Comparing Complex Multiple Linear Models for Toronto and Mississauga House Prices

David Pham, 1005349053

December 5, 2020

I. Data Wrangling

In the third and final assignment for STA302, I extend the work done in Assignment 2 and try to find a multiple linear regression model that home buyers can use to predict the sale price of single-family, detached homes in two neighbourhoods in the Greater Toronto Area.

The dataset that I will be using was provided by the STA302 team. The data file is called real203.csv, and it contains 192 observations. I will be randomly sampling 150 data points from the dataset.

To begin, here is the list of the randomly sampled ID's:

```
##   [1]   5  42 102   7  92 125 188  39 110 142   6 166 185  25  71 107 180  13
##  [19]  12  22  41   4  90  86  21 144  75 117   3 177  53  51 193 178 122  54
##  [37]  83 218 201  67  20 132  26 114 155 157  55  29 227  81  66  85  24   2
##  [55]  70 116 172 183 181 207  38 173  87  16  62 109  96  45 179 118 133 103
##  [73] 143 147  48 113  69  97 161 134 137  58 158 176 205 190 175  77 138 174
##  [91] 169  31  91  28  33  65 104 186 131 204  73 119 115  11  79  88  40 154
## [109] 151 189  52 168  63  80 126 171   9  68  27  47  84  15 195  61  23  78
## [127]  72  89 139 146 165  43 160  19  32  98  76 145 159 150 112  99  10  14
## [145]  57  56  17 182 141 162
```

Next, I removed the *maxsqfoot* variable entirely because at least half the data from this column had missing entries. Furthermore, below is the list of cases with missing values, after removing the predictor. I decide to remove these because I would prefer clean data with no NA values anywhere. Unfortunately, my sample had exactly 11 cases with missing data, so I did not have enough room to remove any influential points.

```
##       ID    sale    list bedroom bathroom parking     taxes location  lotsize
## 21    41 1440000 1500000       7        4       4  4623.000        T       NA
## 44   114 1570000 1599000       3        4       1        NA        T 1703.090
## 47    55 1185000 1198000       3        3      NA  4011.000        T 2278.000
## 50    81  860000  868900       1        2      NA  3676.000        T  703.409
## 66   109 1075000  979900       3        2      NA     4.375        T 2000.000
## 67    96 5100000 5495000       4        5       4 23592.000        T       NA
## 76   113 1410000 1375000       5        3      NA  6885.000        T 4380.000
## 121   84  805000  799000       2        2      NA  2654.000        T 2040.000
## 124   61  755000  649000       1        2      NA  3160.000        T  297.350
## 128   89 1200000 1149000       3        2      NA  4114.000        T 2825.000
## 137   76  875000  895000       2        2      NA  3150.000        T 1246.500
```

After removing these cases, we now have a squeaky clean sample!

II. Exploratory Data Analysis

Next, let's quickly classify the variables according to type:

Categorical: location

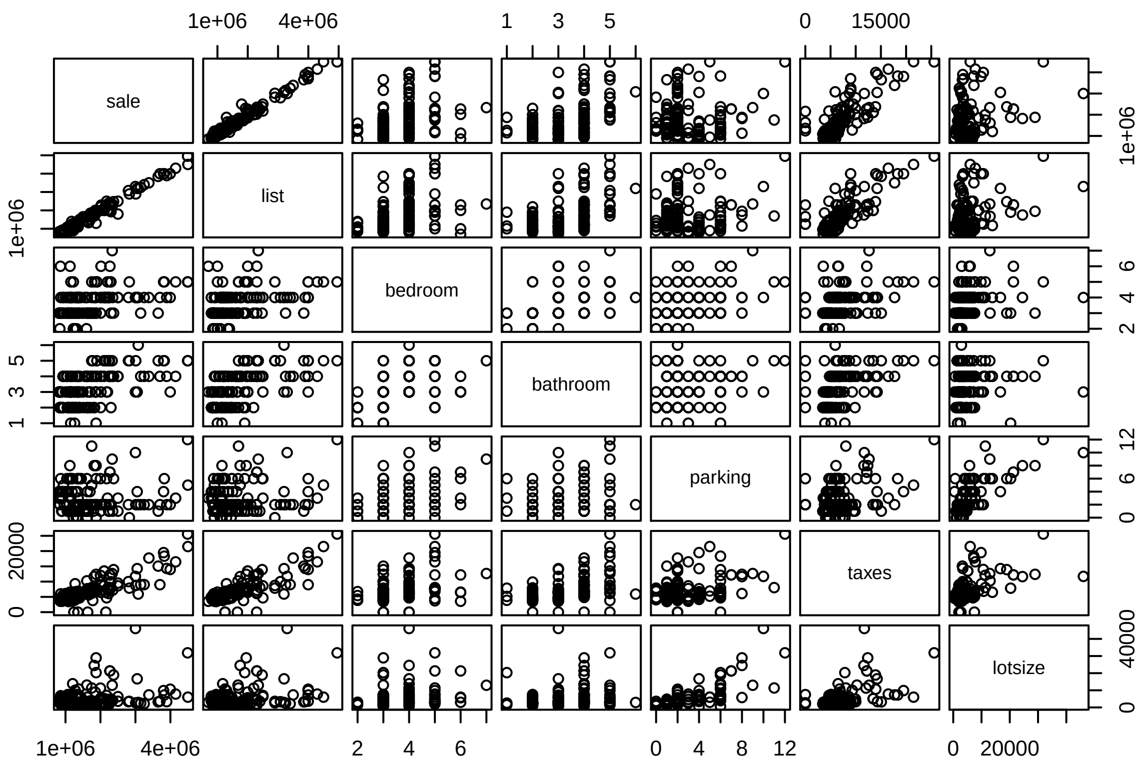
Discrete: ID, sale, list, bedroom, bathroom, parking

Continuous: taxes, lotsize

Most of these are pretty self explanatory, but for the *sale* and *list* variables, I considered them as discrete because they were whole values in the dataset.

Here are all the pairwise correlations and the scatterplot matrix for all the pairs of quantitative variables in the data.

Scatterplots and Correlation Coefficients




```
##             sale   list bedroom bathroom parking  taxes lotsize
## sale      1.0000 0.9861  0.3918   0.5209  0.0846 0.8087  0.3099
## list      0.9861 1.0000  0.3803   0.5333  0.1295 0.8071  0.3409
## bedroom   0.3918 0.3803  1.0000   0.4943  0.3673 0.4034  0.2599
## bathroom  0.5209 0.5333  0.4943   1.0000  0.2660 0.4582  0.1732
## parking   0.0846 0.1295  0.3673   0.2660  1.0000 0.3441  0.7132
## taxes     0.8087 0.8071  0.4034   0.4582  0.3441 1.0000  0.5200
## lotsize   0.3099 0.3409  0.2599   0.1732  0.7132 0.5200  1.0000
```

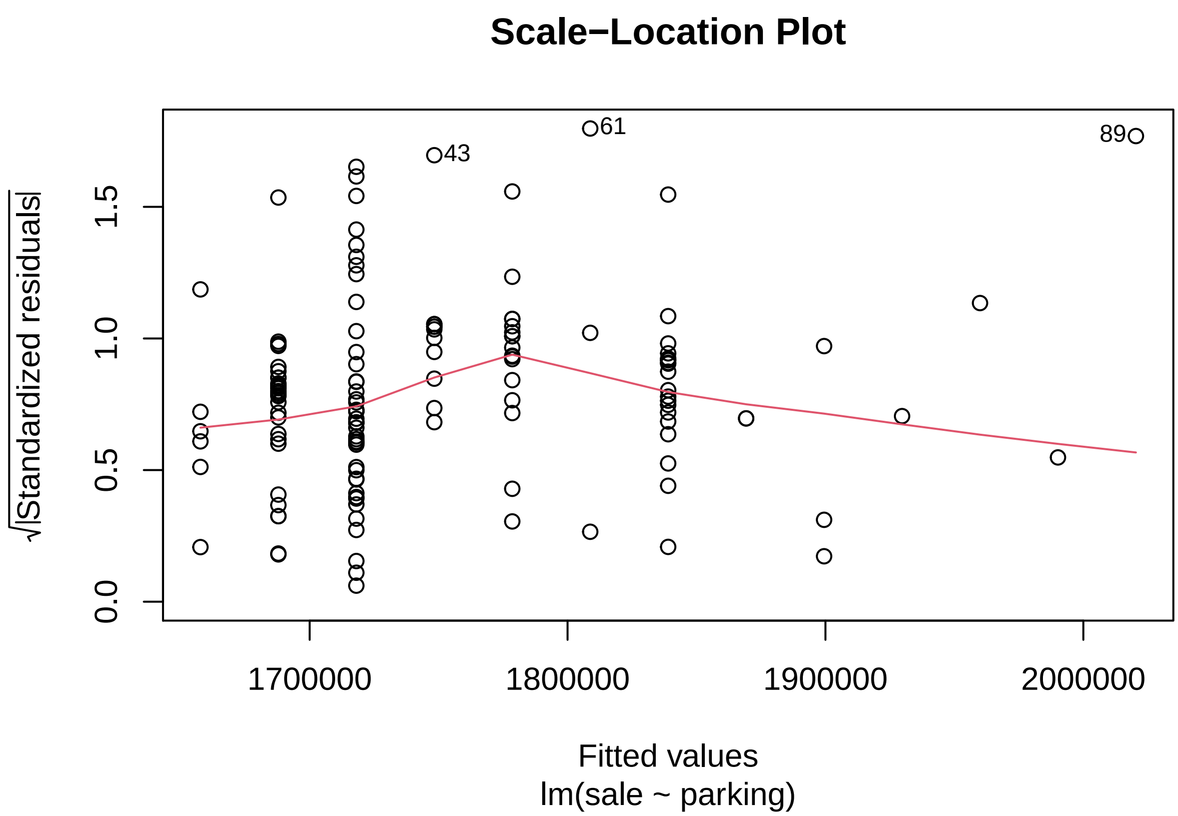
In order, from highest to lowest, the predictors correlate with sale price starting with *list*, then *taxes*, *bathroom*, *bedroom*, *lotsize*, and finally *parking* as the lowest.

Below is a table depicting the strength of the relationships for each predictor and sale price.

```
##      Predictor  Correlation Coefficient for Sale Price
## [1,] "list"     "0.9861"
## [2,] "taxes"    "0.8087"
## [3,] "bathroom" "0.5209"
## [4,] "bedroom"  "0.3918"
## [5,] "lotsize"  "0.3099"
## [6,] "parking"  "0.0846"
```

As a reminder, correlation coefficient measures how well a predictor and response variable form a linear relationship with each other. It can range from -1 to 1, and the closer it is to (+/-) 1, the stronger the relation. We see that the *list* variable almost has a perfect, positive linear relationship with the response, while *parking* almost has no linear relationship with sale price. The *taxes* variable also has a strong positive relationship with sale price, while the others have a weak/moderately positive relationship with the dependent variable.

Lastly for this section, I observe the scatterplot matrix and notice that the *parking* predictor is the most likely to violate the assumption of constant variance (holding all other predictors constant). By looking at the correlation coefficient between it and sale price, as well as the scatterplot between the two variables, it does not appear that they have a linear relationship. The points do not show a clear trend, and the scale-location plot confirms this:



Plotting the square root of the absolute value of the standardized residuals, we see that the constant variance assumption is violated, as the horizontal line bends upwards in the beginning and limps down near the end.

III. Methods and Model

Next, we take a look at the actual multiple linear regression model. I fit an additive linear regression model and have the *location* predictor as an indicator variable (i.e, the additive term).

Below is a table with the estimated regression coefficients, as well as the p-value for the corresponding t-test for that coefficient.

```
##      Regression Coefficient Estimated Regress. Coeff. Value P-value for T-test
## [1,] "Intercept"            "2.86e+04"                      "0.61514"
## [2,] "list"                 "8.24e-01"                      "< 2e-16"
## [3,] "bedroom"              "3.12e+04"                      "0.04326"
## [4,] "bathroom"             "5.49e+03"                      "0.69853"
## [5,] "parking"              "-1.55e+04"                     "0.08317"
## [6,] "taxes"                "1.97e+01"                      "0.00038"
## [7,] "lotsize"              "8.72e-01"                      "0.7602"
## [8,] "locationT"            "9.08e+04"                      "0.02705"
```

Note that the name of the additive regression coefficient for *location* has turned into *locationT*. By interpreting the summary and the table, this means that by holding all other coefficients constant, houses in Toronto are significantly associated with an average increase of 90800 in mean sale price compared to homes in Mississauga.

The p-value for the global F-test is almost 0, so this implies that it is significant. At least one of the slope parameters is not 0. Next, we observe that some of the t-tests are significant. For example, the p-value for the individual t-tests of list price, number of bedrooms, taxes paid and location are all less than the significance level 0.05.

This concludes that there are indeed some useful explanatory variables for predicting the response.

After, we try to find a parsimonious model using stepwise regression with AIC first, and then BIC. Using the step() function, R gives us a final model that is different from the original, full model.

The model went from:

```
## sale ~ list + bedroom + bathroom + parking + taxes + lotsize +
##     location
```

to:

```
## sale ~ list + bedroom + parking + taxes + location
```

It appears that the AIC backward elimination method removed two variables: *bathroom* and *lotsize*. This makes sense because the p-values for these predictors were the largest. Every time the step() function removed a predictor, the AIC value went down slightly (from 3275 -> 3271).

Finally, we will perform backwards elimination with BIC. We use the step() function again, but for the k argument (which represents the multiple of the number of degrees of freedom used for the penalty), we use k = log(n) instead of k = 2, where n is the number of data points.

Interestingly enough, the model is different from both previous parts. The final model is:

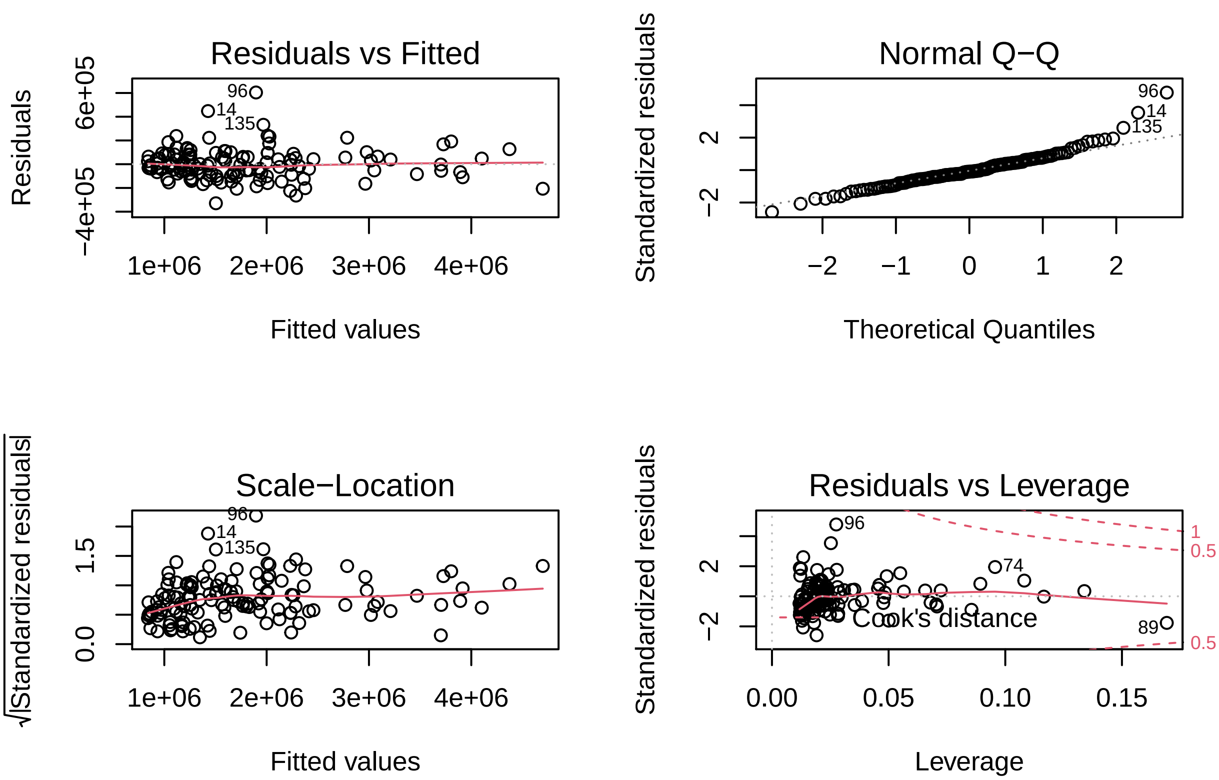
```
## sale ~ list + taxes + location
```

There are only three predictors left: *list*, *taxes* and *location*. It is perfectly possible that AIC and BIC give out different model selections. As a recap, BIC penalizes model complexity more heavily. AIC tends to overfit since the penalty for model complexity is not strong enough. This is sometimes due to the number of estimated parameters being close to a fraction of the sample size. So, the only way the model summaries disagree is if AIC chooses a larger model than BIC, which is indeed what happened. (BIC went from 3298 -> 3286)

As another side note, I'm not sure why the AIC/BIC values produced in the step() function differ from the values obtained from the actual AIC() and BIC() functions. As a result of this, I just referred to the numbers from the step() function.

IV. Discussions and Limitations

Lastly, we will take a look at the diagnostic plots obtained from the reduced model given to us by using the backwards elimination method with BIC.



Starting off with the residuals vs fitted plot, I do not notice a distinctive pattern, and the red line is almost entirely horizontal. So, it is safe to say this is a null plot. There is no trend or pattern anywhere, so linearity is satisfied.

Next, the normal Q-Q plot also looks pretty good. The residuals seem to be very well normally distributed, with the exception of a few points on the top right (Cases 14, 96 and 135). Other than that, normality is satisfied.

Thirdly, the scale-location plot. It also looks okay, with a few noticeable data points (again, cases 14, 96 and 135). The data is a bit clustered to the left of the plot, but the variance of the points is more or less the same. Looking at both the residual vs fitted plot and the scale-location plot, constant variance is satisfied.

Finally, we take a look at the residuals vs leverage plot to see if there are any noteworthy points to take into consideration. Case 96 shows up once again, as well as some newer points (cases 74 and 89). We may consider investigating further into these points to improve upon our model.

In conclusion, I'd say we are very close towards a final model. The next steps would definitely be to take a look at those noteworthy points displayed in the diagnostic plots. They could be heavily affecting the model, so I would double down on that. Overall, in statistics, it is impossible to obtain a 'perfect' model. We just have to try our best to get a really good estimate, and I say I've done a solid job.