MLR Analysis on GTA House Price (.)

Chaerin Song, Id 1005745302 (.)

December 5, 2020 (.)

I. Data Wrangling

Part 1 (.)

House price data was obtained from the Toronto Real Estate Board (TREB) on detached houses in the city of Toronto and the city of Mississauga.

Selecting random 150 cases

These are the sorted ID's of selected 150 samples.

```
##   [1]   1   2   3   4   8   9  10  11  13  14  15  16  17  20  21  22  23  24
##  [19]  26  27  28  29  30  31  32  34  36  38  39  40  41  42  43  44  46  47
##  [37]  48  49  50  51  53  54  55  56  57  58  60  61  62  63  64  65  66  67
##  [55]  69  70  72  74  76  77  78  81  82  83  84  85  86  87  89  90  91  96
##  [73]  97  98  99 100 102 103 104 105 106 107 111 112 113 114 116 117 118 119
##  [91] 122 125 126 132 133 134 135 136 137 139 140 141 142 143 144 145 146 147
## [109] 148 149 150 151 153 154 155 156 158 160 161 162 163 165 166 167 170 171
## [127] 172 173 174 176 177 179 181 182 183 185 187 188 190 191 193 194 195 201
## [145] 204 205 207 218 227 229
```

Creating a new variable

And we hereby transformed "lotwidth" and "lotlength" predictors into a new predictor "lotsize." Lotsize will be representing the size of the house in square feet.

```
##   ID    sale    list bedroom bathroom parking maxsqfoot taxes location lotsize
## 1  1 1265000  999900       2        2       1        NA  4732        T  119.00
## 2  2 2200000 1999900       5        3       3        NA  7712        T  156.61
## 3  3 1225000 1169000       5        3       2        NA  4448        T  145.50
## 4  4 1900000 1995000       5        4       2        NA  6783        T  148.92
## 5  8 3170000 3150000       2        4       1        NA 10549        T   94.00
## 6  9 1510000 1599000       3        2       2        NA  5907        T  160.00
```

Removing a predictor and (up to 11) cases

We will remove "maxsqfoot" predictor from our data since it has too many NA values.

```
##   ID    sale    list bedroom bathroom parking taxes location lotsize
## 1  1 1265000  999900       2        2       1  4732        T  119.00
## 2  2 2200000 1999900       5        3       3  7712        T  156.61
## 3  3 1225000 1169000       5        3       2  4448        T  145.50
```

Now, we will remove every case that has at least one N/A value, which counts to be 10 observations.

```
## [1] "Number of NA rows:"
```

```
## [1] 10
```

We can calculate leverage and Cook's distance for each point with a full numerical additive model.

In fact, there are many points that exceed the threshold and are categorized as leverage points, while there is no point with extremely high Cook's distance. However, we can see that the case of index 106 has a noticeably higher Cook's distance than other points; in addition, it also has the biggest hat value.
We could remove this case from our data.

```
## [1] "Top 5 Cooks distance:"
```

```
##   106    49    25   132   131
## 0.455 0.093 0.062 0.045 0.043
```

```
## [1] "Top 5 hat values that exceeds the threshold:"
```

```
##   106   130    88    96    28
## 0.368 0.280 0.256 0.192 0.164
```

II. Exploratory Data Analysis

Part 2 (.)
Table (.)

Variable Classification

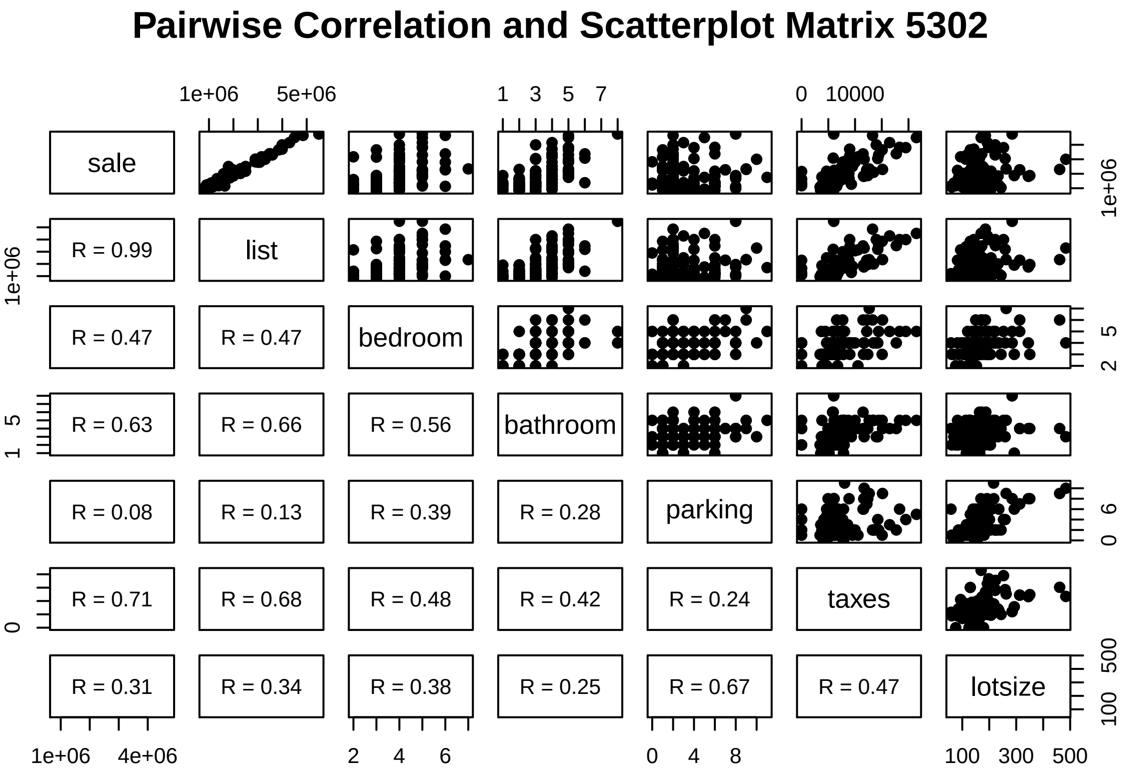
Type of Variables

Variable	Type
sale	continuous
list	continuous
bedroom	discrete
bathroom	discrete
parking	discrete
taxes	continuous
lotsize	continuous
location	categorical

Pairwise Correlation and Scatterplot Matrix

The following matrix plot demonstrates a pairwise correlation and scatterplot of two variables. Variables that are included are: *sale (response variable), list, bedroom, bathroom, parking, taxes, and lotsize.*

The upper diagonal half of the panel demonstrates scatterplots of two corresponding variables, and the lower half demonstrates pairwise correlation of these variables.



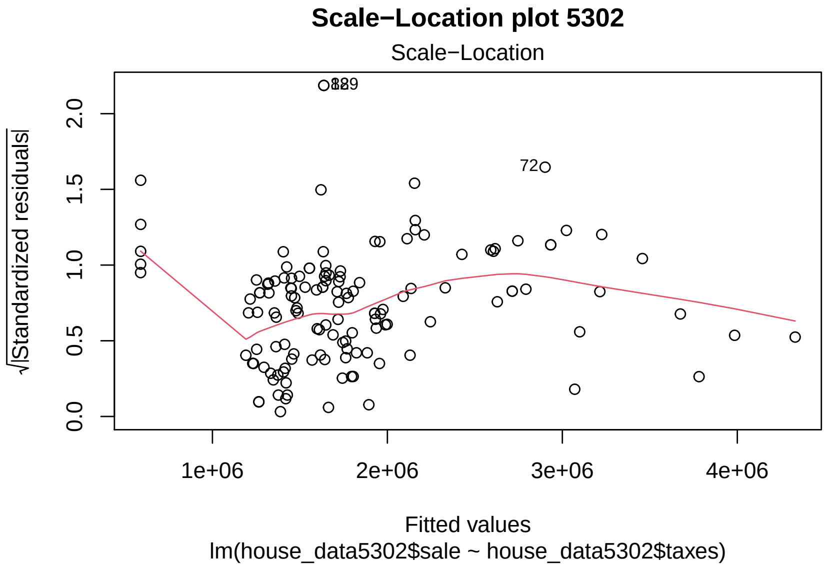
Based on the matrix above, we can rank each quantitative predictor based on their correlation coefficient with the sale variable.
Here is the summary:

Each Quantitative Predictor for Sale Price Rank (base on R value)

Variable	Type	Cor.coeff
1	list	R = 0.99
2	taxes	R = 0.71
3	bathroom	R = 0.63
4	bedroom	R = 0.47
5	lotsize	R = 0.31
6	parking	R = 0.08

Let's see if there is any multicollinearity between any two predictors. Although there are some predictors that have a visibly positive linear relationship, none of their pairwise correlation value indicates that they are strongly correlated.

Based on the scatterplot matrix, it seems like sale~taxes violates the homoscedasticity assumption the most. We shall check this by looking at the scale-location plot.



Since the graph is far from being horizontal, we can confirm that taxes as a predictor of sale price strongly violates the assumption of constant variance.

III. Methods and Model

Part 3 (.)

Table (.)

Fitting a full model

Before we go ahead and fit an additive linear regression model, we shall assign numerical values for location predictor, so that it will be an indicator variable; 1 for T(Toronto), 0 for (Mississauga).

```
##   ID    sale    list bedroom bathroom parking taxes location lotsize
## 1  1 1265000  999900       2        2       1  4732        1  119.00
## 2  2 2200000 1999900       5        3       3  7712        1  156.61
## 3  3 1225000 1169000       5        3       2  4448        1  145.50
## 4  4 1900000 1995000       5        4       2  6783        1  148.92
## 5  8 3170000 3150000       2        4       1 10549        1   94.00
## 6  9 1510000 1599000       3        2       2  5907        1  160.00
```

Now we will fit an additive linear regression model for sale price, including our recoded location variable. Here is the summary of the model:

```
##
## Call:
## lm(formula = sale ~ list + bedroom + bathroom + parking + taxes +
##     lotsize + location)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -415850  -69058  -12413   67629  583707
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.902e+04  6.090e+04   1.298   0.1967
## list         8.333e-01  2.079e-02  40.087  < 2e-16 ***
## bedroom      1.990e+04  1.518e+04   1.310   0.1924
## bathroom    -6.120e+02  1.356e+04  -0.045   0.9641
## parking     -1.203e+04  8.207e+03  -1.465   0.1452
## taxes        2.097e+01  4.357e+00   4.814 4.02e-06 ***
## lotsize     -7.362e+01  2.678e+02  -0.275   0.7838
## location     8.295e+04  3.746e+04   2.214   0.0285 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 128800 on 131 degrees of freedom
## Multiple R-squared:  0.9816, Adjusted R-squared:  0.9806
## F-statistic: 999.9 on 7 and 131 DF,  p-value: < 2.2e-16
```

Our fitted model is:
$\hat{y} = 79020 + 0.8333x_1 + 19900x_2 - 612.0x_3 - 12030x_4 + 20.97x_5 - 73.62x_6 + 82950x_7$

where each explanatory variable represents:

x_1 = list price in CAD
x_2 = number of bedrooms
x_3 = number of bathrooms
x_4 = number of parking spots
x_5 = taxes in CAD
x_6 = lot size in square feet
x_7 = 1 for Toronto neighborhood, and 0 for Missisauga neighborhood.

Our global F-test turned out to be significant with an extremely small p-value.
Now, let’s look at the estimated regression coefficients and the p-values for individual t-tests.

Full additive fitted model

Predictor	Est. Coeff.	p-value
list	0.8333	0.0000
bedroom	19900.	0.1924
bathroom	- 612.0	0.9641
parking	-12030	0.1452
taxes	20.97	0.0000
lotsize	-73.62	0.7838
location	82950.	0.0285

With a benchmark significance level of 5%, list price, taxes, and location can be used as factors to help predict the sale price *over and above* other predictors.
Because our model’s global F-test is significant and some of the t-tests are significant, we can assume that there are some useful explanatory variables in our model for predicting the sale price.

Stepwise regression with AIC

Now, we are going to use a stepwise regression with backward AIC from our original model.

Our final model is:
$\hat{y} = 71150 + 0.8296x_1 + 19760x_2 - 12930x_3 + 20.68x_4 + 85090x_5$
where each explanatory variable represents:

x_1 = list price in CAD
x_2 = number of bedrooms
x_3 = number of parking spots
x_4 = taxes in CAD
x_5 = 1 for Toronto neighborhood, and 0 for Mississauga neighborhood.

In this model, there are only 5 predictors used to predict the sale price: bedroom, parking, location, taxes, and list.

Below is the table demonstrating estimated regression coefficients and p-values for these four predictors.

Fitted model using backward AIC

Predictor	Est. Coeff.	p-value
list	0.8296	0.0000
bedroom	19760.	0.1584
parking	- 12930.	0.0846
taxes	20.68	0.0000
location	85090.	0.0156

This model is not only different in the number of predictors from our original model, but also in its common predictors' estimated coefficients.
In short, the results are not consistent with our original (full) model.

Stepwise regression with BIC

Now, we will use BIC to repeat our what we did above.

Our final model is:
$\hat{y} = 76440 + 0.8313x_1 + 21.21x_2 + 125000x_3$
where each explanatory variable represents:

x_1 = list price in CAD
x_2 = taxes in CAD
x_3 = 1 for Toronto neighborhood, and 0 for Mississauga neighborhood.

In this model, there are only 3 predictors used to predict the sale price: location, taxes, and list.

Below is the table demonstrating estimated regression coefficients and p-values for these four predictors.

Fitted model using backward BIC

Predictor	Est. Coeff.	p-value
list	0.8313	0.0000
taxes	21.21	0.0000
location	125000.	0.0000

The results are not consistent with both our original (full) model and AIC model. This model has the least predictors and the smallest individual t-test p-values. Predictors' estimated coefficients are different from the previous two models as well, although with AIC model they aren't numerically too different.

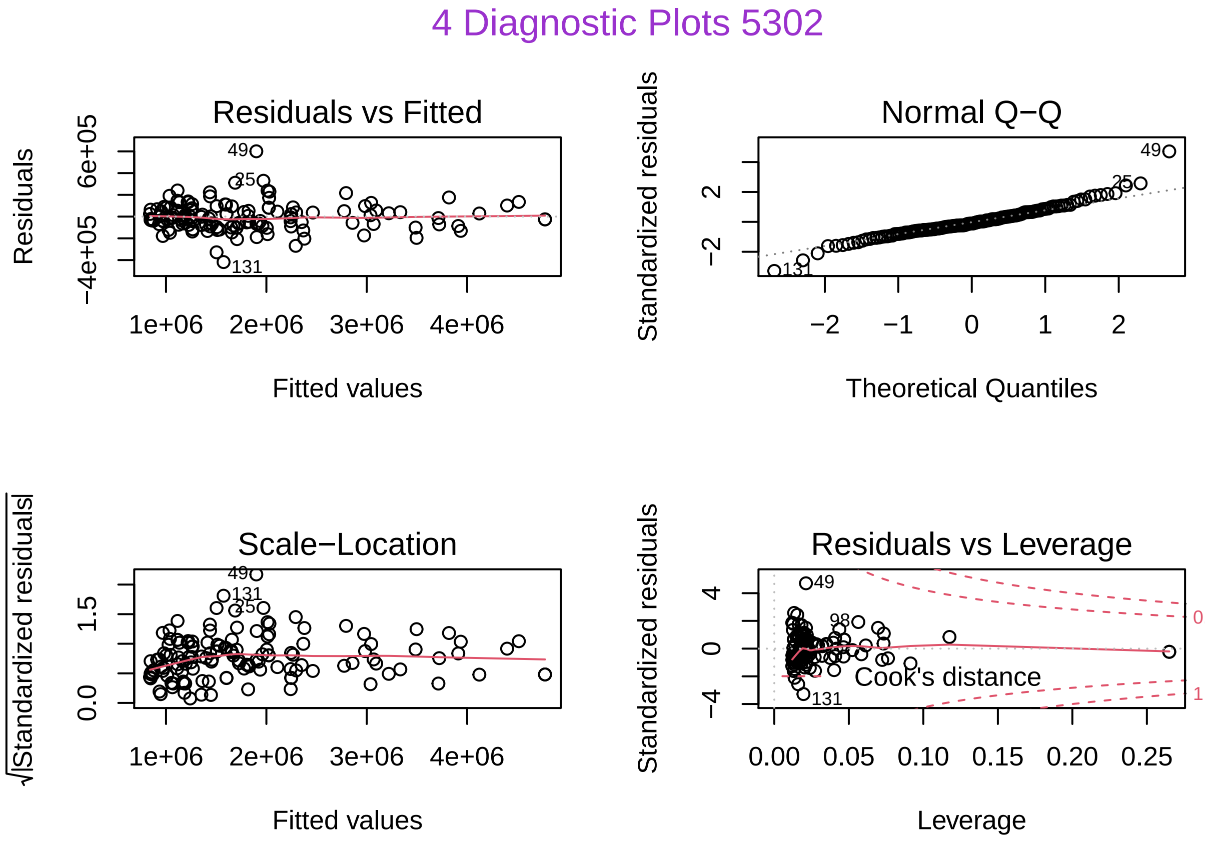
IV. Discussions and Limitations

Part 4 (.)

Plot (.)

Check MLR assumptions

Below are the 4 diagnostic plots for our BIC model.



1. **Residuals vs Fitted**: We obtained a horizontal line without distinct patterns, which is an indication for a linear relationship.

2. **Normal Q-Q**: Most of the residuals are normally distributed, as they follow the straight dashed line. However, some points with large absolute theoretical quantiles are far from the dashed line, which indicates that the normal error MLR assumption is not completely satisfied.

3. **Scale-Location**: Although there is a positive slope with the smaller values of sale price and the points are not too equally spread, there is a generally horizontal line. We are mostly satisfying the homoscedasticity MLR assumption.

4. **Residuals vs Leverage**: There are some points with large absolute standardized residuals, but none of the points' Cook's distance exceeds 1, nor are any of them beyond the dashed lines.

Towards our final model

First, as shown in the scatterplot matrix in part (i), some of our explanatory variables are not normally distributed. In the future, we could conduct a Box-Cox transformation to transform these variables to be close to normally distributed.

Right now, how much have we come to fit a valid model?
We drew scatterplots of the data and removed a bad influential point. We are mostly satisfying our MLR assumptions, and we fit 3 different models to end up with a model with 3 predictors, all statistically significant.
To better satisfy the homoscedasticity assumption, we could possibly use the bootstrap for inference and refit the model.