Exploration of Factors that Influence the Final Sold Price of One-Family Dwelling in the City of Vancouver Between 2016-2019

Ivan Gao 81581944, Zunquan Wang 73127698

Abby Hong 99385726, Minghao Wang 56536469

12 Aug 2022

1 Introduction

1.1 Background and Motivation

Vancouver, a city located in British Columbia, Canada is ranked the 5th most livable city worldwide according to a recent published report by the EIU. It is well known for its beautiful environment, nature and prosperity. However, the housing price is one of the major issues that withholds its rank. One housing advocate says, "Cities are about bringing people together, but if people can't live here because of waiting lists, high rents, or high prices, then you don't have a city anymore, so I don't see how you can call it 'livable.'" (Charach, 2022)

In this project, we want to explore the factors that influence the final sold price of one-family dwelling in Vancouver.

1.2 Selection of Potential Variables

The goal of our project is to investigate different factors that may influence the housing price in Vancouver. Therefore our response variable is housing price. We choose six factors from various aspects as our predictor variables based on living experiences and common sense.

We suspect that the housing price is influenced by its previous price and tax levy as they enlarge the buyer's burden on the purchase and reduce their willingness to reach a decision. Tax increases every year and affects the housing price.

Building year and big improvement year influence the utility lines and other built-in equipment since early building may not be able to catch up with daily technology improvement.

1.3 Data Collection

Based on our target variables, we used the tax report published by the City of Vancouver (2021) at Open Data Portal.

In terms of data accuracy, it is dependent on the matching of records between multiple agencies including non-City sources.

2 Analysis

2.1 Preprocessing data

After retrieving the dataset, we first delete all private information to protect personal privacy.

We also removed the geographic information such as the land coordinate, legal address of the housing, and zoning district since it requires special software and theory and this is beyond the scope of this course.

After we remove the unprocessable columns, we then filter out all the NA rows since they are deemed as incomplete data.

We choose to combine the land value and improvement value as the response variable: final sold price. We also combined the previous land value and improvement value as one explanatory variable which is the previous housing price.

This leaves us with six explanatory variables and one response variable as shown in the table below.

Name	Description	Unit
Housing Price (y)	The sum of current land value and current improvement value.	CAD
Legal Type (x_1)	There are three different legal types(Strata/ Land/Other) in property.	/
Tax Assessment Year (x_2)	The year in effect for current land value, current improvement value and tax levy.	/
Year Built (x_3)	The year that the property is built	/
Previous Price (x_4)	The sum of land value and improvement value in the previous assessment year.	CAD
Big Improvement Year (x_5)	Year of major improvement to the property.	/
Tax Levy (x_6)	This is the total taxes printed on the most recent tax notice. It includes the City's general levy, levies for all taxing authorities, utilities, local improvements and miscellaneous charges.	CAD

Table 1: Table of response variable and explanatory variables used

2.2 Data Analysis

2.2.1 Preliminary Analysis of Data

To determine which model to use, we draw the scatter plot of two quantitative variables versus price to observe their relationship.

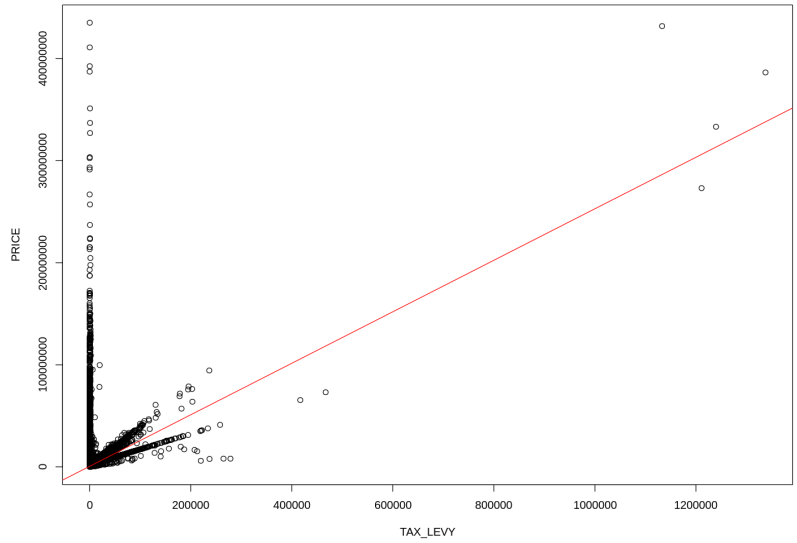

Figure1: Scatter plot of TAX_LEVY versus PRICE

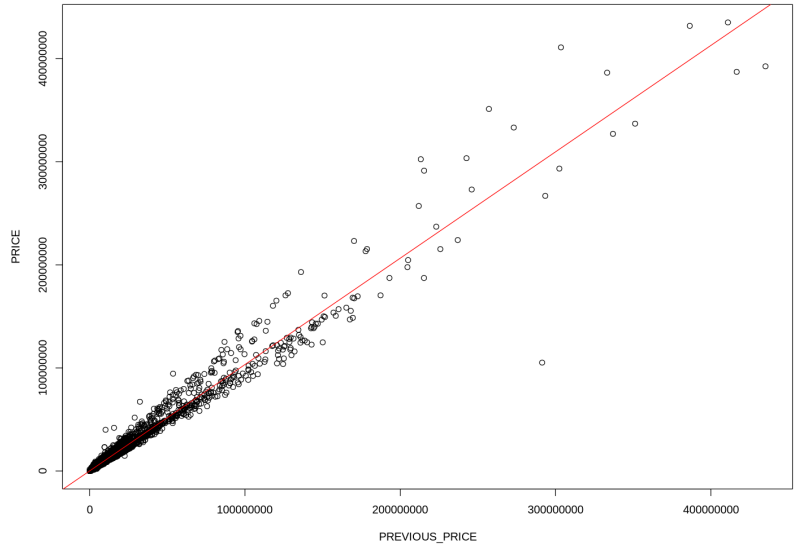

Figure 2: Scatter plot of PREVIOUS_PRICE versus PRICE

In Figure 1, despite the data cluster at TAX_LEVY zero, the overall pattern is still linear as indicated by the line of best fit. Figure 2 suggests that we have a strong linear relationship between PREVIOUS_PRICE and PRICE. Hence, a linear model seems reasonable. As a result, we decide to apply a simple multiple linear regression model to our data.

2.2.1 Model and Variable Selection

In order to perform model selection, we need to calculate the generalized variance inflation factor (GVIF) to detect potential collinearity problems. If GVIF > 5, then there is evidence that suggests a collinearity is significant.

	GVIF	Df	GVIF^(1/(2*Df))
	<dbl>	<dbl>	<dbl>
LEGAL_TYPE	1.0437091	2	1.0107526
YEAR_BUILT	2.1100937	1	1.4526161
BIG_IMPROVEMENT_YEAR	2.1540191	1	1.4676577
PREVIOUS_PRICE	1.2231207	1	1.1059479
TAX_LEVY	1.2019588	1	1.0963388
TAX_ASSESSMENT_YEAR	1.0249602	3	1.0041174

Table 2: GVIF of the data

Table 2 shows the GVIF value of each variable. Since all GVIF belows 3, therefore there is no obvious collinearity between each parameter and we can apply model selection for this dataset.

To select the model that best fit to our data, we used the stepwise selection method, where we iterate between forward and backward selection and validate its performance with k fold cross validation with k =10.

	nvmax	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	873180.56	0.967335981	337464.62	218953.78	0.015229662	6690.6774
2	2	826161.69	0.970877872	266237.33	227891.33	0.015256007	5040.2441
3	3	804008.90	0.972543931	223159.37	235220.01	0.015525202	5720.0828
4	4	2734438.25	0.501194785	725218.93	2059598.65	0.497314802	526947.1732
5	5	2682659.36	0.501626269	709701.12	2010243.10	0.497453139	528378.3257
6	6	4619133.38	0.031428403	1213665.81	283484.55	0.024313501	18088.0454
7	7	795860.73	0.972693572	206875.81	242723.47	0.016072415	4905.1104
8	8	795781.37	0.972698040	207033.56	242736.14	0.016072913	4884.6865

Table 3: performance table of nvmax number of predictors' models

Stepwise selection gives us the table 3 above. Within it, nvmax is the number of variables in the model. For example nvmax = 2, specify the best 2-variables model.

In order to find the model that fits the best, we choose three measurements to inspect closely: the RMSE and MAE, and Mallow's Cp.

RMSE is the root mean square error of the model and MAE is the mean absolute error of the model. The less, the better the model. From Table 3 we can conclude that line 8 with 8 predictors is the best fitted model since it has the least RMSE and MAE altogether when comparing to other predictors along the lines.

To determine which one is the best fit, we applied Mallow's Cp to assess their performance. The best fitted model always has Cp $\rightarrow$ p where p is the number of predictors. The Cp versus p plot is displayed below.

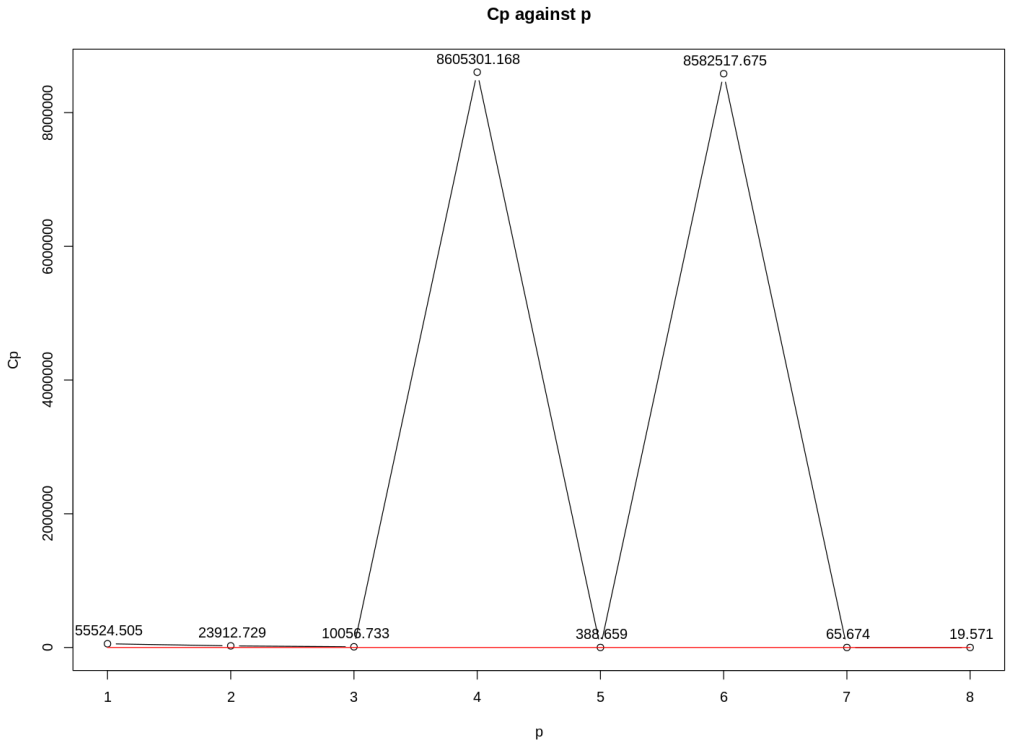

Figure 3: P versus CP graph for the eighth model

Red line is the line Cp=p. We can see that the model with the closest Cp is model 8 with Cp = 19.571. Hence, after careful examination, we choose model 8 as our best fitting model.

2.2.2 Model Validation and Criticism

To validate the model selected we need to check all assumptions of the model by examining three plots regarding the model.

To detect any error in measurement, we drew the index plot of standardized residuals with respect to the index (Figure 4 below). There is one value that lies far from others. After a close examination, this point is neighbourhood V6R 4L9, the Vancouver Jericho Sailing Centre. It is not a one-family dwelling. Hence this point must be an error data which requires removal.

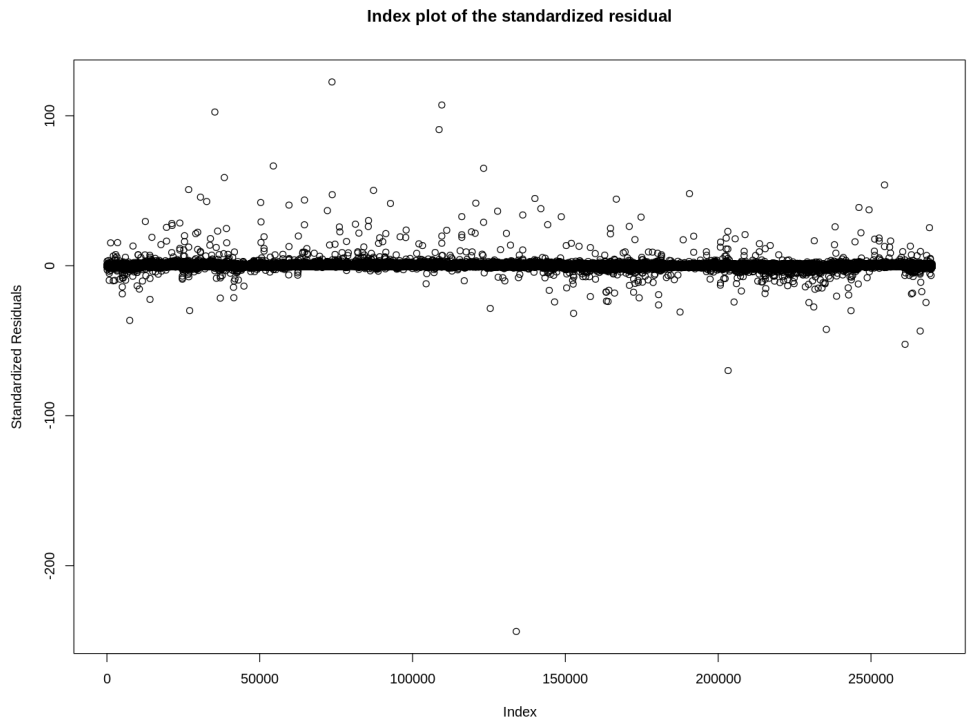

Figure 4: The index plot of standardized residual

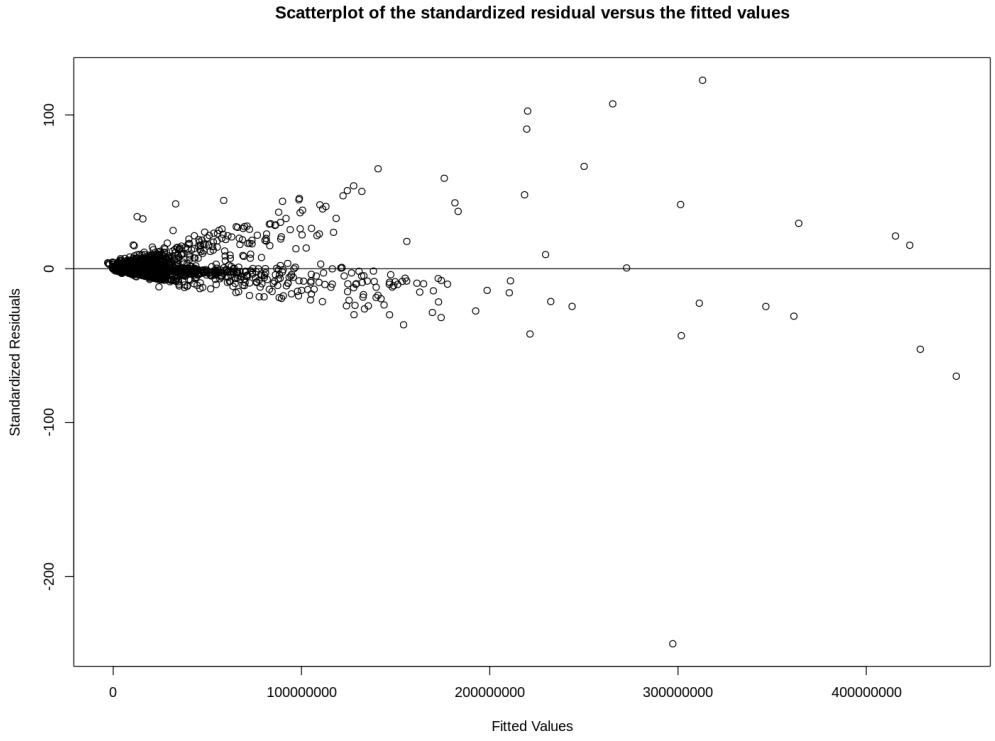

Figure 5: Scatterplot of the Standardized Residual vs Fitted Values

However, from Figure 5, the standardized residuals plot of the selected model shows increasing of value as fitted value increases. This could be a potential heteroscedasticity problem.

2.2.3 Heteroscedasticity Problem

In order to determine whether heteroscedasticity exists or not, the Breusch-Pagan test is applied (Breusch & Pagan, 1979). Since the p-value from the test is close to zero (p-value < 2.22e-16), the null hypothesis is rejected and this concludes that heteroscedasticity is present within this model. To address this issue, log transformation has been introduced along with weighted least square methods.

We apply log on all three quantitative variables: PRICE, response variable, TAX_LEVY, and PREVIOUS_PRICE, the predictor. Despite the fact that we treat YEAR_BUILT and BIG_IMPROVEMENT_YEAR as numeric integer variables, we can not take log on them as it doesn't make sense to have log(year).

The standardized residual plot of the model refitted after the log transformation on data through weighted least squares methods has been shown below (Figure 6).

By comparing this to the previous residual plot. The standardized residual of the model after transformation and weighted is much better as it shows no cone shape pattern.

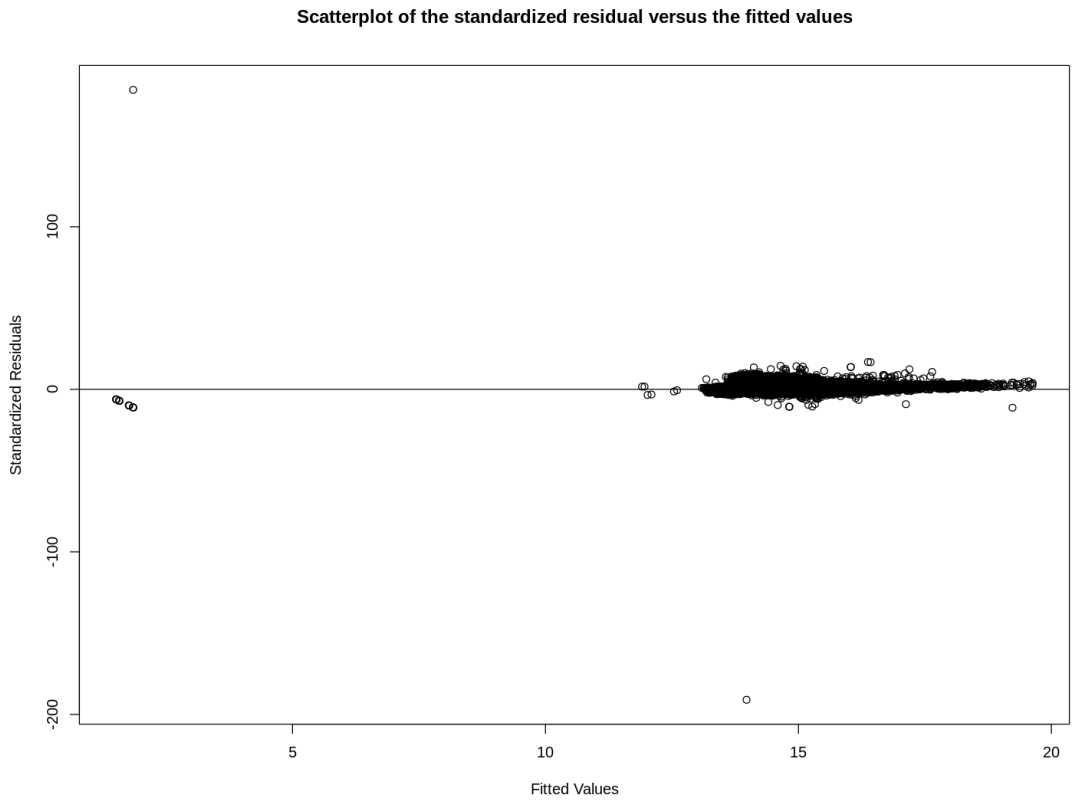

Figure 6 : Scatterplot of the Standardized Residual vs the Fitted Values

2.2.4 Outlier

Despite most standardized residuals clustering around line zero, some points still lie beyond the data cluster (Figure 6). They are two points with standardized residuals above and below 100, denoted by point 1 and point 2. A small group of points where fitted values are below 5 is called group A.

To analyze what makes them different. We retrieve the raw data of those points and carefully examine them individually. Some surprising factors show up.

We realize that point 1 and point 2 come from the same neighbourhood, V5P 2X3. This phenomenon could indicate some factors changes within this neighbourhood that make them

different from the rest of Vancouver. This change is unique and makes the house price in that region unique. Hence they are outliers and should be removed. This is the same for the points in group A as they have evenly distributed within four neighbourhoods: four points in V5N 3Z3, four points in V5V 2R9, four points in V6K 2J9.

Except for those points, to ensure there are no other influential points. We calculated the cook's distance to all data points and compared it with the threshold of 1 as suggested by the textbook (Collins, 2022). As a result, there are no other influential points detected.

2.2.5 Refit model after removing outliers

We noticed that after removing the outliers and refitting our model, the plot became significantly better than previous ones by examining the standardized residuals plot.

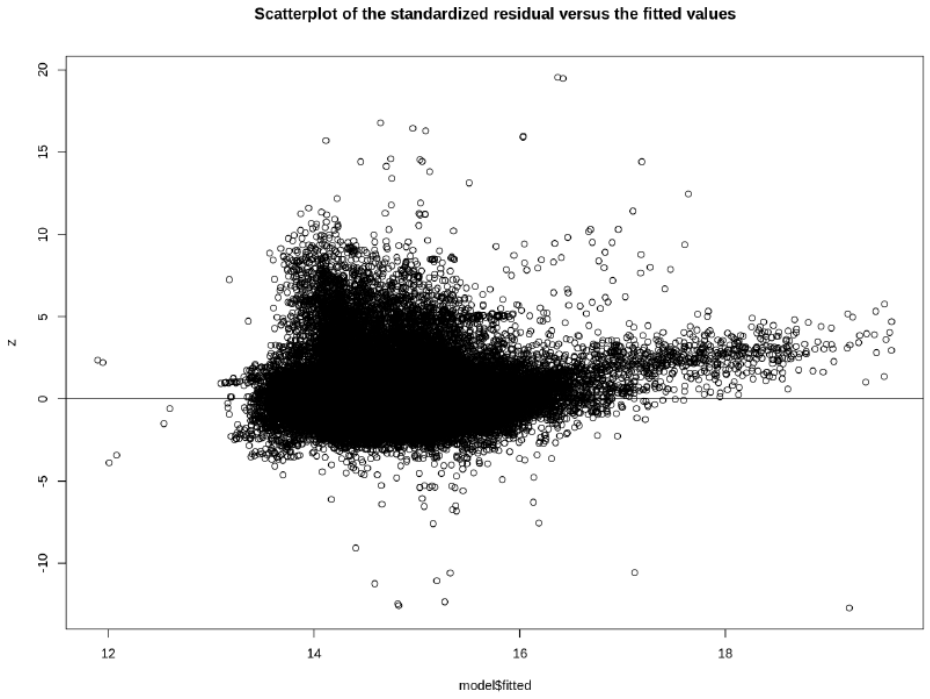

Figure 7 : Scatterplot of the Standardized Residual vs the Fitted Values after refit

By plotting standardized residuals against fitted values, we can observe that the residuals are clustering around 14-16 along model$fitted with a linear pattern extending toward the upper right. The existence of such a pattern suggests a potential undetected variable exists. One valid guess would be due to the lack of geographical indicators such as land coordinates or postal code since house price is highly related to its location and community. However, as mentioned in the introduction, due to the restriction of software and the lack of potential theory, we have to drop the discussion of such variables.

Also, to assess the quality of selected variables, we take the summary table of the model listed below (table 4).

```
Coefficients:
                            Estimate    Std. Error  t value  Pr(>|t|)
(Intercept)              0.9906636614  0.0106696230   92.849 < 2.2e-16 ***
LEGAL_TYPETRUE          -0.0576548943  0.0056564577  -10.193 < 2.2e-16 ***
TAX_ASSESSMENT_YEAR2017  0.0781986452  0.0003064382  255.186 < 2.2e-16 ***
TAX_ASSESSMENT_YEAR2018 -0.1950635593  0.0003398046 -574.046 < 2.2e-16 ***
TAX_ASSESSMENT_YEAR2019 -0.2516817018  0.0003324097 -757.143 < 2.2e-16 ***
YEAR_BUILT               0.0000577395  0.0000050518   11.430 < 2.2e-16 ***
BIG_IMPROVEMENT_YEAR    -0.0002083789  0.0000077930  -26.739 < 2.2e-16 ***
TAX_LEVY                 0.0047138600  0.0003393165   13.892 < 2.2e-16 ***
PREVIOUS_PRICE           0.9629297044  0.0003544509 2716.680 < 2.2e-16 ***
---
```

Table 4 : Summary Table of the Final Model

As shown above, the p value for each parameter in the final model is small and close to zero. Therefore there is enough evidence to reject the null hypothesis and conclude that each parameter is significant to the final sold housing prices.

Hence, here we conclude that our selected mode is the final model based on the current observation dataset.

2.2.6 Final Model

Our final model after refit data without outlier is:

$$\begin{aligned} Log(y) =& 0.9906636614 - 0.0576548943 * x_1 + 0.0781986452 * x_2 - 0.1950635593 * x_3 \\ & - 0.2516817018 * x_4 + 0.00005773 * x_5 - 0.0002083789 * x_6 \\ & + 0.0047138600 * Log(x_7) + 0.9629297044 * Log(x_8) \end{aligned} \quad (1)$$

$$\begin{aligned} y =& e^{0.9906636614} e^{-0.0576548943 * x_1} e^{0.0781986452 * x_2} e^{-0.1950635593 * x_3} e^{-0.2516817018 * x_4} e^{0.00005773 * x_5} \\ & e^{-0.0002083789 * x_6} * (x_7^{0.0047138600}) * (x_8^{0.9629297044}) \end{aligned} \quad (2)$$

Where,

$x_1 = LEGAL\_TYPE_{true}$, $x_2 = TAX\_ASSESSMENT\_YEAR_{2017}$

$x_3 = TAX\_ASSESSMENT\_YEAR_{2018}$, $x_4 = TAX\_ASSESSMENT\_YEAR_{2019}$

$x_5 = YEAR\_BUILT$, $x_6 = BIG\_IMPROVEMENT\_YEAR$

$x_7 = TAX\_LEVY$, $x_8 = PREVIOUS\_PRICE$

Notice that, before model selection there are three levels for legal type variables, they are strata, land, and other. After model selection we determined that it does not matter if legal type equals land and strata. All it matters is that legal type is equal to others. Hence, we transformed this categorical variable into a binary variable where true represents legal type is others and false otherwise. This gives us one dummy variable called LEGAL_TYPE_TRUE with a baseline at legal type not equal to others.

Also, x_2, x_3, x_4 are three dummy variables for a four level categorical variable TAX ASSESSMENT YEAR with baseline 2016.

3 Discussion

We want to extract the value of y hence, we applied the e(natural log) to both sides of the equation and found out that the PREVIOUS_PRICE is the most influential predictor on the final price. We discovered that the value of the factors cannot be less than one. This may happen because we did not include the geographical variable because the method that will be used is beyond the scope of this course.

By removing each predictor variable, we observed that the PREVIOUS_PRICE caused the maximum change in R-square value. This suggests that PREVIOUS_PRICE is the most influential predictor among the rest.

Because all the coefficients exist as a power of exponential and are smaller than 1, the marginal loss of all predictors, except x_8, for 1 unit change is significant. However, x_8 has a coefficient of 0.96 as power hence it loses the least as it changes. For a 20k value of the previous price, it will respond to the house price by 19k CAD. Others are very small compared to x_8 in terms of their influence.

4 Conclusion

Throughout our model selection, we have used different methods to determine the best fitting model. By obtaining the values of RMSE and MAE, and Mallow's Cp, we selected the model with eight explanatory variables including two dummy variables. The R squared value we obtained is 0.98425 which indicates a very good fit. The summary result of our model suggests that predictors are significant, which means they are important to the final price.

There are several limitations for our data analysis. Just as discussed above, our data analysis may contain missing variables such as geographic region, which could influence our response variable tremendously. On top of that, due to our limited knowledge on some fields and with the lack of enhanced data analysing skills such as using the time theory method, some techniques are not being performed and explained. With these, we conclude that this report demonstrates a reasonable and well fitted model with a good explanation of the current data. More theory and advanced software is needed for future development.

References

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, 1287-1294.

Charach, K. (2022, June 24). *New Report ranks Vancouver as 5th most livable city in the world*. iheartradio. Retrieved from https://www.iheartradio.ca/virginradio/new-report-ranks-vancouver-as-5th-most-livable-city-in-the-world-1.18073195

City of Vancouver. (2021, May 8). *Property tax report 2016-2019,* Retrieved from https://opendata.vancouver.ca/explore/dataset/property-tax-report-2016-2019

Collins, G. (2022, May 28). *Vancouver housing market predictions 2021: 2022 home price outlook Houses Condos BC Canada*. Housing Market and Stock Market Forecasts. Retrieved August 2, 2022, from https://gordcollins.com/real-estate/greater-vancouver-housing/

Kassambara, Gorenc, J., Priya, & Visitor. (2018, March 11). *Stepwise regression essentials in R*. STHDA. Retrieved from http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/