In addition to the several factors that affect housing prices mentioned in this study, many other key factors also deserve serious consideration. Hoping that every intentional buyer can relieve the anxiety of buying a house through comprehensive understanding and learning, and finally get the one they love.

References

[1] Duca, J. V., & Murphy, A. (n.d.). Why House prices surged as the COVID-19 pandemic took hold. Dallasfed.org. Retrieved October 21, 2022, from https: //www.dallasfed.org/ research/economics/2021/1228.aspx.

[2] CBC/Radio Canada. (2022, March 16). Average house price up 20% in past year -and new listings aresurging | CBC News. CBCnews. Retrieved October 21, 2022, from https: //www.cbc.ca/news/business/crea-housing-february-1.6385274.

[3] Mejia-Dorantes, L., Paez, A., & Vassallo, J. M. Transportation infrastructure impacts on firm location: Theeffect of a new metro linein the suburbs of Madrid. Journal of Transport Geography, 2012, 22, 236 – 250. https: //doi.org/10.1016/j.jtrangeo. 2011. 09. 006.

[4] Yiu, C. Y. Building depreciation and Sustainable Development. Journal of Building Appraisal, 2007, 3 (2), 97 – 103. https://doi.org/10.1057/palgrave.jba.2950072.

[5] Chiang, Y.-H., Peng, T.-C., & Chang, C.-O. The nonlineareffect of convenience stores on residential property prices: A casestudy of taipei, Taiwan. Habitat International, 2015, 46, 82 – 90. https: //doi.org/10.1016/j.habitatint.2014.10.017.