MLR Data Analysis for the First-Time Home Buying Data

Yu-Chun Chien, 1005194380

December 5, 2020

I. Data Wrangling

Sampling Data

The IDs of the sample selected:

```
##   [1]  48 136 144  20 108  31  68  69 131 154  78 112  38  35  80 175 182   5
##  [19] 180 140 102 166  56 126 186  89 179 106 227  97 193 189  66  34 153 146
##  [37]  41 170 158  14  11 115  91 132   7 110 122  72 191 104  85 161  81  44
##  [55] 183 165  62  88 114 133 125  24 150   9 103 142 138  64  94  43 111 229
##  [73] 151 178 157  87 109 117  74 152   8  19  13 169  12  77  54  30  25 205
##  [91]  45  32  95 159   6 135  42 212 156 218  22  16  23  27  29   2 176 201
## [109]  60  90  76 145 173  26  28 162  93  46 116  49  50  52 155  75   3 148
## [127]  84  61  10 188   4 172  86  99 149 194 107  21  15  71  36  83 113  96
## [145] 100 185 187 105 163  98
```

```
##   [1]   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  19  20  21
##  [19]  22  23  24  25  26  27  28  29  30  31  32  34  35  36  38  41  42  43
##  [37]  44  45  46  48  49  50  52  54  56  60  61  62  64  66  68  69  71  72
##  [55]  74  75  76  77  78  80  81  83  84  85  86  87  88  89  90  91  93  94
##  [73]  95  96  97  98  99 100 102 103 104 105 106 107 108 109 110 111 112 113
##  [91] 114 115 116 117 122 125 126 131 132 133 135 136 138 140 142 144 145 146
## [109] 148 149 150 151 152 153 154 155 156 157 158 159 161 162 163 165 166 169
## [127] 170 172 173 175 176 178 179 180 182 183 185 186 187 188 189 191 193 194
## [145] 201 205 212 218 227 229
```

Data Cleaning

First, since there is too many “NA”s in the predictor “maxsqfoot”, it will not provide too much information to the data. Thus, removing this variable will make further data analysis more efficient.

Next, when we fit a linear model for the data, the entries with “NA” will be automatically removed. Thus, to make our data more concise, remove any remaining data points that have “NA”. By inspection, there are 8 data points with “NA”, which are the ones with ID 79, 76, 61, 84, 89, 96, 109, 113.

II. Exploratory Data Analysis

Classification of Variables

sale: discrete

list: discrete

bedroom: discrete

bathroom: discrete

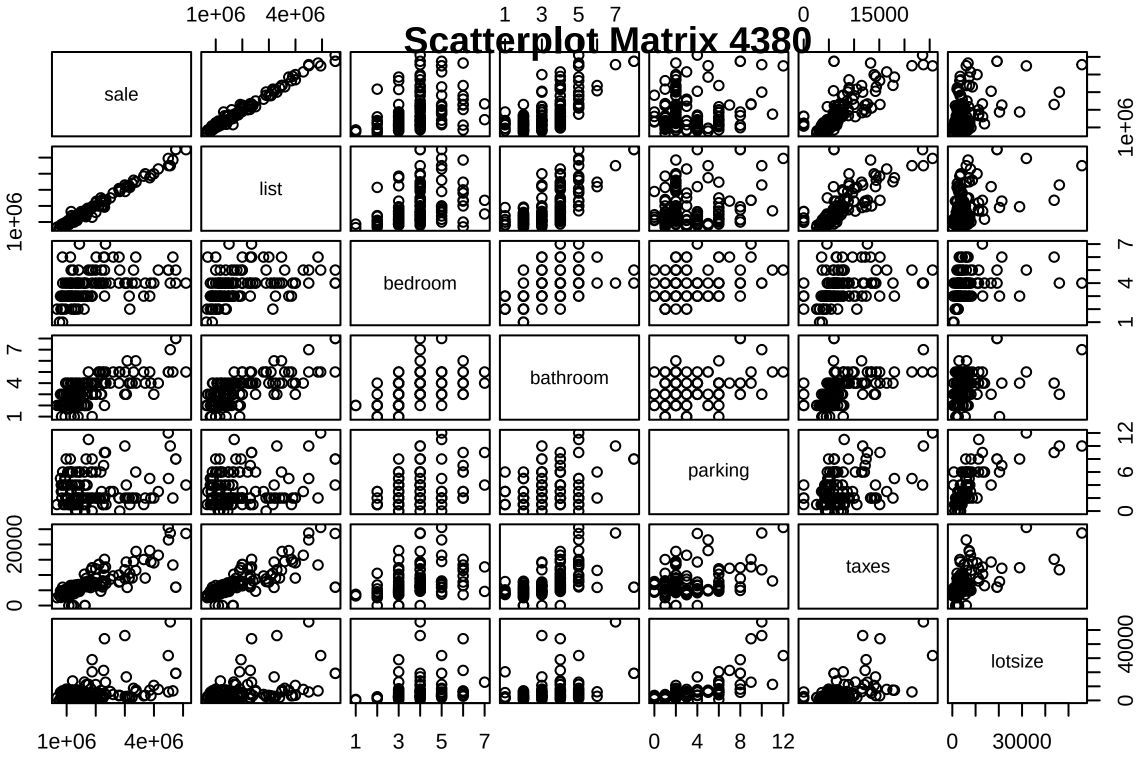
parking: discrete

taxes: continuous

location: categorical

lotsize: continuous

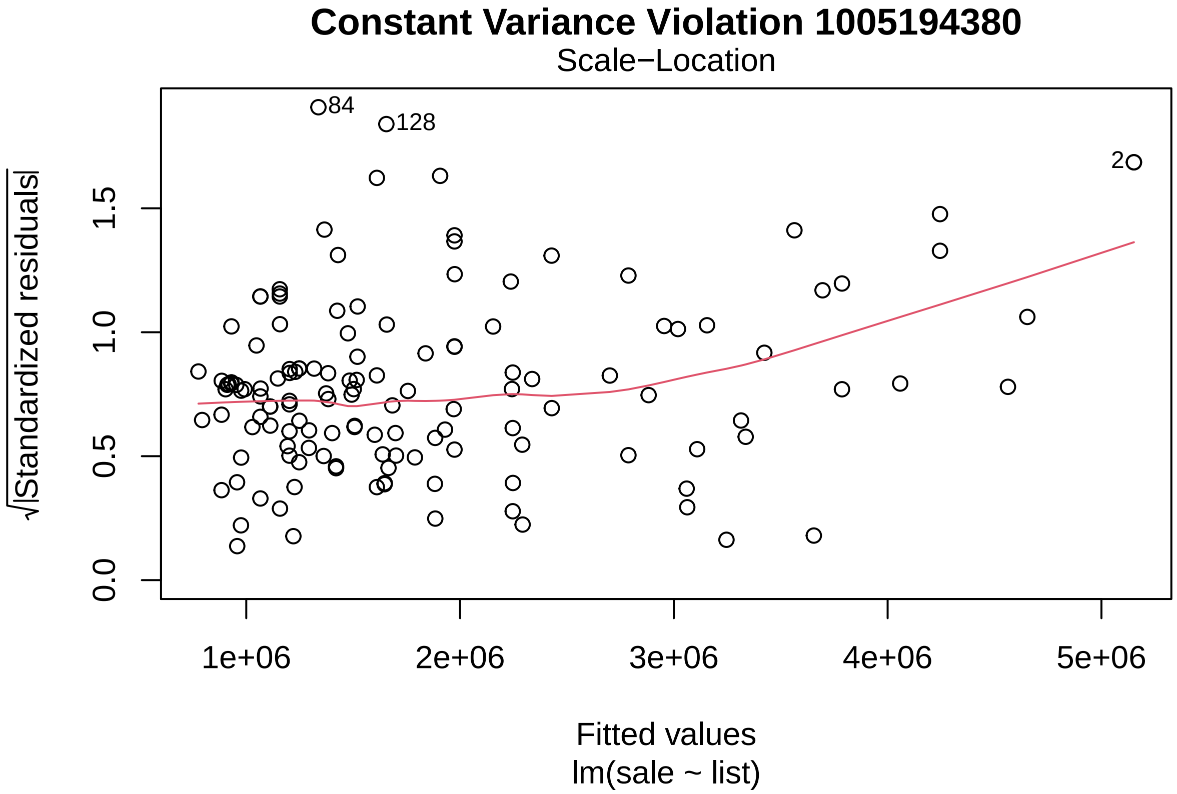
Pairwise Correlations and Scatterplot Matrix



```
##            sale   list bedroom bathroom parking  taxes lotsize
## sale     1.0000 0.9886  0.4217   0.6761  0.1931 0.7628  0.4122
## list     0.9886 1.0000  0.4193   0.6958  0.2382 0.7393  0.4291
## bedroom  0.4217 0.4193  1.0000   0.5204  0.3502 0.3945  0.2847
## bathroom 0.6761 0.6958  0.5204   1.0000  0.3646 0.5258  0.3538
## parking  0.1931 0.2382  0.3502   0.3646  1.0000 0.3829  0.7121
## taxes    0.7628 0.7393  0.3945   0.5258  0.3829 1.0000  0.5705
## lotsize  0.4122 0.4291  0.2847   0.3538  0.7121 0.5705  1.0000
```

Among all quantitative predictor for sale price, list price is the most highly correlated predictor while parking is the least correlated predictor. Rank of correlation coefficient (from highest to lowest): list, taxes, bathroom, bedroom, lotsize, parking

Violation of Constant Variance



Based on the scatterplot matrix, the predictor list would violate the assumption of constants variance. In fact, it is confirmed by the scale-location plot shown above. According to the plot, there is a linear trend between fitted values and the square root of standardized residuals, which indicates that the assumption of constance variance are violated.

III. Methods and Model

Additive Linear Model

Estimated Regression Coefficients for the predictors:

```
##          Coefficient p-value
## list     "0.8324"    "< 2e-16"
## bedroom  "14450"     "0.3209"
## bathroom "3452"      "0.8097"
## parking  "-9934"     "0.3089"
## taxes    "22.1"      "2.29e-06"
## location "101200"    "0.0207"
## lotsize  "-0.6566"   "0.8035"
```

Interpretation of Estimated Model Coefficient:

Using a benchmark significance level of 5%, three coefficients of the predictors are significant: list, taxes, and location.

For the coefficient of list, $p < 2 \times 10^{-16}$, meaning that the probability that the coefficient is equal to zero is less than 2×10^{-16}, which is small. Thus, we can conclude that it is unlikely that the coefficient is equal to zero and that list price predicts sale price. With other predictors having the same value, when the list price increase for 1 unit, it is estimated that the sales price will increase 0.8324 unit on average.

For the coefficient of taxes, $p = 2.29 \times 10^{-6}$, which means that the probability that the coefficient is equal to zero is equal to 2.29×10^{-6}, which is small. This implies that it is unlikely that the coefficient is equal to zero and that taxes predicts sale price. With other predictors having the same value, when the taxes increase for 1 unit, it is estimated that the sales prices will increase 22.1 unit on average.

For the coefficient of location, $p = 0.0207$, which means that the probability that the coefficient is equal to zero is equal to 0.0207, which is smaller than the benchmark $\alpha = 0.05$. Therefore, it is unlikely that the coefficient is equal to zero and that location predicts sale price. With other predictors having the same value, the sales price in Toronto is on average 101200 units more than the sale price in Mississauga.

Backward AIC

The final model obtained by using backward elimination with AIC:

$Y = 6.360 \times 10^4 + 8.340 \times 10^{-1} x_1 + 2.147 \times 10^1 x_2 + 1.372 \times 10^5 d$

Y: sale price

x_1: list price

x_2: taxes

d: location, where “T” = 1 and “M” = 0

The results are consistent with the linear model observed in the full model. In the full model, although there are 7 predictors, only three p-values for the t-tests conducted for the coefficients are significant, which indicates that only three predictors are good predictors.

Backward BIC

The final model obtained by using backward elimination with BIC:

$Y = 6.360 \times 10^4 + 8.340 \times 10^{-1} x_1 + 2.147 \times 10^1 x_2 + 1.372 \times 10^5 d$

Y: sale price

x_1: list price

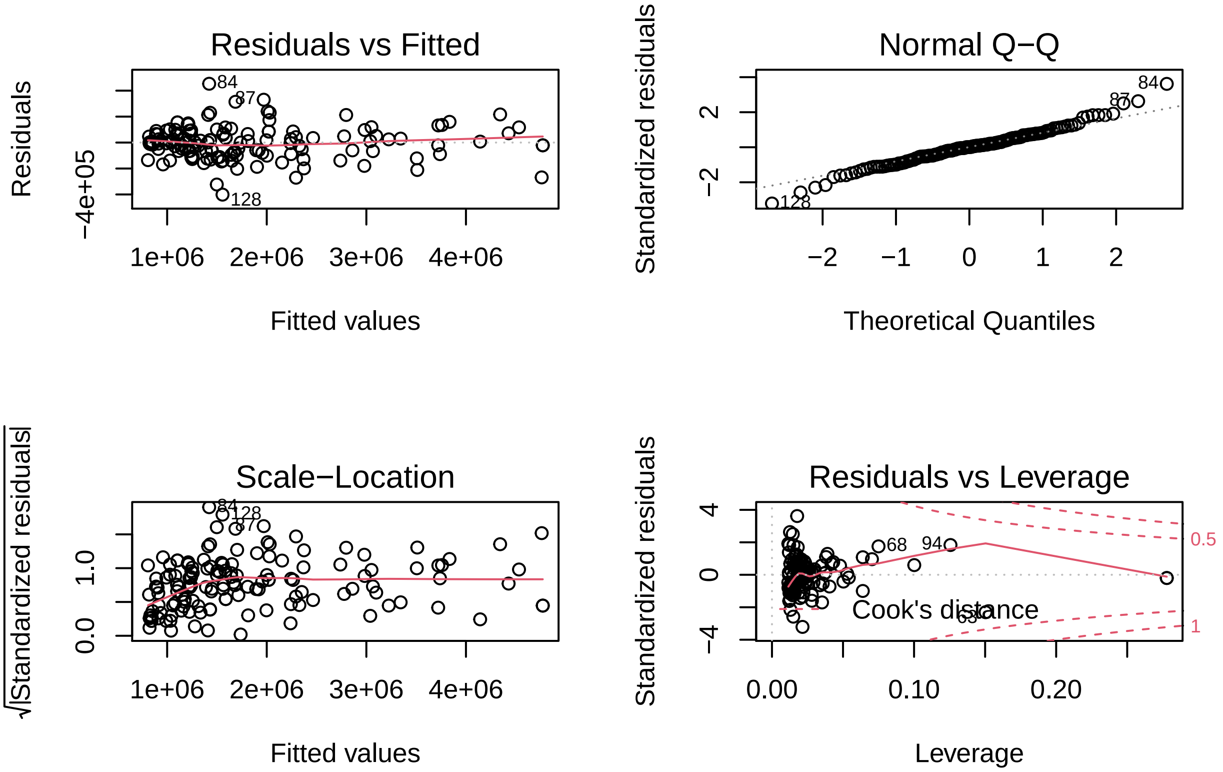
x_2: taxes

d: location, where “T” = 1 and “M” = 0

This model is identical to the one obtained from using AIC. Namely, only list price, taxes, and location are not redundant and are good predictors of sale price. As mentioned in the AIC section, the model obtained by AIC and BIC is consistent with the t-tests of the predictor of the full model, where only list price, taxes, and location are significant under a benchmark of 0.05.

IV. Discussions and Limitations

Diagnostic Plots from Backward BIC Model



Interpretation of the Diagnostic Plots

Residuals vs. Fitted

No distinct pattern is observed in the residual vs. fitted plot. The residuals are equally spread around the horizontal line (residuals = 0), which indicates that there is no non-linear relationship between the predictors and the outcome. .

Normal Q-Q

The majority of the points follow a straight line in general, with only a few points in the two tails deviating from the line. This indicates that the residuals are normally distributed.

Scale-Location

A horizontal line and randomly spread points are observed, which means that the residuals are spread equally along the range of predictors. It can be inferred that the plot follows the assumption of equal variance, which is also called homoscedasticity.

Residuals vs. Leverage

There is no point at the upper and lower right corner and all points are within the dashed line. This indicates that no points have high Cook's distance score, and therefore there is no influential points to the regression line though there might be some outliers. In other words, outliers might exist but there existence did not change the regression line significantly.

Normal Error MLR Assumptions

According to the diagnostic plots, the normal error MLR assumptions are all satisfied:

1. The error terms have mean zero: observed in the residual vs. fitted plot, where the horizontal line has intercept equals to zero.
2. The error terms have constant variance: observed in the scale-location plot, where we observed a line that is roughly horizontal.
3. The errors are uncorrelated: observed in the residual vs. fitted plot, where we observe a null plot with the residuals having no pattern.
4. Jointly Normal - the error terms follow a multivariate normal distribution: observed in the Normal QQ plot, where the majority of the points follow a straight line.

Next Steps

The model obtained by BIC and AIC follows the MLR assumptions according to the diagnostic plots and it has no influential points. The next steps that I would take towards finding a valid 'final' model will be to check for the results of Global F-test and Individual t-tests to get a better sense of whether or not there really are some useful explanatory variables for predicting sales price and if there might be multicollinearity among the predictors. If the comparison of F and t-tests indicates that there might be multicollinearity, I will use the variance inflation factors to check the degree of multicollinearity and determine whether or not to fix it with any transformation methods.