House Price Analysis in Vancouver, Canada

Kuanjie Chen - 1005707610

4/14/2022

Introduction

A countries' standard of living is often reflected by housing availability. By investigating housing prices we look to make insightful conclusions on countries' standard of living - an important statistic that could be used in a wide variety of other statistical researches. Such statistic could also be used for personal gains; it could be useful to those who are considering relocation. This paper will investigate the estimation of property prices and try to answer the following question using a multi-variable linear regression model: How does measurements of land statistics affect property price? A quick search on the articles related to the proposed research question yielded 192,000 results on Google scholar; a common occurrence is observed whilst investigating these articles. An overwhelming majority of these articles uses predictors such as: number of bedrooms, room size, etc. to predict house price. Gupta's model, for example, suggests that room size depicts a strong relationship with house prices ((Manasa, Gupta, R., & Narahari, N. S. (2020))).

Methods

To achieve and validate our model we will accompany the following methods. We first start by separating our data into two randomly chosen dataset with equal number of variables; one is our training dataset while the other is our testing dataset. The training data will be used to create a number of potential models, and we will then fit these models with our testing data to identify the most valid model by comparing the adjusted R square, regression coefficient, predictor and model assumptions. What we are looking for is minimal deviations of these qualities between the model fit using the training data and the testing data.

We start by performing an exploratory data analysis (EDA) on the data, that is, investigating the distribution of all the predictor variables and the scatterplot of the response V.S each predictor. In the distribution we are looking for any skews and in the scatterplot we are looking for any obvious patterns.

We then fit all of our predictors in the initial model and immediately check condition 1 by observing the scatterplot of response (building price) V.S fitted response.To check condition 2 we plot the scatterplot between all of the predictors. The model is refit by adding/subtracting or transforming predictors if any one of the two conditions fail to hold.

If both conditions hold, we can start checking the assumptions by plotting the following three plots: residuals V.S predictor, residuals V.S fitted values plots and normal quantile-quantile (QQ) plots. If we see no discernible patterns in the first two of the aforementioned plot we can conclude that linearity, uncorrelated errors and constant variance assumptions hold. If we see that the points on the QQ plot follow a diagonal line with minimal deviations at both ends, we can conclude that the normality assumption holds.

When all four of these assumptions hold we calculate the VIF value to see if it is under 5; if it is over 5 we will again refit the model. We then conduct a T-test on all of the predictors to identify those which are significant; if there are any insignificant predictors we will remove said variable and use the F-test to determine whether the predictor removal was justified. Lastly, we will calculate the leverage statistic, cook's distance, DFFITS and DFBETAS in order to determine the outliers and conclude whether or not these outliers have negative impacts on the model.

Results

EDA Analysis

We begin to fit our model by first taking a look at their histograms and scatter with the response value.

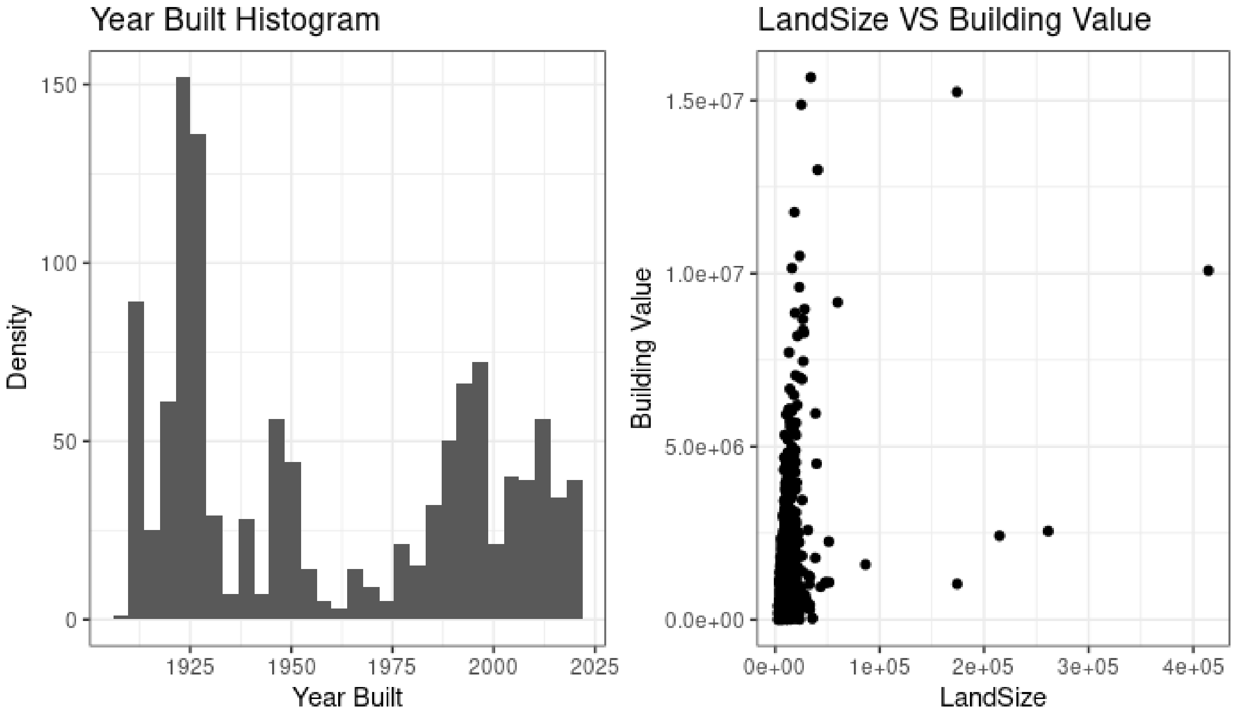


Figure 1: EDA Analysis

All histograms of predictors follow a unimodal pattern without any skews, however, a symmetrical pattern can be observed in the year built histogram; such pattern suggests that the normality assumption may be violated if we choose to keep this variable in the final model. The scatter plots between all predictors and the suggests all of the variables are linearly related to the response, and thus we can say they are all appropriate for our model. Plots and summaries of other variables can be found in the appendix.

Analysis Process & Results

Model 1

As aforementioned in the methods section, we begin our analysis by fitting a model with all of the variables as the response (Total_value, Land_value, Year_Built, Land_Size, PricePerSqFt); condition 1 was checked and a complete one to one pattern was found between the response and the fitted response. Moreover, the summary of the model gave an adjusted R square of 1. We conclude that, due to the results above, condition 1 fails and we start to refit the model. Since we had an adjusted R square of 1 and we are starting with all of the predictors, the only next step would to be remove one of the predictos. We found that removing the Land_Value predictor solves the one to one problem and gave the highest adjusted R square.

Model 2

Now we start to fit our second model without the Land_Value predictor; condition 1 was satisfied as the response V.S fitted response scatterplot depict a clean linear relationship. Condition 2 was also satisfied as all the scatterplots between each predictor all showed no significant pattern. By plotting the fitted response V.S residual, the scatterplot between the residual V.S all of the predictors we conclude that the linearty, uncorrelated error and constant variance assumptions all hold. However, the QQ plot had a non-diagonal line with major deviations at both ends. We attempt to fix this issue by arranging a transformation on the PricePerSqFt predictor; to be more specific we transformed the predictor to a power of 0.5 and refitted the model.

Model 3

The third iteration of the model is fitted as the following; we use Total_value, Year_Built, Land_Size, PricePerSqFt as predictors with the PricePerSqFt transformed as described in the section above. Condition 1 & 2 was satisfied and all of the assumptions hold. By calculating the VIF value we see two predictors with a VIF value over 5: Total_Value and Land_Size. By removing each of these predictors separately we get that removing the Land_Size predictor results in a higher adjusted R square, and thus we decide to remove it and refit the model.

Model 4

The fourth iteration of the model is fitted as the following; we use Total_value, Year_Built, PricePerSqFt as predictors with the PricePerSqFt transformed. Condition 1 & 2 and all of the assumptions hold. The VIF value of all predictors all just hover above 1 and thus we move on to check whether all predictors are significant. A simple T-test suggests that all of the predictors are significant. Since we did not remove any predictors here, we do not need to do a F-test. Finally, a calculation of the leverage statistic, cook's distance, DFFITS and DFBETAS suggests that there are some outliers and leverage points present in the data - the impact of which will be discussed in later sections.

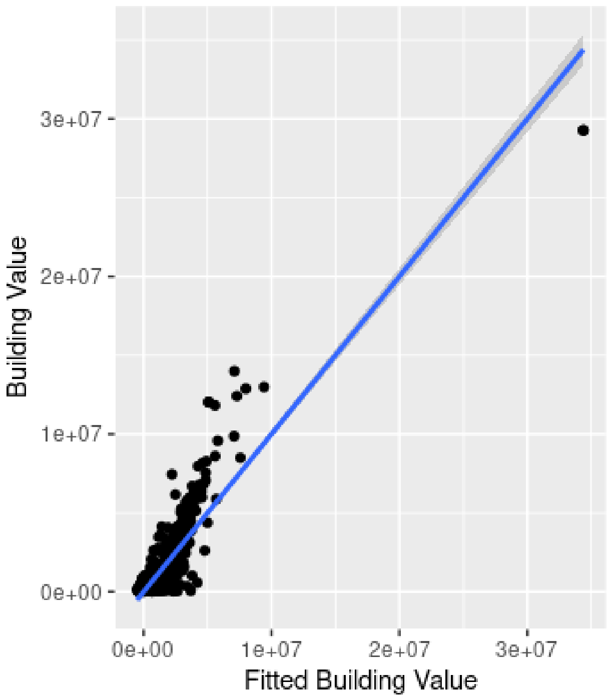
;-;	Adjusted R Square	Leverage Points	Outliers	Influential Points
Model 4	0.4897	54	92	112

Model Validation

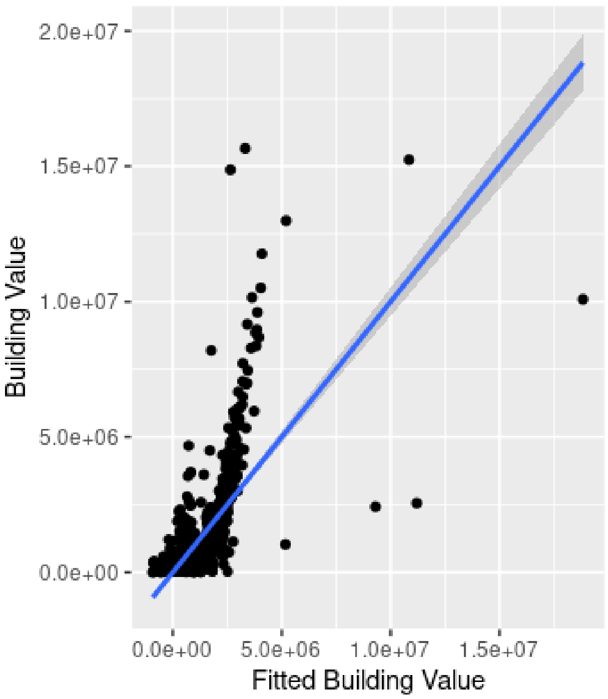
Now we perform our model validation by fitting model 4 with the test data. We see a general similarity between the regression coefficients between the model fitted using training data and the model using test data; thus we can conclude we are estimating a similar relationship. The T-test also suggests that all predictors are significant. All of the assumptions and conditions hold as no violations were observed. Finally, we see that although there are some differences in the adjusted R square it is very minimal. We can conclude, then, that our model is valid.

Dataset	Adjusted R Square	Coef 1 (Total_Value)	Coef 2 (Year_Built)	Coef 3 (PricePerSqFt)
Train	0.4897	1.192e-01	2.109e+04	-2.316e+03
Test	0.798	3.062e-01	1.538e+04	-3.154e+04

Response V.S Fitted Response (Test)



Response V.S Fitted Response (Train)



Discussion

Model Interpretation & Importance

In our final model (Model 4) we have 3 predictors: Total_Value, Year_Built and PricePerSqFt. Total value is total value of the property, that is, the sum of land and house price. Year built denote the year of which the house was built. PricePerSqFt denote the price per square feet of the house.

-.-	Intercept	Coef 1 (Total_Value)	Coef 2 (Year_Built)	Coef 3 (PricePerSqFt)
Final Model	-3.993e+07	1.192e-01	2.109e+04	-2.316e+03

Our intercept here gives little insight; it essentially says that when total value of property hits 0 we have that the house price is a negative value. Realistically this is never going to happen. We can, based off of the coefficients of the final model, then, make the following conclusions when the predictors are held constant; for every unit change in house price we can expect to observe: a 1.192e-01 unit change in total value, a 2.109e+04 unit change in year built and a -2.316e+03 unit change in price per square feet.

The total value interpretation gives us the most insightful information on house price; one could potentially use this statistic to estimate house price based on current or predicted trends using total value.

Limitation

While model 4 gives us a lot of insight, it has several flaws. As aforementioned in the introduction, due to our limited dataset we are not able to use most of the variables that established researches use in their model analysis. Although the diversity in our predictor selection leads to incongruous model interpretations it also results in potential skew in our final model. Moreover, a lot of problematic observations were present in our dataset; although we determined that these problematic observations would most likely not affect our model it could potentially lead to inappropriate model analysis.

Appendix

Variable	Mean	Standard Deviation
Total Value	6041561	6591431
Land Value	4939557	5868769
PricePerSqFt	441.847	123.955
Year Built	1958.905	37.131
Land Size	12392.48	17712.23

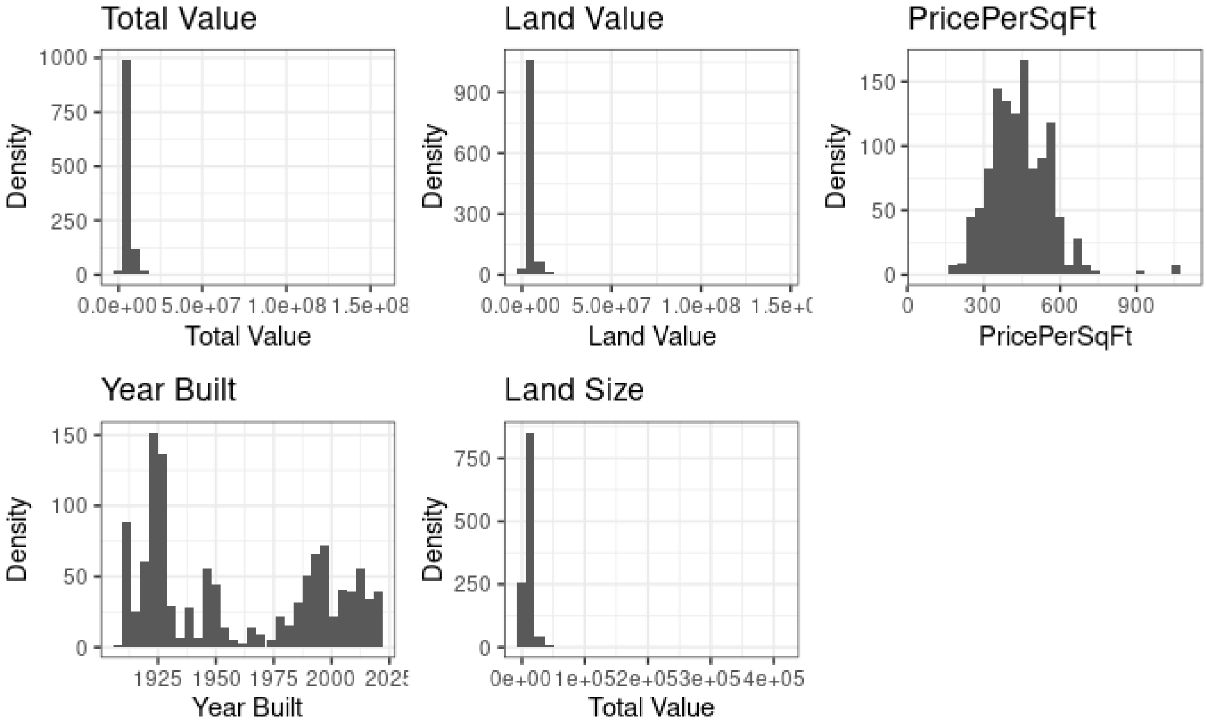


Figure 2: Variable Histograms

Bibliography

Madhuri, Anuradha, G., & Pujitha, M. V. (2019). House Price Prediction Using Regression Techniques: A Comparative Study. 2019 International Conference on Smart Structures and Systems (ICSSS), 1–5. https://doi.org/10.1109/ICSSS.2019.8882834

Varma, Sarma, A., Doshi, S., & Nair, R. (2018). House Price Prediction Using Machine Learning and Neural Networks. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 1936–1939. https://doi.org/10.1109/ICICCT.2018.8473231

Manasa, Gupta, R., & Narahari, N. S. (2020). Machine Learning based Predicting House Prices using Regression Techniques. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 624–630. https://doi.org/10.1109/ICIMIA48430.2020.9074952