The Real Estate Price Prediction Based on Multivariate Linear Regression Models

Yucheng Zhang[1, *]

[1]Stanstead College, Stanstead, Canada

*Corresponding author: Yzhang20@stansteadcollege.ca

Abstract. The price of households had increased greatly as the demands in the market increased. Real estate had become the most popular method to resist inflation. Thus, more experts and scholars are working on the research in this area, including but not limited to find the future trends of housing prices and the possible bubble of housing prices. This paper will use the basic data regression model in Excel, powered by Microsoft, to find the factors that lead to the appreciation or depreciation of real estate, and three different methods of dealing with the data will be applied which are standardization, interception, and normal normalization. As a result of complex processes, normalization is the best way to deal with the data that was chosen from Kaggle. In the meantime, a linear correlation was done for checking the independency of each variable, and turned out to be successful. According to the analysis, an inverse relationship between the variables house age and distance nearest to the store was found, and the lower the house age is, the higher the price is, higher the house age is, lower the price is. The main aim of this paper is to investigate the relationship of the factors and their impact of them on the price of the house, was achieved and solved mathematically, and therefore, this research paper has found its goals. These results shed light on guiding further exploration of house price estimation.

Keywords: Prediction; Regression Model; Linear Correlation.

1. Introduction

Nowadays, the price of purchasing a house had become insanely expensive, according to Statista's data in 2020 [1], property average price in thousand United States dollars had four cities which were above 750 thousand dollars Hong Kong, Shanghai, Shenzhen, and Beijing. Based on another Statista research on Hong Kong, 2020 [2], this city cost the most dollars for a housing property, the property price most more than 1.25 United States dollars. The booming of house prices is not just happening in China, in fact, many developed countries also suffer from this situation. Forbes advisor also gives their opinion on the housing price where the demands were expected to raise, and it is believed that the high price may cause inflation and geopolitical uncertainty [3]. To take a specific look, in thousands United States dollars, the average price of property in Hongkong is 1254.44, Munich is 1,000, Singapore is 915.6, Shanghai is 905.83, Shenzhen is 783.86, Beijing is 763.5, Vancouver is 754.62, Los Angeles is 717.58, Paris is 650.55, and New York is 646.03. The list was mostly large urban areas in many countries and not a single city has an apparent low household price. Therefore, for the final project, the focus on how certain factors could cause a significant impact on the prices of real estate would be done, and how will these factors be clearly demonstrated mathematically.

Due to the high popularity of real estate-related topics, a lot of seniors had written a lot of papers regarding this topic. In fact, there was a research paper on the market of households in China, by Chang Liu and Wei Xiong in November 2018 [4]. This paper was published on nber.org, a national bureau of economic research website. According to the analysis, the market of real estate is a part of China's financial system and the key part of the economy of China. Shocking enough, in 2017, housing sales totaled 13.37 trillion RMB, equivalent to 16.4% China's GDP. The paper, in general, analyzed the reasons behind the booming market of real estate in China regarding to the government's policies and discussed the hidden threat of bubble which the authors found a challenging issue. In the meantime, a passage was found that illustrates the trending situations of real estate, from realtor.com [5]. The author of this claimed that there is still a high trend and market needs for real estate, despite the fact of tighter financial conditions.

This research paper aims to find out the relationship between different factors so that in further research of real estate prediction, the authors may find a ready-made citation. Furthermore, it is vital to understand the tendency of the price so that people can make the right choice on the time, plot, and other relevance that will affect the price of the house as for the majority of individuals in the society. Monetary saving is one of the most important factors and it is hoped that the readers will have a better understanding on the house price and the relationship between each factor and the price. The rest part of the paper is organized as follows. The second part includes data and methods to approach the graphing results and to gain the formulas of each processing method. Follow-up is the results and discussion for the third part, containing correlation analysis to test the validity of variables, and contains specific graphs and tables to show the results of each model, with the relationship between x and y. In the comparison part, the values of the three models will be put together to figure out the best model that fits into the data. Then, the limitations of this paper would be discussed as well as the prospects of the topic of real estate. Finally, the conclusion part will draw a summary of the main points brought up in this paper and discuss whether the main objective is achieved.

2. Data & Method

2.1 Data

The dataset that has been applied in this study originated from Kaggle [6], a public online statistical data research website. The model that has been used in this paper is the basic regression model. It is important to understand the meaning of R-squared and P-value as it will indicate the importance of a model and whether a model is suitable for the data or not. In simple definitions, the R-squared value must be within the range between zero and one, and as the value reaches closer to one, the basic model of regression is more suitable to put into the dataset and therefore will provide a better result for the relationship. According to simply psychology [7], the P-value determines the suitability of a factor, and if the P-value of a factor is too small, typically smaller or equal to 0.05, that indicates the value does not cause a significant impact on the function of x and therefore could be ignored in the paper. Regarding to the relevance of factors, the linear correlation is also significant, as it will test every factor that has been used in the data and determine how strong its multicollinearity is, to show a good utilization of factor, every factor compared to another must be smaller than 0.5.

2.2 Method

The meaning of standardizing is letting a value to subtract a measure of location, and then it is divided by a scale measure. Standardize number should be within the range of 0 to 1, Definitions from Towardsai.net [8]. The variables used were house age as x_2, distance to the nearest MTR station as x_3, and convenience stores as x_4. Three models were followed with different methods of data processing, the first was done by standardizing the variables into a range between zero to one. The second model was done by multiplying the variables and the last one was done by z-score and normal normalization. The formulas of each model to present their relationship were listed below:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (1)$$

$$y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_4 \quad (2)$$

$$y = \beta_0 + \beta_1 x_1' + \beta_2 x_2' + \beta_3 x_3' + \beta_4 x_4' \quad (3)$$

Here, y_1 stands for the house price, β_0 stands for the intercept coefficient, β_1 stands for coefficient of x_1 and every x stands for the factor that will affect the value of y.

During the selection of the data, the first step is to select 80 percent of the train dataset and standardize the x's so the values will be unified, then the x and y were applied to the regression model in Excel and the simulation was gained. Afterward, the p-value and R-squared were gained. The same step of selecting 80 percent of the train dataset was done for models 2 and 3, and the only difference between them is the different ways of processing. Then, all three groups of data were gained, and a conclusion may be given by the relationship shown by the data.

3. Results & Discussion

3.1 Correlation analysis

Meanwhile, the linear correlation also needs to be done to show the multicollinearity of variables. According to questionpro [9], the strength of association depends on the coefficient, when the coefficient is greater than 0.1, it is positive, and the greater the coefficient is, the stronger the strength of association is. On the other hand, if the number is smaller than -0.1, then it is negative and smaller the coefficient is, the less the association is. As a result, all three variables used in the process of modeling were analyzed and the specific values were listed in the Table. 1. Overall, the results of linear correlation show low dependency which means variables were all good to go as they were all smaller than 0.5, and the normalization process model again was the best to fit in the regression model.

Table 1. Correlation Analysis

	x2 standard	x3 standard	x4 standard
x2 standard	1		
x3 standard	0.015728279	1	
x4 standard	0.06672117	-0.584201475	1

3.2 Regression

3.2.1 Standardization model

As mentioned above in the part of methods, part of the data was selected to run into the model of standardization. Tables. 2, 3 and 4 summarize and list of the first run of model 1. To notice that the P-value of standard x_3 was highlighted, this was done since the P-value of x_3 is 3.18×1021, which is the distance nearest to the MTR Station, was too small. Therefore, it can be ignored as it will not make a huge impact on the y value and the whole relationship. Then, the dataset was re-run with house age, number of convenience stores, and the y value.

Table 2. Regression analysis for model 1

Regression statistics	
Multiple R	0.617068
R Square	0.380773
Adjusted R Square	0.376986
Standard Deviation	10.93889

Table 3. Variance analysis for model 1

	df	SS	MS	F	Significance F
Regression analysis	2	24060.85	12030.43	100.539	9.28E-35
Residual	327	39128.59	119.6593		
Total	329	63189.44			

Table 4. Model summary for model 1

	Coefficients	Standard error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	32.05264	1.400233	22.89093	6.83E-70	29.29804	34.80725
x2 standard	-12.6461	2.32625	-5.43627	1.07E-07	-17.2224	-8.06982
x4 standard	27.30081	2.068694	13.19712	3.57E-32	23.23118	31.37044

The x of Model one was processed by (x-minimum) divided by maximum, and the data were selected to run the model which we got r-squared as 0.380773. By the second run of the model, one may conclude that the relationship of this formula is:

$$y = 32.05264 + 27.30081 \times store - 12.6461 \times house\ age \quad (4)$$

Then one input the values into the rest of the 20 percent of the data, and the values of $\hat{y}$ and y were all gained, and the error propagation was known.

3.2.2 Interception

After the processing of the first model was done, we now have the interception model. The intercept in a basic regression model stands for the mean value of the response variable when all of the predictor variables in the model are equal to zero. The form is given as follows [10]:

$$\hat{y} = \beta 0 + \beta 1(x) \quad (5)$$

Therefore, for the interception regression model, x3×x4 was used as the x value where y value, the housing price, remains the same. Again, 80 percent of the data was selected to input into the model while the other 20 percent was used to test y hat and y. The regression analysis, variation analysis and summary of the models are presented in Tables 5, 6, and 7. The data was applied and the r-squared gained from the results is 0.042295. According to the relation formula, the relationship of this formula is

$$y = -96.9637 \times x3 \times x4$$

Table 5. Regression analysis for model 2

Regressions statistics	
Multiple R	0.205658
R Square	0.042295
Adjusted R Square	0.039384
Standard deviation	13.63014

Table 6. Variance analysis for model 2

	df	SS	MS	F	Significance F
Regression analysis	1	2699.323	2699.323	14.52962	0.000165
Residual	329	61121.84	185.7807		
Total	330	63821.16			

Table 7. Model summary for model 2

	Coefficients	Standard error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	40.9615	1.099094	37.26841	4.1E-120	38.79936	43.12364
x3*x4	-96.9637	25.43795	-3.81177	0.000165	-147.005	-46.9222

Table 8. Regression analysis for model 3

Regression statistics	
Multiple R	0.715306
R Square	0.511663
Adjusted R Square	0.507182
Standard deviation	9.762667

Table 9. Variance analysis for model 3

	df	SS	MS	F	Significance F
Regression analysis	3	32654.9	10884.97	114.2063	1.32E-50
Residual	327	31166.26	95.30967		
Total	330	63821.16			

Table 10. Model summary for model 3

	Coefficients	Standard error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	37.39242	0.552185	67.71722	1.8E-194	36.30613	38.4787
z-score(Standard Normal Distribution)x2	-3.66511	0.844285	-4.34108	1.89E-05	-5.32603	-2.0042
z-score(Standard Normal Distribution)x3	-5.59901	0.722888	-7.74533	1.21E-13	-7.02111	-4.17691
z-score(Standard Normal Distribution)x4	5.117698	0.755074	6.777744	5.7E-11	3.632282	6.603114

3.2.3 Normalization

The meaning of Normalization is to alter the values in the dataset down to a same scale without destroying the differences in the values' ranges. Finally, the normalization model was handed with z-score, in which the process was done (x-mu)/sigma, and y value remains the same. The same process of selecting data was done and the statistical summary of the results are demonstrated in Tables. 8, 9 and 10. According to the results, the r-squared of this model is 0.511663. The relationship of this formula is

$$y = -3.66511 \times x_1 - 5.59901 \times x_2 + 5.117698 \times x_3 \quad (6)$$

3.3 Comparison

The values of r-squared were 0.380773 for the standardize model , 0.042295 for interception model, and 0.511663 for model 3. As mentioned above, the evaluation standard for judging how good or bad the model fits the data is by observing the values of the R-squared, where the bigger the R-squared is close to one, the better the model fits into the data, and if it gets closer to zero, the worse it fits. From the worst to the best, model 2 was the poorest fit of the regression model as the r-squared needs to be big to show the well-fit of the model with the numbers, runners-up were model 1 and the champion is model 3, normalization.

4. Limitations & Prospects

In this paper, one of the main limitations is the lack of data. Not a lot of data on real estate were put online, hence it is hard to pick suitable data that may be run perfectly into the regression model. In the meantime, the data that was chosen does not have a clear indication of which country, or city it is, although one may suspect that it is in the United Kingdom yet there is a lack of information here from the word MTR, which might cause confusion and misleading. Moreover, the application used in this paper Microsoft Excel is not as powerful as other applications (e.g., Java, Python, C++ etcetera), which may potentially cause to have less information shown in the output analysis. In the future, it is hoped that more papers can focus on the prediction of house prices, so that this field of study may be more complete than current.

5. Conclusion

In conclusion, this paper investigates house price relations toward different factors based on regression model analysis. Specifically, three processing regression models are constructed with different relationships and formulas, and the best model was found to predict the housing price tendency. Granted, it is true that the application used in this paper is basic, yet all information required to find the main target of this paper was achieved. In the future, it is hoped that more and more researchers can focus on housing prices worldwide so that more people can save their time on finding a suitable household themselves as the relationship will show everything. Back to the main goal of this paper, the relationship between y and x was also analyzed, which is when house age is lower, the number of convenience stores is higher, and is an inverse relationship. When there are more convenient stores, the price of the house went up. Overall, the main target was achieved by the application and statistical results of the model.

References

[1] Most Expensive Housing Markets Globally 2020. *Statista*, Statista Research Department, 9 June 2022, Retrieved from: https://www.statista.com/statistics/1040698/most-expensive-property-markets-worldwide/.

[2] Topic: Global Housing Market. *Statista*, Statista Research Department, 2 June 2022, Retrieved from: https://www.statista.com/topics/5466/global-housing-market/#topicHeader__wrapper.

[3] Campisi Natalie. Will the Housing Market Crash? Experts Give 5-Year Predictions. *Forbes*, Forbes Magazine, 16 Aug. 2022, Retrieved from: https://www.forbes.com/advisor/mortgages/real-estate/will-housing-market-crash/.

[4] Liu Chang, and Wei Xiong. "China's Real Estate Market National Bureau of Economic Research." *Https://Www.nber.org/System/Files/working_papers/w25297/w25297.Pdf*, Nber.org, Nov. 2018, https://www.nber.org/system/files/working_papers/w25297/w25297.pdf.

[5] Hale Danielle, and Sabrina Speianu. 2022 Housing Market Predictions and Forecast. *Realtor.com Economic Research*, Realtor.com, 6 July 2022, Retrieved from: https://www.realtor.com/research/2022-national-housing-forecast-midyear-update/.

[6] Bruce, Algor. Real Estate Price Prediction. *Kaggle*, 8 Dec. 2018, Retrieved from: https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction.

[7] McLeod S. A. 2019, May 20. What a p-value tells you about statistical significance. Simply Psychology. Retrieved from: www.simplypsychology.org/p-value.html

[8] Team, Towards AI. "How, When, and Why Should You Normalize / Standardize / Rescale Your Data?" Retrieved from: *Towards AI*, 29 May 2020, Retrieved from: https://towardsai.net/p/data-science/how-when-and-why- should-you-normalize-standardize-rescale-your-data-3f083def38ff.

[9] "Pearson Correlation Coefficient: Free Examples." *QuestionPro*, 22 June 2022, Retrieved from: https://www.questionpro.com/blog/pearson-correlation-coefficient/.

[10] Zach. "How to Interpret the Intercept in a Regression Model (with Examples)." *Statology*, 22 June 2021, Retrieved from: https://www.statology.org/intercept-in-regression/.