MLR for Homes in GTA

Eddie Moon, Id 1004161916

December 4, 2020

In this assignment we will explore and analyze first-time home buying in our neighbourhoods, a current major issue. Especially during the COVID-19 crisis, prices for detached houses have been at all-time high. For this assignment the data was obtained from the Toronto Real Estate Board (TREB), provided in the assignment handout.

There are many variables in our data: ID, sale, list, bedroom, bathroom, parking, maxsqfoot, taxes, lotwidth, lotlength, and location. The goal of this assignment will be to utilize our variables and data to develop a complex linear model in which home buyers can use to predict the sale price of single-family detahced homes in the two neighbourhoods in the GTA.

I. Data Wrangling

First we will load our data in and randomly select a sample of 150 cases.

Then we will create a new variable named “lotsize” that will multiply lotwidth by lotlength and replace them in the data.

Now we can clean the data by removing at most eleven cases and one predictor. For the predictor variable we will choose to remove the maxsqfoot variable. This is because it seems that approximately half or maybe more of the cases have missing values (NA) for the maxsqfoot variable so it will not be a very nice predictor to work with to build a linear model. Therefore we can choose to remove this one.

After removing this predictor we can see a couple of cases still with missing values mostly in the parking variable column and one in lotsize so we will go ahead and remove those.

Now we have removed 8 cases with “NA” values, we are allowed to remove three more, so we will remove the cases where there is an extreme outlier. There were many cases where the taxes variable had an unabsurdly low number, so we will remove three of those. Now we have our dataset we can use for the rest of the assignment.

	ID <int>	sale <int>	list <int>	bedroom <int>	bathroom <int>	parking <int>	taxes <dbl>	location <fctr>	lotsize <dbl>
1	162	1039000	1049000	3	4	2	4955.000	M	5471.2500
2	91	1500000	1650000	3	3	2	6320.000	T	2600.0000
3	62	1860000	1898000	3	3	2	6736.000	T	3500.0000
4	53	1850000	1999000	4	3	2	6982.000	T	8984.7200
5	115	1738000	1698000	3	4	2	6109.000	T	2180.0000
6	36	1415000	1480000	4	3	2	4820.000	T	3498.2400
7	7	1281000	1199000	3	2	2	4230.000	T	1412.0000
8	102	2240000	2399000	4	4	1	8097.000	T	3498.7500

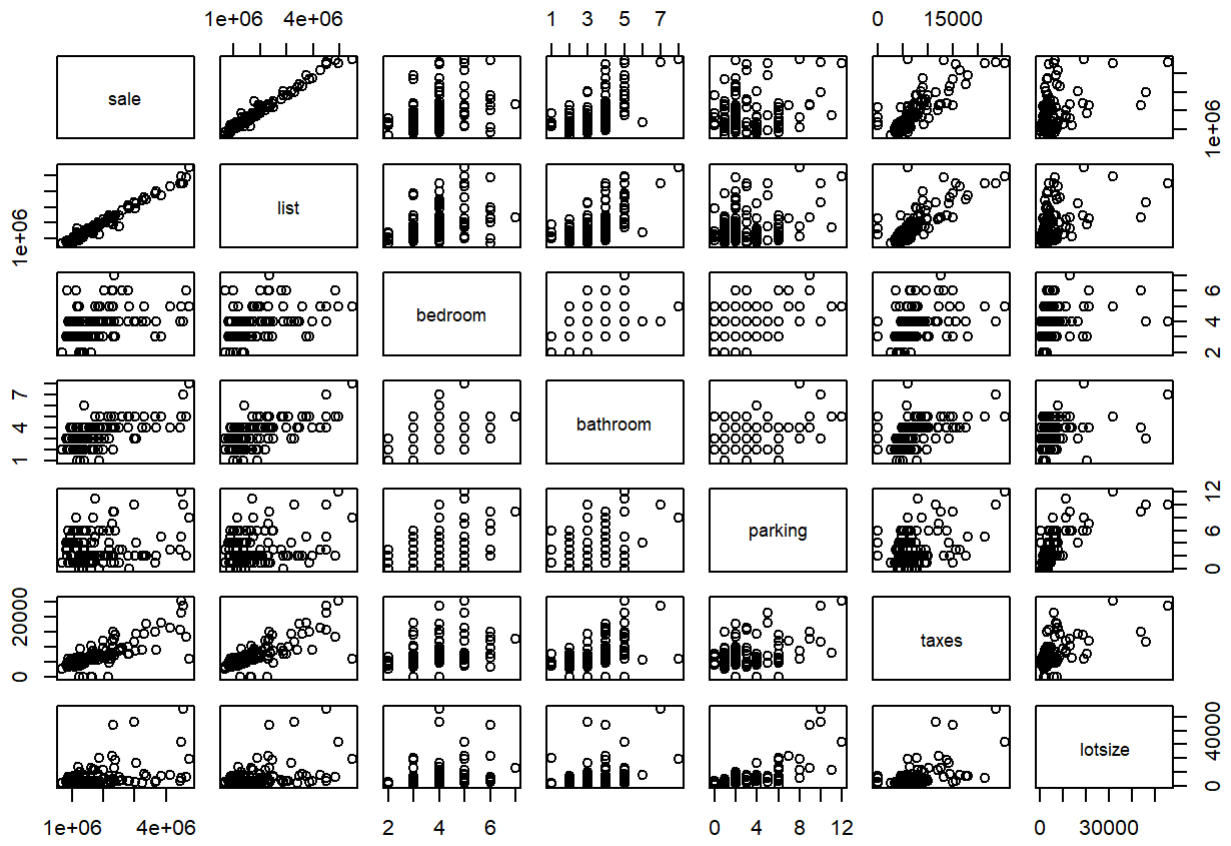
	ID	sale	list	bedroom	bathroom	parking	taxes	location	lotsize
	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<fctr>	<dbl>
9	103	1715000	1850000	3	2	2	7741.000	T	2185.8200
10	132	837000	850000	3	2	3	3917.000	M	2500.0000

1-10 of 140 rows Previous **1** 2 3 4 5 6 … 14 Next

II. Exploratory Data Analysis

Our categorical variables are: ID, location Our discrete variables are: sale, list, bedroom, bathroom, parking, taxes Our continuous variables are: lotsize, lotwidth, lotlength, maxsqfoot

Now we will create a scatterplot matrix and produce the pairwise correlations for all pairs of quantitative variables in our data.

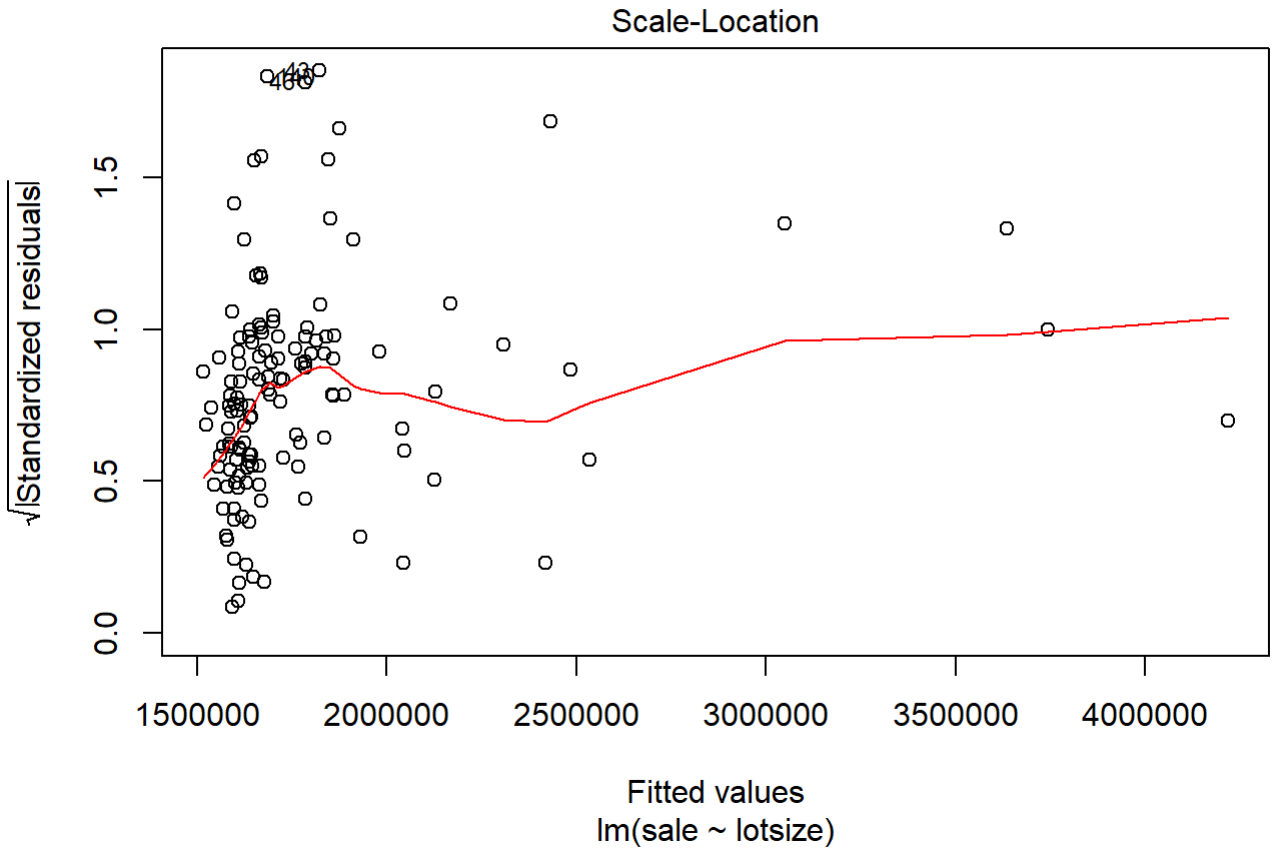



```
##        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
## [1,] 1.0000 0.9860 0.4223 0.6067 0.1947 0.7912 0.4229
## [2,] 0.9860 1.0000 0.4271 0.6216 0.2396 0.7748 0.4400
## [3,] 0.4223 0.4271 1.0000 0.5078 0.3662 0.3663 0.2978
## [4,] 0.6067 0.6216 0.5078 1.0000 0.3586 0.4755 0.3435
## [5,] 0.1947 0.2396 0.3662 0.3586 1.0000 0.3921 0.6976
## [6,] 0.7912 0.7748 0.3663 0.4755 0.3921 1.0000 0.5947
## [7,] 0.4229 0.4400 0.2978 0.3435 0.6976 0.5947 1.0000
```

We can clearly see that the variables list and sale have the highest correlation coefficient of 9.872. The ranking for predictor variables of sale price from highest coefficient to lowest is:

1. list
2. taxes
3. bathroom
4. lotsize
5. bedroom
6. parking

Based on our scatterplot matrix, the single predictor violating the assumption of constant variance the most seems to be the lotsize variable.



We can check with a plot and see that this is true.

III. Methods and Model

Now we will fit an additive linear regression model.

```
##
## Call:
## lm(formula = sale ~ list + bedroom + bathroom + parking + taxes +
##     lotsize, data = cleandata1916)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -407502  -82541  -26616   54547  577417
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.237e+05  4.906e+04   2.522   0.0129 *
## list         8.401e-01  2.190e-02  38.353  < 2e-16 ***
## bedroom      1.722e+04  1.434e+04   1.201   0.2319
## bathroom     7.535e+03  1.392e+04   0.541   0.5893
## parking     -2.969e+04  7.156e+03  -4.150 5.90e-05 ***
## taxes        2.246e+01  5.016e+00   4.477 1.61e-05 ***
## lotsize      1.109e+00  2.381e+00   0.466   0.6421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137400 on 133 degrees of freedom
## Multiple R-squared:  0.9782, Adjusted R-squared:  0.9772
## F-statistic: 995.5 on 6 and 133 DF,  p-value: < 2.2e-16
```

Now we will do stepwise regression with AIC.

```
## Start:  AIC=3319.35
## sale ~ list + bedroom + bathroom + parking + taxes + lotsize
##
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 4.0944e+09 2.5141e+12 3317.6
## - bathroom  1 5.5277e+09 2.5156e+12 3317.7
## - bedroom   1 2.7220e+10 2.5373e+12 3318.9
## <none>                   2.5100e+12 3319.4
## - parking   1 3.2500e+11 2.8350e+12 3334.4
## - taxes     1 3.7827e+11 2.8883e+12 3337.0
## - list      1 2.7761e+13 3.0271e+13 3665.9
##
## Step:  AIC=3317.58
## sale ~ list + bedroom + bathroom + parking + taxes
##
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 4.8971e+09 2.5190e+12 3315.9
## - bedroom   1 2.5954e+10 2.5401e+12 3317.0
## <none>                   2.5141e+12 3317.6
## - taxes     1 4.4438e+11 2.9585e+12 3338.4
## - parking   1 4.6012e+11 2.9743e+12 3339.1
## - list      1 2.8024e+13 3.0538e+13 3665.2
##
## Step:  AIC=3315.86
## sale ~ list + bedroom + parking + taxes
##
##           Df  Sum of Sq        RSS    AIC
## - bedroom  1 3.5422e+10 2.5545e+12 3315.8
## <none>                   2.5190e+12 3315.9
## - taxes    1 4.3948e+11 2.9585e+12 3336.4
## - parking  1 4.6152e+11 2.9806e+12 3337.4
## - list     1 3.4970e+13 3.7489e+13 3691.9
##
## Step:  AIC=3315.81
## sale ~ list + parking + taxes
##
##           Df  Sum of Sq        RSS    AIC
## <none>                   2.5545e+12 3315.8
## - parking  1 4.2632e+11 2.9808e+12 3335.4
## - taxes    1 4.2947e+11 2.9839e+12 3335.6
## - list     1 3.8721e+13 4.1276e+13 3703.4
```

```
##
## Call:
## lm(formula = sale ~ list + parking + taxes, data = cleandata1916)
##
## Coefficients:
## (Intercept)         list      parking        taxes
##   1.796e+05    8.535e-01   -2.477e+04    2.262e+01
```

The final model here is sale as the dependent variable and list + parking + taxes as the explanatory predictors. So, lm(formula=sale~list+parking+taxes). The results are somewhat consistent with those in the previous fullmodel, in that they have similar coefficients.

Now we will do BIC instead of AIC.

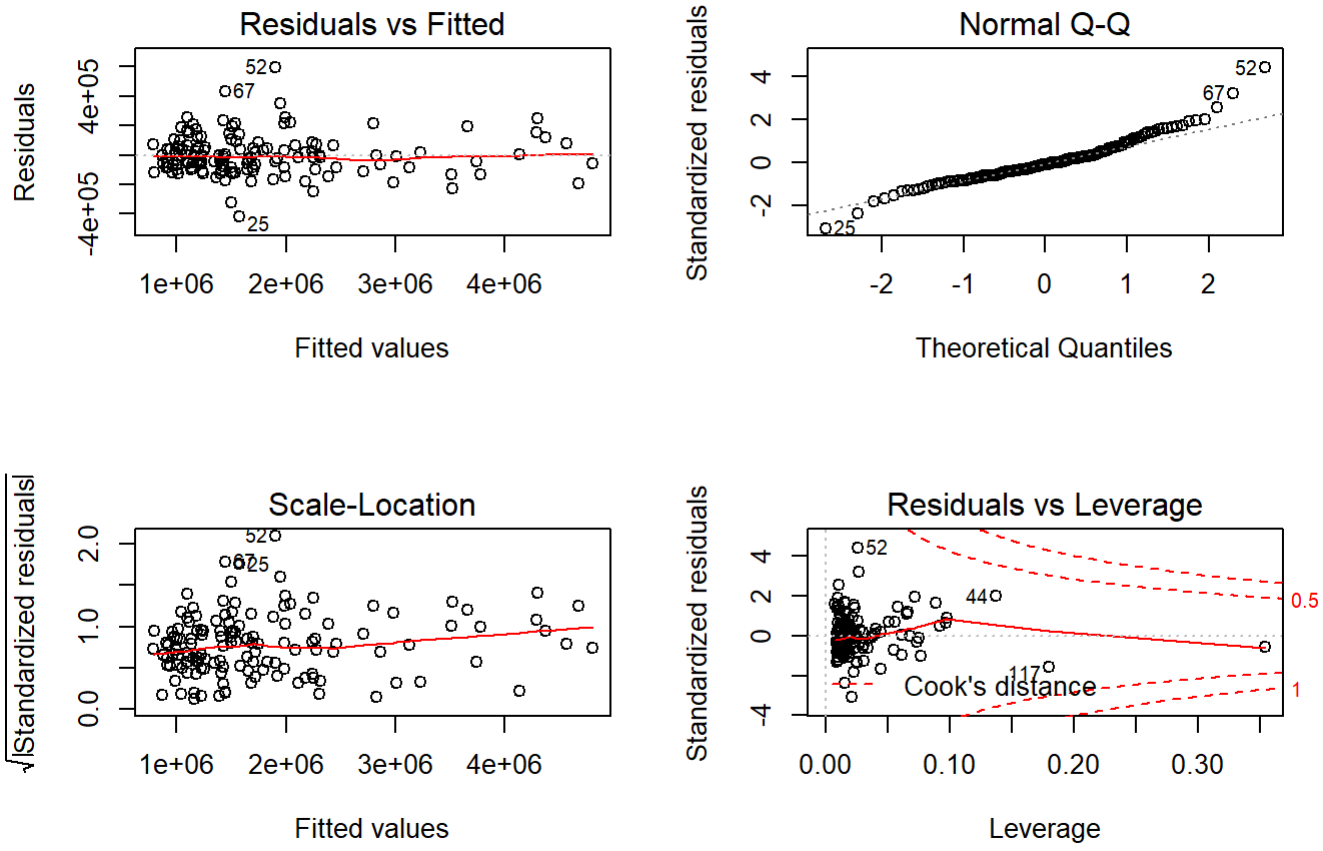
```
## Start:  AIC=3339.95
## sale ~ list + bedroom + bathroom + parking + taxes + lotsize
##
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 4.0944e+09 2.5141e+12 3335.2
## - bathroom  1 5.5277e+09 2.5156e+12 3335.3
## - bedroom   1 2.7220e+10 2.5373e+12 3336.5
## <none>                   2.5100e+12 3339.9
## - parking   1 3.2500e+11 2.8350e+12 3352.1
## - taxes     1 3.7827e+11 2.8883e+12 3354.7
## - list      1 2.7761e+13 3.0271e+13 3683.6
##
## Step:  AIC=3335.23
## sale ~ list + bedroom + bathroom + parking + taxes
##
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 4.8971e+09 2.5190e+12 3330.6
## - bedroom   1 2.5954e+10 2.5401e+12 3331.7
## <none>                   2.5141e+12 3335.2
## - taxes     1 4.4438e+11 2.9585e+12 3353.1
## - parking   1 4.6012e+11 2.9743e+12 3353.8
## - list      1 2.8024e+13 3.0538e+13 3679.9
##
## Step:  AIC=3330.56
## sale ~ list + bedroom + parking + taxes
##
##           Df  Sum of Sq        RSS    AIC
## - bedroom  1 3.5422e+10 2.5545e+12 3327.6
## <none>                   2.5190e+12 3330.6
## - taxes    1 4.3948e+11 2.9585e+12 3348.1
## - parking  1 4.6152e+11 2.9806e+12 3349.2
## - list     1 3.4970e+13 3.7489e+13 3703.6
##
## Step:  AIC=3327.58
## sale ~ list + parking + taxes
##
##           Df  Sum of Sq        RSS    AIC
## <none>                   2.5545e+12 3327.6
## - parking  1 4.2632e+11 2.9808e+12 3344.2
## - taxes    1 4.2947e+11 2.9839e+12 3344.4
## - list     1 3.8721e+13 4.1276e+13 3712.2
```

```
##
## Call:
## lm(formula = sale ~ list + parking + taxes, data = cleandata1916)
##
## Coefficients:
## (Intercept)         list      parking        taxes
##   1.796e+05    8.535e-01   -2.477e+04    2.262e+01
```

Here we can see that the final model is same as the AIC part, and the results are consistent with that.

IV. Discussions and Limitations

We will now show the 4 diagnostic plots for the final model from part III.



From our first plot we can see a few large outliers, but the red line is fairly horizontal so we can assume linearity.

From our second plot the normal qq plot, we can see that most of the points fall approximately along the reference line so we can assume normailty.

From our third flow of scale-location, we can see that there is a horizontal line with approximately equally spread points so we can asume homoscedasticity and assume constant variance

Looking at our fourth and last plot, we can see that there are three points of interest 52, 44, and 117. Point 52 exceeds 3 standard deviations with a value of 4, so that is not good, meaning it is a high leverage point. However it is within the dashed lines of Cook's distance so it is not an influential point.

The next steps we can take towrads finding a valid ‘final’ model could be Cross Validation, or even going back and doing better variable selection with likelihood.