Data Analysis on house sale price in Toronto and Mississauga

Chaeyeon Stella Bae

December 4, 2020

In this paper, we are going to look at more complex data analysis on predicting sale price of detached homes in two different neighbourhoods, Mississauga and Toronto, using Multiple Linear Model.

I. Data Wrangling

We will use randomly selected samples of 150 cases of the following IDs.

```
##   [1]   1   3   4   5   6   7   8   9  11  12  13  14  15  16  17  20  21  22
##  [19]  23  24  25  26  27  29  30  32  33  34  35  36  38  39  40  41  42  45
##  [37]  47  48  49  51  52  53  55  56  57  58  60  61  62  63  64  65  66  67
##  [55]  68  69  70  71  72  73  75  77  78  81  82  83  89  90  91  94  96  97
##  [73]  99 100 101 102 105 106 108 109 110 112 114 116 118 119 122 125 126 131
##  [91] 132 134 135 136 137 139 141 142 143 145 147 148 149 150 151 152 154 155
## [109] 156 157 158 159 161 162 163 164 165 166 167 168 169 172 173 174 175 176
## [127] 177 178 179 180 181 182 183 185 186 187 188 189 190 191 193 194 195 196
## [145] 201 204 205 218 227 229
```

Cleaning Data

To begin with the cleaning process of data, we need to identify missing values.

Since there are too many missing values of maximum square footage(maxsqfoot), if we remove all of the cases containing missing values, we will end up omiting too many cases. Therefore, we will remove the maximum square footage variable which contains the most missing values. Then, we identify the missing values as shown below. Now we only got 7 cases containing missing values to omit.

```
##      ID    sale    list bedroom bathroom parking     taxes lotwidth lotlength
## 107 109 1075000  979900       3        2      NA     4.375    20.00    100.00
## 40   41 1440000 1500000       7        4       4  4623.000       NA        NA
## 87   89 1200000 1149000       3        2      NA  4114.000    25.00    113.00
## 94   96 5100000 5495000       4        5       4 23592.000       NA        NA
## 79   81  860000  868900       1        2      NA  3676.000    16.10     43.69
## 59   61  755000  649000       1        2      NA  3160.000    19.00     15.65
## 54   55 1185000 1198000       3        3      NA  4011.000    17.00    134.00
## 112 114 1570000 1599000       3        4       1        NA    23.33     73.00
##     location  lotsize
## 107        T 2000.000
## 40         T       NA
## 87         T 2825.000
## 94         T       NA
```

```
## 79          T  703.409
## 59          T  297.350
## 54          T 2278.000
## 112         T 1703.090
```

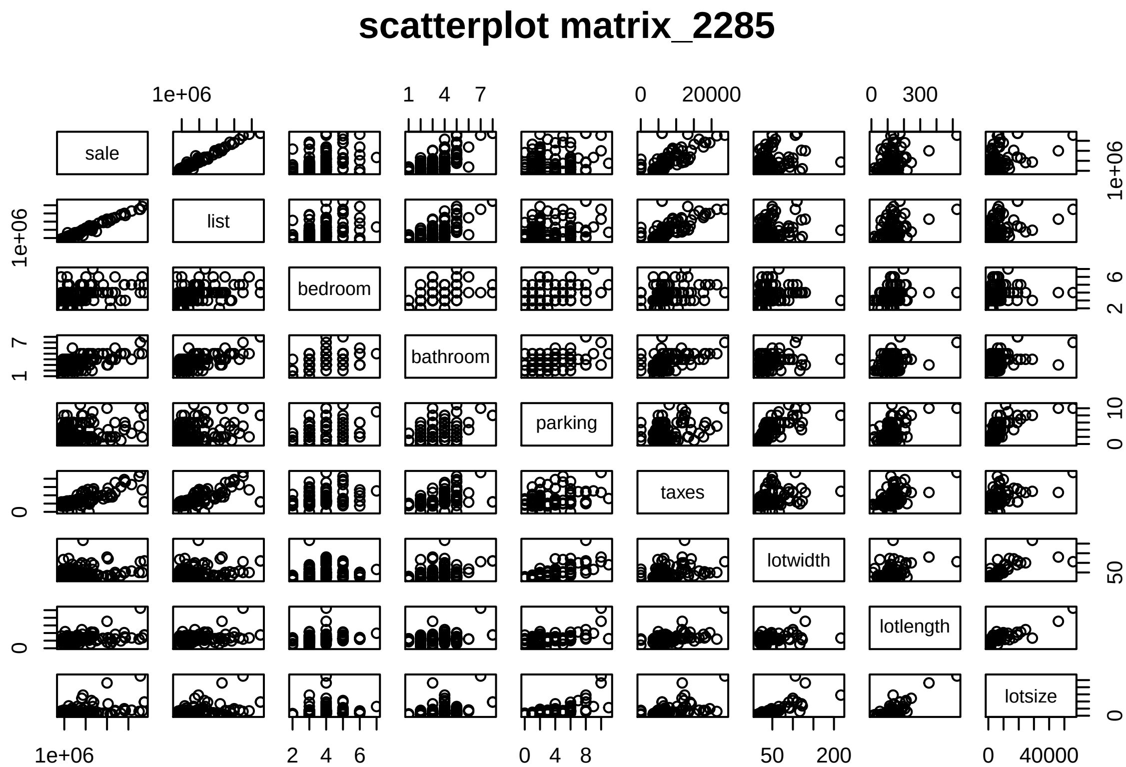
II. Exploratory Data Analysis

a. Classify variables

To classify each variables in this dataset; discrete variables are ID, number of bedrooms, number of bathrooms and number of parking spots. Continuous variables are sale price of property, last list price of property, previous year's property taxes, width, length and size of property. A categorical variable is location of neighborhood.

b. Pairwise correlations and scatterplot matrix

```
##                sale      list   bedroom  bathroom   parking     taxes  lotwidth
## sale      1.0000000 0.9868505 0.4085449 0.6641468 0.1691756 0.7461814 0.2641505
## list      0.9868505 1.0000000 0.4146378 0.6867011 0.2172145 0.7143954 0.2960767
## bedroom   0.4085449 0.4146378 1.0000000 0.5407135 0.3619200 0.3455984 0.2395233
## bathroom  0.6641468 0.6867011 0.5407135 1.0000000 0.4020838 0.4839778 0.3670764
## parking   0.1691756 0.2172145 0.3619200 0.4020838 1.0000000 0.3080157 0.7343786
## taxes     0.7461814 0.7143954 0.3455984 0.4839778 0.3080157 1.0000000 0.3757826
## lotwidth  0.2641505 0.2960767 0.2395233 0.3670764 0.7343786 0.3757826 1.0000000
## lotlength 0.3972230 0.4015202 0.2206447 0.3289028 0.4671357 0.5134494 0.3714166
## lotsize   0.4025891 0.4183451 0.2165965 0.3930542 0.6803074 0.5107855 0.7751052
##           lotlength   lotsize
## sale      0.3972230 0.4025891
## list      0.4015202 0.4183451
## bedroom   0.2206447 0.2165965
## bathroom  0.3289028 0.3930542
## parking   0.4671357 0.6803074
## taxes     0.5134494 0.5107855
## lotwidth  0.3714166 0.7751052
## lotlength 1.0000000 0.8225289
## lotsize   0.8225289 1.0000000
```

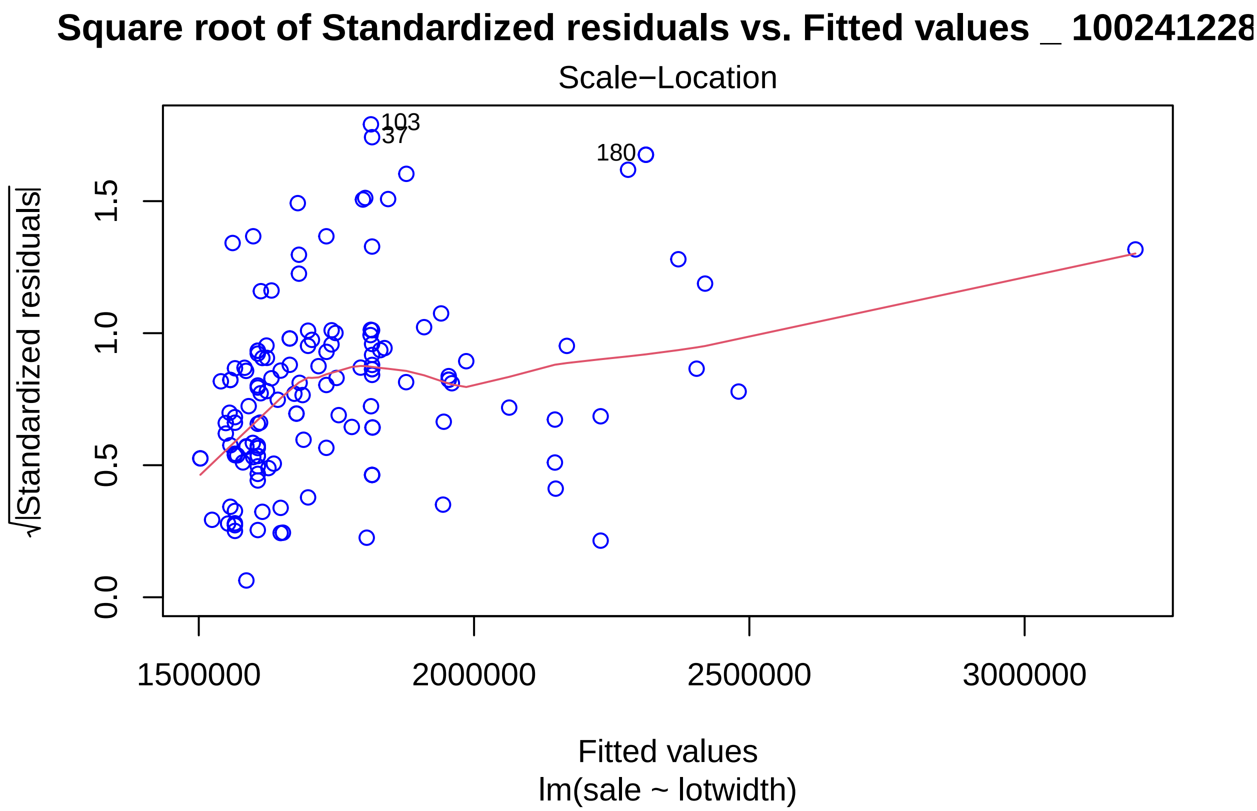


rank	1	2	3	4	5	6	7	8
predictors	list price	taxes	bathroom	bedroom	lotlength	lotsize	lotwidth	parking
correlation coefficient	0.9875	0.7546	0.6614	0.4558	0.4307	0.4215	0.3464	0.2561

As shown in the correlation matrix and scatterplot matrix above, all predictors showed positive correlation with sale price. List price ranked the highest correlation with sale price, which is 0.987 of correlation coefficient. Previous year's property taxes has the second highest correlation with sale price of 0.755, then followed by number of bathroom with correlation of 0.661. Number of bedroom ranked forth highest with correlation of 0.4558, followed by lotlegth, lotsize, lotwidth of property, with correlation of 0.431, 0.422, and 0.346. Lastly, parking ranked the lowest correlation with sale price, of 0.256.

c. Identifying assumption of constant variance violation

By looking at the scatterplot matrix, the assumption of constant variance for lotwdith of sale price would be strongly violated. This can be proved by the standardized residual plot below since it does not show a random/equal spread around the red horizontal line.



III. Methods and Model

a. Multiple linear regression model

Now, we will look at the multiple linear regression with all available predictors for sale price.

	(Intercept)	list	bedroom	bathroom	parking	taxes	lotwidth	lotlength	locationT
estimated regression coefficient	1.166e+05	8.133e-01	4.043e+03	1.837e+04	-1.818e+04	-1.818e+04	2.226e+01	-3.409e+02	-4.021e+02
p-values	0.2103	<0.0001	0.7740	0.1770	0.0478	<0.0001	0.7877	0.4895	0.0178

In accordance with the p-values, list price, taxes, parking, locationT has the significant t-test results. For fixed amount of all other predictors, for every 1 dollar increase in list price leads to an increase in sale price by 81.33 cents on average. The parking coefficient suggests that every 1 unit increase in number of parking spot will result a decrease in sale price by 18,180 dollar on average, holding all other predictors fixed. Also, For every 1 dollar increase in taxes, sale price increase by 22.26 dollars, on average, holding all other predictors constant. Last of all, 1 unit increase in difference between the means of locationT and locationM leads to an increase of 106,500 dollar in sale price on average, for all other predictors fixed.

b. Backward elimination model using AIC

```
## Start:  AIC=3341.27
## sale ~ list + bedroom + bathroom + parking + taxes + lotwidth +
##     lotlength + location + lotsize
##
##             Df  Sum of Sq        RSS    AIC
## - bedroom    1 1.0964e+10 2.0531e+12 3340.0
## - lotwidth   1 1.5440e+10 2.0576e+12 3340.3
## - bathroom   1 2.0788e+10 2.0630e+12 3340.7
## - lotlength  1 2.7688e+10 2.0699e+12 3341.2
## <none>                    2.0422e+12 3341.3
## - lotsize    1 4.5108e+10 2.0873e+12 3342.4
## - parking    1 4.6700e+10 2.0889e+12 3342.5
## - location   1 1.0739e+11 2.1496e+12 3346.5
## - taxes      1 5.7242e+11 2.6146e+12 3374.4
## - list       1 2.0846e+13 2.2888e+13 3682.4
##
## Step:  AIC=3340.03
## sale ~ list + bathroom + parking + taxes + lotwidth + lotlength +
##     location + lotsize
##
##             Df  Sum of Sq        RSS    AIC
## - lotwidth   1 1.1948e+10 2.0651e+12 3338.9
## - lotlength  1 2.2398e+10 2.0755e+12 3339.6
## <none>                    2.0531e+12 3340.0
## - parking    1 3.7198e+10 2.0903e+12 3340.6
## - lotsize    1 3.7245e+10 2.0904e+12 3340.6
## - bathroom   1 3.8323e+10 2.0915e+12 3340.7
## - location   1 1.3202e+11 2.1851e+12 3346.9
## - taxes      1 5.8603e+11 2.6392e+12 3373.7
## - list       1 2.0857e+13 2.2910e+13 3680.6
##
## Step:  AIC=3338.85
## sale ~ list + bathroom + parking + taxes + lotlength + location +
##     lotsize
##
##             Df  Sum of Sq        RSS    AIC
## - lotlength  1 1.0470e+10 2.0755e+12 3337.6
## <none>                    2.0651e+12 3338.9
## - lotsize    1 3.2416e+10 2.0975e+12 3339.1
## - bathroom   1 4.0306e+10 2.1054e+12 3339.6
## - parking    1 4.8523e+10 2.1136e+12 3340.1
## - location   1 1.6718e+11 2.2323e+12 3347.9
## - taxes      1 5.7414e+11 2.6392e+12 3371.7
## - list       1 2.0952e+13 2.3017e+13 3679.2
##
## Step:  AIC=3337.57
## sale ~ list + bathroom + parking + taxes + location + lotsize
##
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 2.3700e+10 2.0992e+12 3337.2
## <none>                   2.0755e+12 3337.6
## - bathroom  1 3.8638e+10 2.1142e+12 3338.2
```

```
## - parking   1 4.4212e+10 2.1198e+12 3338.6
## - location  1 1.6050e+11 2.2360e+12 3346.1
## - taxes     1 5.6431e+11 2.6399e+12 3369.7
## - list      1 2.1158e+13 2.3234e+13 3678.6
##
## Step:  AIC=3337.18
## sale ~ list + bathroom + parking + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## - parking   1 2.5850e+10 2.1251e+12 3336.9
## <none>                   2.0992e+12 3337.2
## - bathroom  1 3.1621e+10 2.1309e+12 3337.3
## - location  1 1.4602e+11 2.2453e+12 3344.7
## - taxes     1 6.5506e+11 2.7543e+12 3373.7
## - list      1 2.2346e+13 2.4445e+13 3683.8
##
## Step:  AIC=3336.92
## sale ~ list + bathroom + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   2.1251e+12 3336.9
## - bathroom  1 3.1686e+10 2.1568e+12 3337.0
## - location  1 5.2706e+11 2.6522e+12 3366.4
## - taxes     1 6.3634e+11 2.7614e+12 3372.1
## - list      1 2.3277e+13 2.5402e+13 3687.2
```

```
##
## Call:
## lm(formula = sale ~ list + bathroom + taxes + location, data = newreal203)
##
## Coefficients:
## (Intercept)         list     bathroom        taxes    locationT
##   2.913e+04    8.041e-01    1.832e+04    2.464e+01    1.459e+05
```

Using backward elimination with AIC, we choose the model with list, parking, taxes, and location predictors. Our final model looks like following:

$$\hat{sale} = 118,500 + 0.8352 list - 12,510 parking + 22.64 taxes + 87,610 locationT$$

The results are consistent with part a above that the relevant predictors were also found significant.

c. Backward elimination model using BIC

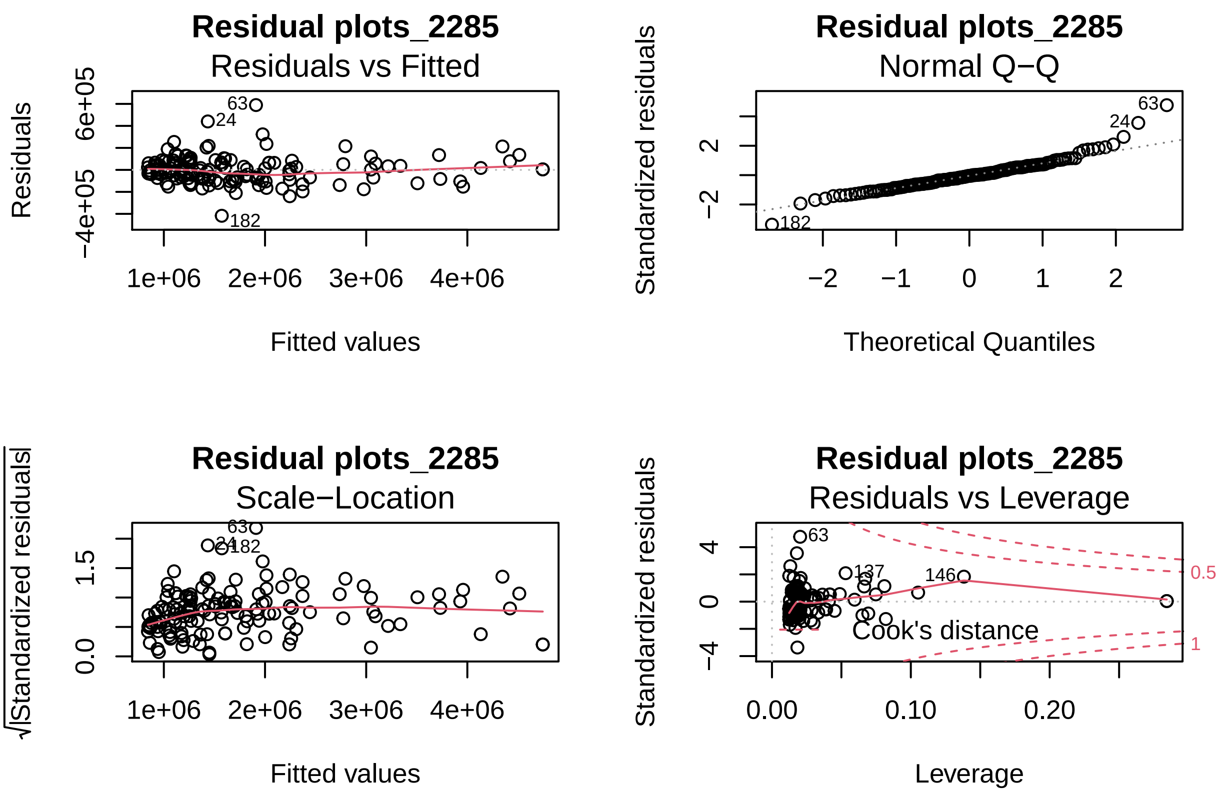
Using backward elimination with BIC, we end up with the model with list, taxes, and location predictors. Our final model looks like following:

$$\hat{sale} = 73,340 + 0.8270 list + 21.51 taxes + 132,200 locationT$$

The results are not consistent with part a and b above. List price, taxes and locationT are found as relevant predictors, however, number of parking spots resulted as one of the predictors in part a and b is eliminated with backward elimination using BIC here.

IV. Discussions and Limitations

a. Diagnostic plots



b. Interpretation of residual plots

We can conclude whether the normal error MLR assumptions are satisfied by interpreting the plots above. Residuals vs Fitted plot shows data spread around a horizontal line without a pattern, but points are not quite equally spread around the line. Therefore, we can see that the multicolinearity assumption is violated. Normal Q-Q plot shows fairly good alignment with the line, indicating that the errors are normally distributed. The Scale-Location plot appear to have higher density in residuals at lower fitted values, not equally spread residuals. This indicates that the assumption of constant variance (homoscedasticity) is violated. Finally, Residuals vs. Leverage plot shows no influential point to exclude that no point is beyond Cook's distance.

c. Further Steps

Since our model have violated multicolinearity and homoscedasticity assumption, we can further try Box-Cox Transformations or Partial F-test to find a valid final model.