Analysing the sale price of houses in GTA

Saksham Ahluwalia

December 7, 2020

Code supporting this analysis is available at: https://github.com/sakshamahluwalia/sale-price-of-houses

Key words: MLR, Stepwise BIC, BIC, Backward BIC, Observational Study, Real Estate.

Real estate should be relevant to everyone because it is considered a good investment and we all need a place to stay. For quite some time now, there has been a growing concern that real - estate prices in the Greater Toronto Region (GTA) are rising at exponential rates. This is a cause of concern for anyone who is looking to buy a house. Having a model that estimates the sale price of a house can help buyers in their decision of making their purchase. One way to come up with a model is to apply MLR. It is better suited for this situation compared to regular SLR since there can be multiple variables that affect the sales price of a house. In this report we focus on detached houses from two neighborhoods - Toronto and Mississauga. A MLR model was used to check if and how the sale price of a detached house is linked to other variables such as the last list price, number of bedrooms/bathrooms etc. This is done using a stepwise BIC backward variable selection procedure. The relevant variables are then used to fit a model for sale price in GTA. In the Methodology section, I describe the data, and the model that was used in this study. Results of the model are provided in the Results section, and inferences of this data along with conclusions are presented in the Conclusion section.

DATA

The Data used in this study was obtained from the Toronto Real Estate Board. It includes the following variables:

- ID: id corresponding to the property
- Sale: actual sale price in CAD
- List: last list price in CAD
- Bedroom: number of bedrooms
- Bathroom: number of bathrooms
- Parking: number of parking spaces
- Maxsqfoot: maximum square footage of the property
- Taxes: Previous year's taxes
- Lot-size: area in feet
- Location: Toronto (T) or Mississauga (M)

Note: There are a lot of NULL/missing values for the **maxsqfoot** variable. This variable has been be dropped from the data set. I will also remove other cases containing NULL/missing values. This leaves us with 181 unique cases.

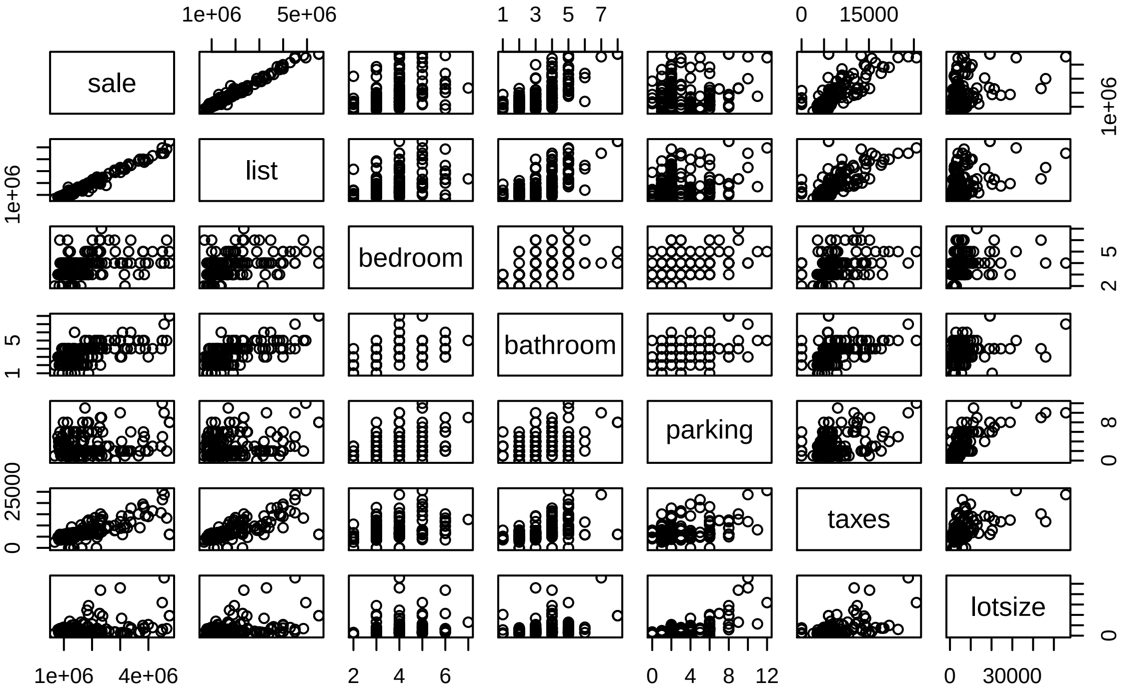
Table 1 includes some baseline characteristics of the variables in our data.

Table 1: Table 1 (3058)

Parameter	Data Type	Mean	Sd[s]	Frequency
ID	Discrete	-	-	-
bedroom	Discrete	3.727	0.995	-
bathroom	Discrete	3.399	1.249	-
parking	Discrete	3.295	2.386	-
sale	Continuous	1827446	970866.6	-
list	Continuous	1834405	1054994	-
taxes	Continuous	7305.584	4312.684	-
lotsize	Continuous	6080.673	7449.693	-
location	Categorical	-	-	T: 0.625 M: 0.375

Pairwise Correlation table Using the scatterplot Matrix we can see: that there is a strong positive linear trend between sale price and list price. There is also a strong positive linear relationship between list price and taxes. Correlation among other variables is not clear from the scatterplot matrix.

Fig 2 Scatterplot Matrix (3058)



MODEL

An additive multiple linear regression was used with Sale price being the response variable. The model can be defined as follows:

$$y_F = 52752.3036 + 0.8253x_1 + 13907.1387x_2 + 11139.0194x_3 - 16396.9727x_4 + 1.7465x_5 + 21.9112x_6 + 92139.8946x_7$$

The predictor variables in the above equation are defined as follows:
x_1: The last list price of the property in Canadian dollars.
x_2: The total number of bedrooms.
x_3: The total number of bathrooms.
x_4: The total number of parking spots.
x_5: New variable Lot Size.
x_6: Previous year's property tax.
x_7: Located in Toronto.

y_F **(response variable)**: Sale price

BIC Backward elimination To make the above model simpler a BIC backward selection is performed. Backward elimination procedure starts with all the variables in the model. At each step, the variable with the highest p-value from the individual T test is removed until all variables have been deleted from the model or the information criterion increases [1, pg 236]. We choose BIC backward selection procedure because it favors simpler models compared to the AIC forward selection which can cause overfitting.

After performing BIC backward elimination on the additive linear regression model stated above we get the following model:

$$y_{BIC} = 64056.020 + 0.835x_1 + 21.587x_2 + 125438.7143x_3$$

The variables in the above equation are defined as follows:
x_1: The last list price of the property in Canadian dollars.
x_2: Previous year's property tax.
x_3: Located in Toronto.

Furthermore, two different models are fit to find the expected sale price of detached houses in each neighborhood. The model representing the sale price of detached houses in Toronto is defined below:

$$y_T = 173216.651 + 0.8299x_1 + 25.2149x_2$$

Similarly the model representing the sale price of detached houses in Mississauga is defined as:

$$y_M = 91206.3519 + 0.8319x_1 + 18.5417x_2$$

The variables in the above equation are defined as follows:
x_1: The last list price of the property in Canadian dollars.
x_2: Previous year's property tax.

Results

Our Global F-test is significant at 95% level which suggests that there are one or many predictor variables which can be used to estimate Sale price of a property in our data set. Looking at the individual T tests of List price, Parking and Taxes we can see that they are significant at 95% and therefore using these variables in our model is appropriate. The rest can be dropped or ignored because of a non-significant T test at a 95% threshold.

Below in Table 2 we can see the estimated regression coefficients and the p-values for the corresponding t-tests for the coefficients expected by the additive model mentioned above.

Table 2: Table 2 (3058)

	Estimate	t value	Pr(>|t|)
(Intercept)	52752.3036	1.1019	0.2720
list	0.8253	43.5803	0.0000
bedroom	13907.1387	1.1345	0.2582
bathroom	11139.0194	0.9430	0.3470
parking	-16396.9727	-2.1401	0.0338
taxes	21.9112	5.7256	0.0000
lotsize	1.7465	0.8473	0.3980
locationT	92139.8946	2.7472	0.0066

Using the above mentioned significant predictor variables we fit two different models one for each neighborhood. From the individual models we can see that there is a higher tax rate observed in Toronto compared to Mississauga. Ignoring list price and taxes there is a difference of approximately 82,000 CAD in between the two neighborhoods. To conclude, according to our sample and models detached houses are more expensive in Toronto compared to Mississauga.

Discussions

With real-estate prices soaring in GTA tools are needed to assess the value of properties. Using a BIC backward selection procedure to select variables and data obtained from TREB a MLR model is fit to find the expected Sales price of a detached house in two different neighborhoods Toronto and Mississauga. We found that there is a difference in the expected price in the two neighborhoods. We found that houses in Toronto are on average more expensive than houses in Mississauga. The tax rate also seems to be more in Toronto compared to Mississauga. These results can be used by individuals looking to buy a detached house in either Toronto and Mississauga to evaluate the value of houses they are interested in. For example we saw that detached houses in Toronto can appreciate faster overtime and therefore make prime investment opportunities.

Weaknesses Only looks at 2 neighborhoods within GTA. We also had a limited amount of variables to work with. For example we did not have square footage or crime rate for the neighborhood which can be lurking variables. [2]

Next Steps Next steps can include performing some more analyses:

- We should perform Cross validation on data from a different year.
- We look at diagnostic plots to look for any violations in our assumptions.
- We can extend the study to different neighborhoods.
- A pooled two-sample t-test can be used in this scenario to determine if there is a statistically significant difference between the slopes of the simple linear models for the two neighborhoods. This is because our X variable (location) is categorical with two levels rather than quantitative.
- We can look into lurking variables like crime rate and square footage.

References

- Sheather, S. J. (2009). A modern approach to regression with R. New York: Springer. doi:https://doi.org/10.1007/978-0-387-09608-7 [1]

- https://upside.com.au/articles/selling-your-property/selling-guide/9-surprising-factors-affect-home-value [2]