What affect the sale prices of Houses in the Great Toronto Area

Zishu Zhu, Id 1002897813

December 4, 2020

I. Data Wrangling

(a) Here are the IDs of the samples:

```
##   [1] 111 204 122  66  95 229 207 151  14 115 157 117 185 168 196  97 174 110
##  [19]  80  35 173  25 188 137 186  26  98   1   9 163  99  22 133 181 180 100
##  [37]  68  49  17 154 177 126  94 155  27  39   5 108  69 139  74  56  53 143
##  [55] 134  90  47  52 161  83  19 187  44  34  65  55  54  91 190 150 218 149
##  [73] 171  57  28  40  78 142  16  24  61  58  60 144  88 138 176  45  62 179
##  [91] 162 159 193  15  50 158 132  67 178 136 131  10 183  75   8  32 227 153
## [109] 101 182  92  46 189 152 166   3  86  36 119  79 170 160 172 103 169 175
## [127]  81 194 125  29  93  21 116  64  71  33 113 114  42  72 109  89 112  12
## [145] 212 147  70  38  20 156
```

(b) We create a new variable with the name "lotsize" by multiplying "lotwidth" by "lotlength". The new variable "lotsize" represents the total area of the properties. We update the new data set with this new variable.

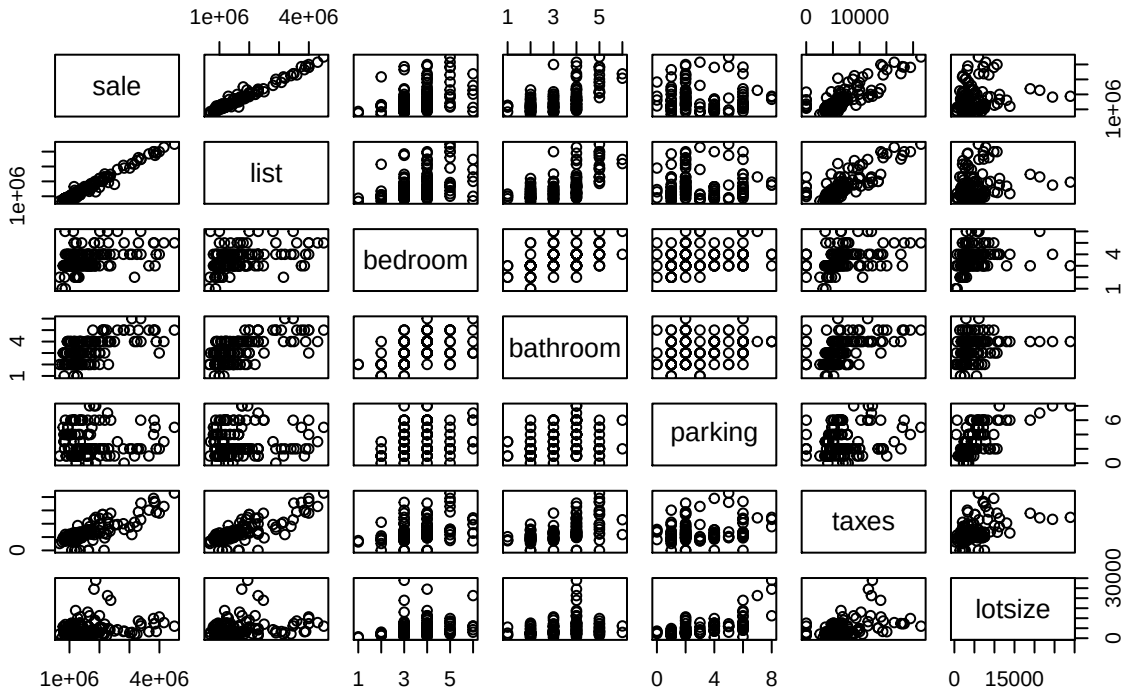
(c) Now we remove some extreme observations. Firstly, we remove three houses whose number of parking spots are the largest. In those points we removed, there are over 10 parking spots in each household, and this is not a representative sample to predict the sale price. Secondly, we remove one house whose number of bathrooms is the largest. Since it differs significantly from other observations, we remove this sample from the data set. Thirdly, we remove two houses which owns the largest number of bedrooms, as those data points are also not representative. Finally, we remove one predictor which variable name is "maxsqfoot" in the data set. The reason why we remove it is that there are too many missing values in this column, so it is hard to use "maxsqfoot" to predict the sale prices of the households. To sum up, we clean the data by removing 6 cases and one predictor. We now have created a new data set called "updated_dataset".

Table 1: Pairwise Correlations - 7813

	sale	list	bedroom	bathroom	parking	taxes	lotsize
sale	1.000	0.985	0.401	0.592	-0.033	0.807	0.206
list	0.985	1.000	0.400	0.614	0.010	0.802	0.228
bedroom	0.401	0.400	1.000	0.498	0.290	0.373	0.272
bathroom	0.592	0.614	0.498	1.000	0.296	0.526	0.297
parking	-0.033	0.010	0.290	0.296	1.000	0.220	0.658
taxes	0.807	0.802	0.373	0.526	0.220	1.000	0.459
lotsize	0.206	0.228	0.272	0.297	0.658	0.459	1.000

II. Exploratory Data Analysis

FIgure 1: Scatterplot Matrix − 7813

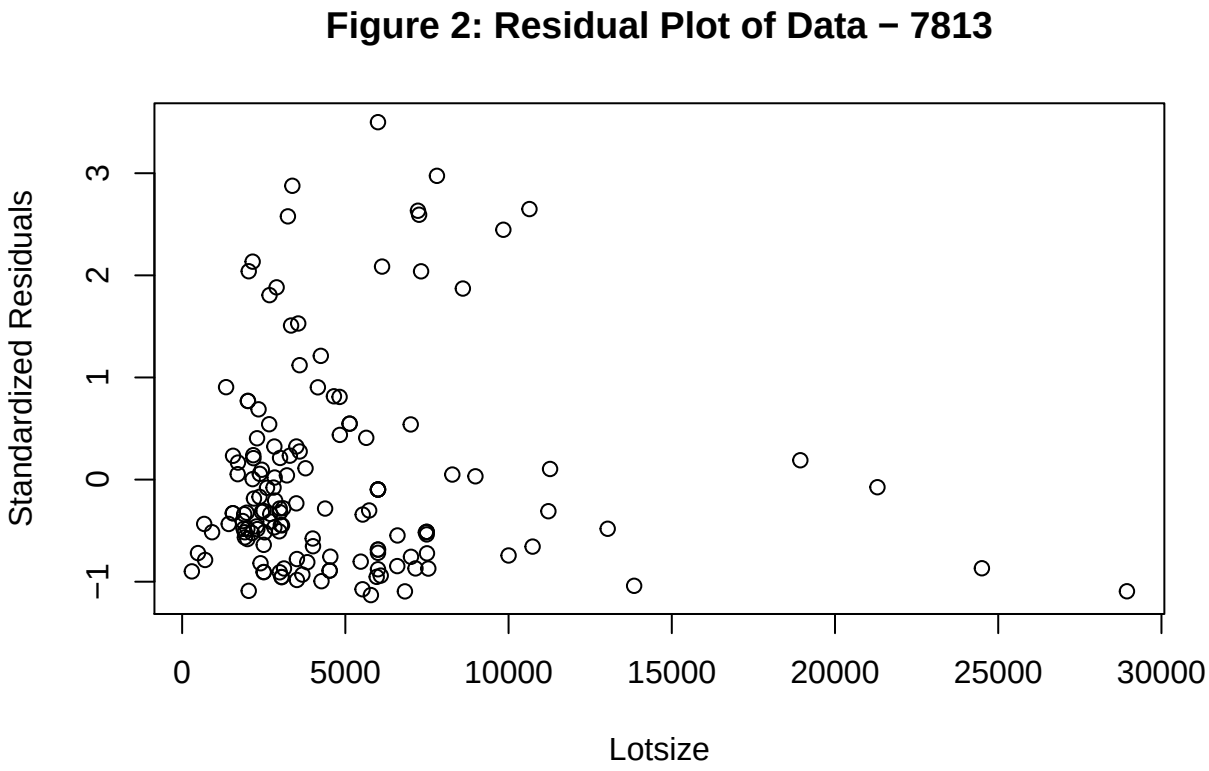


(a) There are eleven variables in the data set. The continuous variables are sale, list, maxsqfoot, taxes, lotwidth, lotlength, and lotsize. The discrete variables are bedroom, bathroom, and parking. There is only one categorical variable "location" in the data.

(b) Table 1 is the pairwise correlations for all pairs of quantitative variables in the data. We can see that from highest correlation coefficients to the lowest correlation coefficients for sale price rank is list, taxes, bathroom, bedroom, lotsize, and parking.

(c) Figure 1 is the scatterplot matrix for all pairs of quantitative variables in the data set. And based on the scatterplot matrix, we can see that the variable "lotsize" is strongly violated the assumption of constant variance. We can see a "horn-shaped" pattern in the scatterplot for lotsize vs. sale price.

Figure 2 is a plot of standardized residuals, and it is also horn-shaped.And this indicates that the variance of the error term is non-constant.



III. Methods and Model

Table 2: estimated coefficients and p-value list - 7813

	(Intercept)	list	bedroom	bathroom	parking	taxes	location	lotsize
estimated	4.969e+04	8.294e-01	2.031e+04	8.498e+03	-1.736e+04	2.139e+01	8.154e+04	4.436e-02
p-value	0.4130	$<$ 2e-16	0.1798	0.5898	0.0813	0.0003	0.0511	0.9913

(i) In Table 2 above, it lists the estimated values and the p-values for the corresponding t-tests for these coefficients. There are two significant t-test results, and the corresponding coefficients are list and taxes. Under the condition that everything else stays the same when the list price going up by 1 Canadian dollar, the expected sale price will increase by 0.8294 Canadian dollars. Under the condition that everything else stays the same when the taxes going up by 1 Canadian dollar, the expected sale price will increase by 0.2139 Canadian dollars.

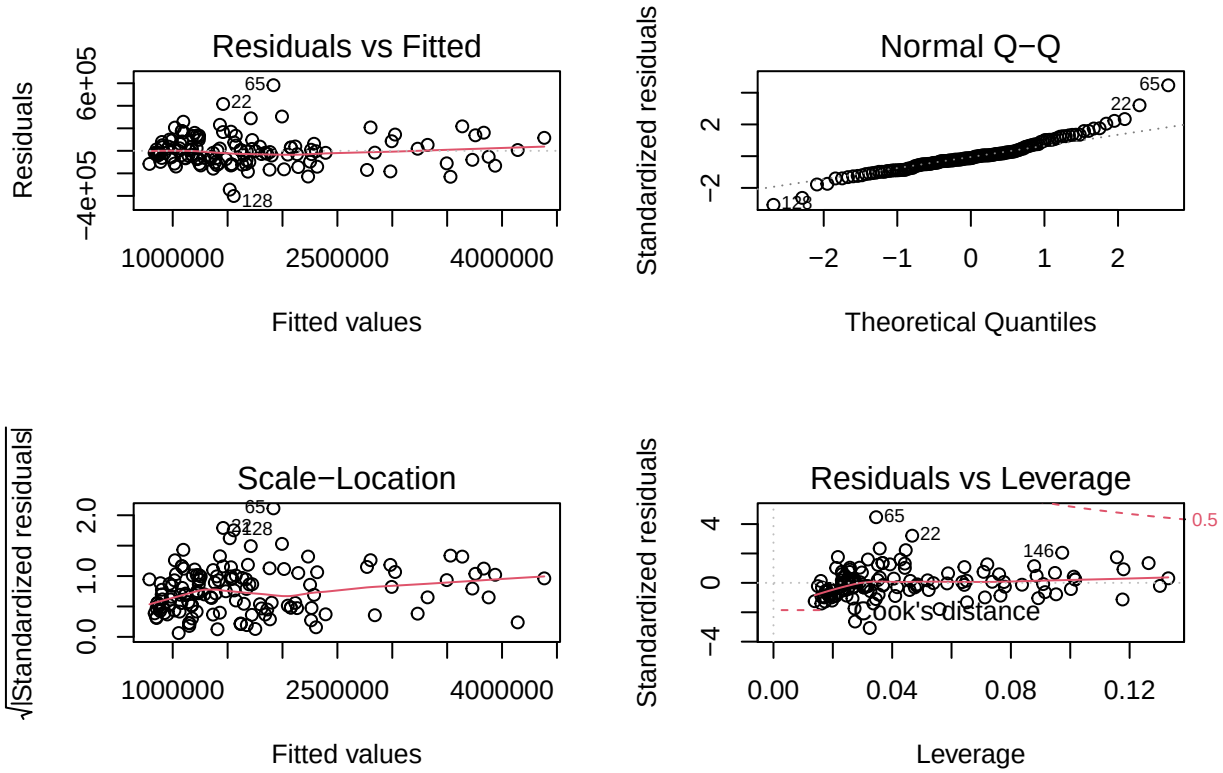
(ii) The final fitted model after using backward elimination with AIC is $\hat{sale} = 62570 + 0.8384 \cdot list + 22720 \cdot bedroom - 17370 \cdot parking + 20.90 \cdot taxes + 73570 \cdot location$ There are five predictors in this fitted model, so it does not consistent with the result by applying t-test on the full model.

(iii) The final fitted model after using backward elimination with BIC is $\hat{sale} = 62570 + 0.8384 \cdot list + 22720 \cdot bedroom - 17370 \cdot parking + 20.90 \cdot taxes + 73570 \cdot location$. The result is consistent with fitted model

after using backward elimination with AIC, but not consistent with the result by applying t-test on the full model.

IV. Discussions and Limitations

(a)



(b) There is no pattern in the first graph "Residuals vs Fitted" and the mean is zero. This indicates that the model is appropriate. In the second graph "Normal Q-Q", the overall appearance is heavy-tailed. Both the ends of the plot deviate from the straight line, but most of the points are on the straight line. Therefore, I think normal error MLR assumptions are not satisfied and there is still room for improvement. In the third graph "Scale-Location", the mean is zero, but we can see a "horn-shaped" pattern. Therefore, it violated the assumption of constant variance. From the fourth graph "Residuals vs Leverage", no case is outside of Cook's distance. So, we can conclude that there is almost no case that can be influential to the regression result.

(c) The next steps I would take towards finding a valid final model are as follow. Firstly, we can apply a transformation on y to achieve normality and constant variance. Secondly, we can also use regression validation to analyze the goodness of fit of the regression.