A Multiple Linear Model for Toronto and Mississauga House Prices

Tanyilan, Id 1004701548

December 5, 2020

I. Data Wrangling

1.1 Sampling Data

We randomly selected 150 cases from the raw data. The id of the cases are

```
##   [1]  51  99 114 131  60 158  17 133  30  92 145 142 150 148 118 117   4 115
##  [19]  64  70  93  15  53  22  83 140 151 119 144 107  39 149 143 146 103  82
##  [37] 106  76   6  55   7 105  85  33  84  80 153 101  87  54  36  10 111  91
##  [55]  68 108  58  61  98 122  79  74  97   1 147 116  77 152 113  37 159  24
##  [73]  63  96   9  88  40  14  62 154 100  71  49   3 136  27  47  56  50  73
##  [91]  72  19  31  81  67 137  35  21  69  13  16  95 160 135 156  32 112  28
## [109]   8  89  90  86 110 157  75  57 102 126 134 138  65  48  66 141  45  11
## [127] 125  78   5   2 155  52 104 132 139  25 109  43  38  44  42  29  46  20
## [145]  94  23  34  41  12  26
```

1.2 Data Cleaning

Independent variable 'maxsqfoot' is removed because

- We don't need two variables representing the size of the property. 'lotsize'is another area variable.
- We choose 'lotsize' instead of 'maxsqfoot' because there are 98 missing values for 'maxsqfoot'.

Also, we remove 11 data sets containing 'na' (missing values).

II. Exploratory Data Analysis

2.1 Classify Variables

Data summary

Name	data
Number of rows	139
Number of columns	9
______________________	
Column type frequency:	
numeric	9
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1	78.29	47.72	1.00	36.5	75	117.5	160	▇▇▇▆▇
sale	0	1	1934252.25	965318.33	672000.00	1249500.0	1635000	2255000.0	4750000	▇▅▁▂▁
list	0	1	1932823.24	1047772.11	679900.00	1172000.0	1599000	2299000.0	5499000	▇▃▁▁▁
bedroom	0	1	3.71	1.02	2.00	3.0	4	4.0	6	▁▇▇▂▁
bathroom	0	1	3.31	1.26	1.00	2.0	3	4.0	8	▇▃▇▁▁
parking	0	1	2.85	2.33	0.00	1.0	2	4.0	12	▇▁▁▁▁
taxes	0	1	7362.69	4397.00	4.46	4741.0	6320	8646.5	25575	▅▇▁▁▁
location	0	1	0.78	0.41	0.00	1.0	1	1.0	1	▁▁▁▁▇
lotsize	0	1	5634.01	7936.96	675.68	2299.5	3060	6000.0	55756	▇▁▁▁▁

- Categorical Variable(s): location
- Discrete Variable(s): Number of bedrooms(bedroom), Number of bathrooms(bathroom), Number of parking spots(parking), Maximum square footage(maxsqfoot, removed)
- Continuous Variables: Sale price(sale), List price(list), Property tax(taxes), Lot size(lotsize), Frontage(lotwidth, removed), Length(lotlength, removed)

2.2 Correlation Matrix

TABLE 2.1: Correlation Coefficient Matrix

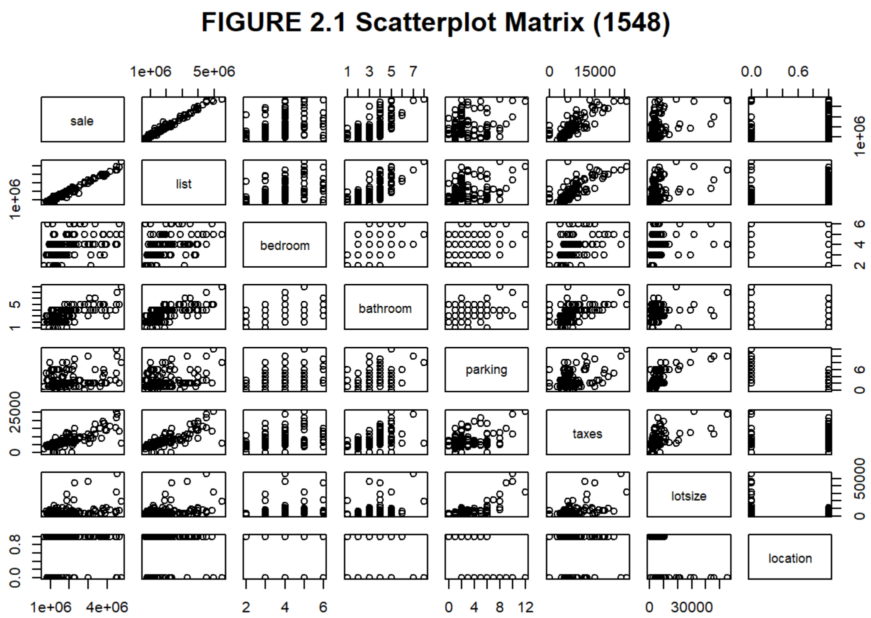
	sale	list	bedroom	bathroom	taxes	parking	lotsize	location
sale	1							
list	0.9874	1						
bedroom	0.4395	0.4436	1					
bathroom	0.6788	0.6955	0.539	1				
taxes	0.8087	0.7926	0.3931	0.5251	1			
parking	0.2486	0.2874	0.3543	0.3654	0.4287	1		
lotsize	0.3714	0.3873	0.2726	0.3175	0.5463	0.741	1	
location	0.0948	0.0585	-0.1491	-0.163	-0.1291	-0.771	-0.5589	1

Highest to Lowest correlation coefficient with 'sale price':

1. Last list price (list): 0.9874,
2. Property tax (taxes): 0.8087,
3. Number of bathrooms (bathroom): 0.6788,

4. Number of bedrooms (bedroom): 0.4395,
5. Lotsize (lotsize): 0.3714,
6. Number of parking spots (parking): 0.2486
7. Location of the property (location): 0.0948

2.2 Scatterplot Matrix

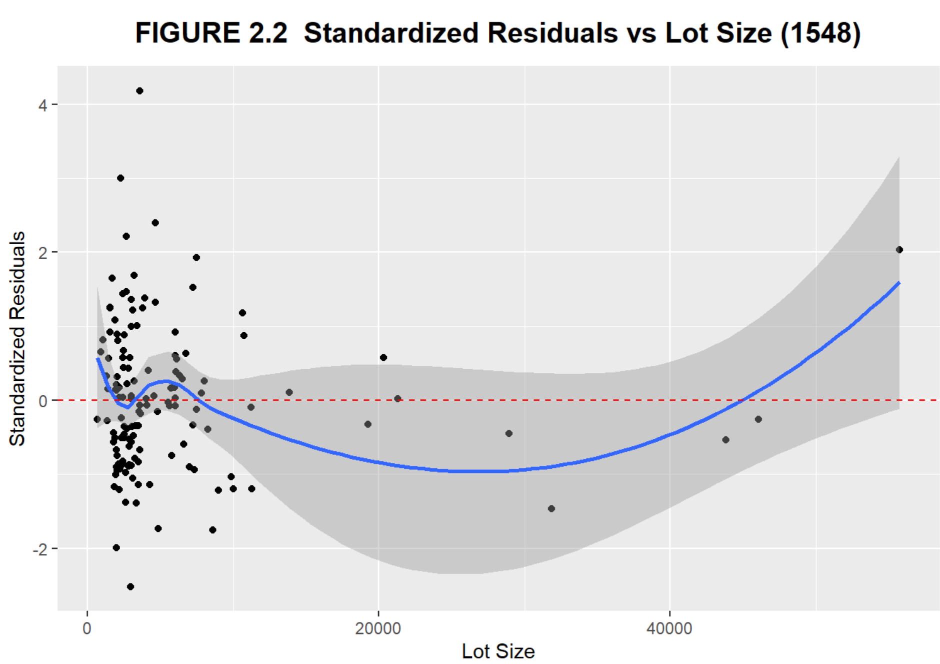


2.2.1 Violate The Constant Variance Assumption

Take a look at the first row except the dummy variable “location”; the scatter plots are approximately positively related except for the “lotsize”. Points on the last graph are centered at the left-bottom corner.

Thus, we guess that the data of ‘lotsize’ violate the constant variance assumption.

2.2.2 The Standardized Residuals Plot of ‘lotsize’



The residuals do not roughly form a horizontal band around the zero line (red), suggesting that the error terms’ variance are not constant. The conclusion proves that the constant variance assumption is not satisfied.

III. Methods and Model

3.1 Fitted Linear Regression Model

3.1.1 Summary Table

```
## 
## Call:
## lm(formula = sale ~ list + bedroom + bathroom + parking + taxes + 
##     lotsize + location, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -346623 -100571  -10163   76779  570434 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.004e+05  6.899e+04   1.456   0.1478    
## list         8.325e-01  2.423e-02  34.353  < 2e-16 ***
## bedroom      1.194e+04  1.432e+04   0.834   0.4060    
## bathroom     7.379e+03  1.470e+04   0.502   0.6166    
## parking     -2.177e+04  1.066e+04  -2.041   0.0432 *  
## taxes        2.228e+01  4.942e+00   4.509 1.43e-05 ***
## lotsize      1.509e+00  2.443e+00   0.618   0.5377    
## location     5.807e+04  5.209e+04   1.115   0.2670    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 139300 on 131 degrees of freedom
## Multiple R-squared:  0.9802, Adjusted R-squared:  0.9792 
## F-statistic: 928.1 on 7 and 131 DF,  p-value: < 2.2e-16
```

TABLE 2.2: Summary of the Linear Regression Model Coefficient

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	100436.1181	68988.8890	1.4558	0.1478
list	0.8325	0.0242	34.3528	0.0000
bedroom	11935.7999	14318.8585	0.8336	0.4060
bathroom	7379.2088	14702.6122	0.5019	0.6166
parking	-21769.6909	10664.8601	-2.0413	0.0432
taxes	22.2828	4.9416	4.5093	0.0000
lotsize	1.5093	2.4429	0.6179	0.5377
location	58072.3904	52089.9799	1.1148	0.2670

3.1.2 Fitted Multiple Linear Regression Model (Full)

We get the full model given by:

$$sale = 100436.12 + 0.83 \times list + 11935.80 \times bedroom + 7379.21 \times bathroom - 21769.70 \times parking + 22.28 \times taxes + 1.51 \times lotsize + 58072.40$$

Note: location = 1 if the property is in Toronto, otherwise 0.

3.1.3 Significance of Variables

List price, parking spots, and property taxes are significant because the p-values of the three variables are smaller than the significance level 0.05. We are able to reject the null hypothesis that the coefficients of these variables are zero.

Keep all other variables constant, as the list price increased by 1, the property's average sale price will increase by 0.83. Keep all other variables constant, as the parking spots increased by 1, the property's average sale price will decrease by 21769.7. Keep all other variables constant, as the property tax increased by 1, the property's average sale price will increase by 22.28.

3.2 Find A Parsimonious Model (Backward, AIC)

```
## Start:  AIC=3300.48
## sale ~ list + bedroom + bathroom + parking + taxes + lotsize +
##     location
##
##             Df  Sum of Sq        RSS    AIC
## - bathroom  1 4.8875e+09 2.5466e+12 3298.8
## - lotsize   1 7.4068e+09 2.5491e+12 3298.9
## - bedroom   1 1.3482e+10 2.5552e+12 3299.2
## - location  1 2.4115e+10 2.5658e+12 3299.8
## <none>                   2.5417e+12 3300.5
## - parking   1 8.0844e+10 2.6225e+12 3302.8
## - taxes     1 3.9452e+11 2.9362e+12 3318.5
## - list      1 2.2897e+13 2.5439e+13 3618.7
##
## Step:  AIC=3298.75
## sale ~ list + bedroom + parking + taxes + lotsize + location
##
##             Df  Sum of Sq        RSS    AIC
## - lotsize   1 6.2773e+09 2.5529e+12 3297.1
## - location  1 2.0890e+10 2.5675e+12 3297.9
## - bedroom   1 2.1321e+10 2.5679e+12 3297.9
## <none>                   2.5466e+12 3298.8
## - parking   1 7.9869e+10 2.6265e+12 3301.0
## - taxes     1 3.8978e+11 2.9364e+12 3316.5
## - list      1 3.3281e+13 3.5827e+13 3664.3
##
## Step:  AIC=3297.09
## sale ~ list + bedroom + parking + taxes + location
##
##             Df  Sum of Sq        RSS    AIC
## - location  1 1.8672e+10 2.5715e+12 3296.1
## - bedroom   1 1.9381e+10 2.5722e+12 3296.1
## <none>                   2.5529e+12 3297.1
## - parking   1 7.5274e+10 2.6281e+12 3299.1
## - taxes     1 4.5146e+11 3.0043e+12 3317.7
## - list      1 3.3325e+13 3.5878e+13 3662.5
##
## Step:  AIC=3296.11
## sale ~ list + bedroom + parking + taxes
##
##            Df  Sum of Sq        RSS    AIC
## - bedroom  1 2.0970e+10 2.5925e+12 3295.2
## <none>                  2.5715e+12 3296.1
## - parking  1 4.0670e+11 2.9782e+12 3314.5
## - taxes    1 4.4992e+11 3.0215e+12 3316.5
## - list     1 3.7097e+13 3.9669e+13 3674.4
##
## Step:  AIC=3295.24
## sale ~ list + parking + taxes
##
##            Df  Sum of Sq        RSS    AIC
## <none>                  2.5925e+12 3295.2
## - parking  1 3.8605e+11 2.9786e+12 3312.5
## - taxes    1 4.4687e+11 3.0394e+12 3315.3
## - list     1 4.0391e+13 4.2984e+13 3683.6
```

```
##
## Call:
## lm(formula = sale ~ list + parking + taxes, data = data)
##
## Coefficients:
## (Intercept)         list      parking        taxes
##   1.954e+05    8.507e-01   -2.520e+04    2.260e+01
```

We get the final model according to the AIC values using backward regression is given by

$$sale = 195400 + 0.85 \times list - 25200.70 \times parking + 22.6 \times taxes$$

3.3 Find A Parsimonious Model (Backward, BIC)

```
## Start:  AIC=3323.96
## sale ~ list + bedroom + bathroom + parking + taxes + lotsize + 
##     location
## 
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 4.8875e+09 2.5466e+12 3319.3
## - lotsize   1 7.4068e+09 2.5491e+12 3319.4
## - bedroom   1 1.3482e+10 2.5552e+12 3319.8
## - location  1 2.4115e+10 2.5658e+12 3320.3
## - parking   1 8.0844e+10 2.6225e+12 3323.4
## <none>                   2.5417e+12 3324.0
## - taxes     1 3.9452e+11 2.9362e+12 3339.1
## - list      1 2.2897e+13 2.5439e+13 3639.2
## 
## Step:  AIC=3319.29
## sale ~ list + bedroom + parking + taxes + lotsize + location
## 
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 6.2773e+09 2.5529e+12 3314.7
## - location  1 2.0890e+10 2.5675e+12 3315.5
## - bedroom   1 2.1321e+10 2.5679e+12 3315.5
## - parking   1 7.9869e+10 2.6265e+12 3318.7
## <none>                   2.5466e+12 3319.3
## - taxes     1 3.8978e+11 2.9364e+12 3334.2
## - list      1 3.3281e+13 3.5827e+13 3681.9
## 
## Step:  AIC=3314.7
## sale ~ list + bedroom + parking + taxes + location
## 
##            Df  Sum of Sq        RSS    AIC
## - location  1 1.8672e+10 2.5715e+12 3310.8
## - bedroom   1 1.9381e+10 2.5722e+12 3310.8
## - parking   1 7.5274e+10 2.6281e+12 3313.8
## <none>                   2.5529e+12 3314.7
## - taxes     1 4.5146e+11 3.0043e+12 3332.4
## - list      1 3.3325e+13 3.5878e+13 3677.1
## 
## Step:  AIC=3310.78
## sale ~ list + bedroom + parking + taxes
## 
##           Df  Sum of Sq        RSS    AIC
## - bedroom  1 2.0970e+10 2.5925e+12 3307.0
## <none>                  2.5715e+12 3310.8
## - parking  1 4.0670e+11 2.9782e+12 3326.3
## - taxes    1 4.4992e+11 3.0215e+12 3328.3
## - list     1 3.7097e+13 3.9669e+13 3686.2
## 
## Step:  AIC=3306.97
## sale ~ list + parking + taxes
## 
##           Df  Sum of Sq        RSS    AIC
## <none>                  2.5925e+12 3307.0
## - parking  1 3.8605e+11 2.9786e+12 3321.3
## - taxes    1 4.4687e+11 3.0394e+12 3324.1
## - list     1 4.0391e+13 4.2984e+13 3692.4
```

```
## 
## Call:
## lm(formula = sale ~ list + parking + taxes, data = data)
## 
## Coefficients:
## (Intercept)         list      parking        taxes  
##   1.954e+05    8.507e-01   -2.520e+04    2.260e+01
```

We get the final model according to the BIC values using backward regression is given by

$$sale = 195400 + 0.85 \times list - 25200.70 \times parking + 22.6 \times taxes$$

3.4 Check Multicollinearity

```
## Analysis of Variance Table
## 
## Response: sale
##            Df     Sum Sq    Mean Sq   F value    Pr(>F)    
## list        1 1.2538e+14 1.2538e+14 6528.9659 < 2.2e-16 ***
## parking     1 1.7371e+11 1.7371e+11    9.0454  0.003142 ** 
## taxes       1 4.4687e+11 4.4687e+11   23.2698 3.747e-06 ***
## Residuals 135 2.5925e+12 1.9204e+10                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## Call:
## lm(formula = sale ~ list + parking + taxes, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -333307  -94932  -14955   78844  583818 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.954e+05  2.618e+04   7.463 9.33e-12 ***
## list         8.507e-01  1.855e-02  45.862  < 2e-16 ***
## parking     -2.520e+04  5.620e+03  -4.484 1.55e-05 ***
## taxes        2.260e+01  4.686e+00   4.824 3.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 138600 on 135 degrees of freedom
## Multiple R-squared:  0.9798, Adjusted R-squared:  0.9794
## F-statistic:  2187 on 3 and 135 DF,  p-value: < 2.2e-16
```

Global F-test for the final model is significant, equals to 2187. One of the partial F-test of three variable, list price, is much more significant than others. This indicates there does not exist multicollinearity, but one variable is more significant than other two.

3.5 Summary

The final model is given by:

$$sale = 195400 + 0.85 \times list - 25200.70 \times parking + 22.60 \times taxes$$

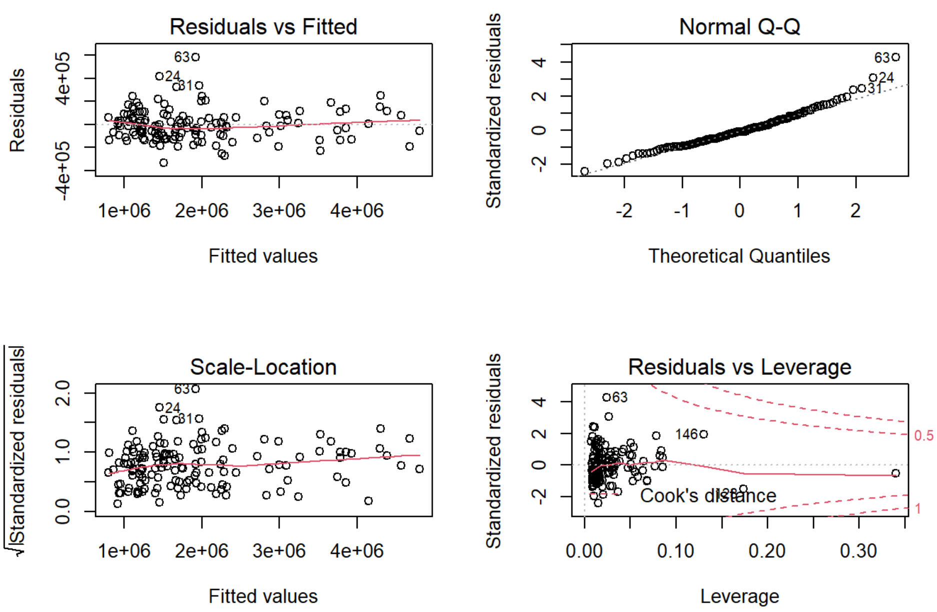
Keep all other variables constant, as the list price increased by 1, the property's average sale price will increase by 0.85. Keep all other variables constant, as the parking spots increased by 1, the property's average sale price will decrease by 25200.7. Keep all other variables constant, as the property tax increased by 1, the property's average sale price will increase by 22.6.

AIC or BIC produces the same model, but different from the full model. Three of the variables are removed, and the coefficients of the remaining three are sightly different. AIC and BIC are both penalized-likelihood criteria. The model given by AIC and BIC is more likely to approximate the true model. In this case, they choose three variables considered to be more influential to the dependent variable. The coefficients are different because three independent variables are deleted from the model.

IV. Discussions and Limitations

4.1 Diagnostic Plots

FIGURE 4.1 Diagnostic Plots (1548)



4.2 Interpretation Diagnostic Plots

4.2.1 Residual vs Fitted

In this case, the residuals are approximately randomly distributed around the horizontal zero line. This indicates that the residuals and the fitted values are uncorrelated. The assumption of equal variance (homoscedasticity) is satisfied.

4.2.2 Normal Q-Q

Most of the points follow the theoretical normal line. This indicates that the residuals are normally distributed. The assumption of the normal error MLR is satisfied.

4.2.3 Scale-Location

There is no obvious pattern shown in the graph. This indicates that the residuals are spread equally along with the ranges of predictors. The assumption of equal variance (homoscedasticity) is satisfied.

4.2.4 Residuals vs Leverage

All of the points are in the red line region. This indicates that there is no leverage point or outlier.

4.3 Future Work

- We could use box cox transformation to improve our model by making some of the variables more normally distributed.
- We could add more independent variables that may affect the sale price of properties.