The MLR research on sale price and its predictors.

Zian Lu, Id 1004710280

December 1, 2020

I. Data Wrangling

a)

```
##   [1] 156 179  10 150 157  40  39   8 229  17   7  71 152   4 125 126  31  14
##  [19]  72   2  57 196 207 148 151 194 187 175  84  55  12 193   9 167  25  90
##  [37] 195 131  99  35 154 173  87 141   5 109  30  49 164 185 146 111  93  94
##  [55]  13 189  42 132  64  78 142  45 159  70  79 149  47  60 107  51 177  97
##  [73]  88 139 136  61  85 108  28 117  69 183  92  50 140  48 112 106  65 122
##  [91] 137 176  26 186 162 165 119  44 138  43 145   6 114 163 227  29 101 170
## [109] 160 153 204  96  67  41 133 178 105  95  56  77  38 158 188 147 155 100
## [127] 103  53  46 166  33 104  66  52  83  22 191  21  58  73 205 190 180  54
## [145]  81 182 212  32  62   3
```

b)

- lotsize=lotwidth*lotlength

c)

- I choose to remove the predictor maxsqfoot because it has 90 NA data points out of 150 cases and I believe the remaining 60 cases cannot reveal the overall maxsqfoot level of detached house.Another reason is that the meaning of maxsqfoot and lotsize are quite similar so one should be enough and lotsize should be better since it has fewer NA values.
- I removed 5 cases who has missing value on parking, 1 case with missing value on taxes and 2 cases with missing cases on lotsize

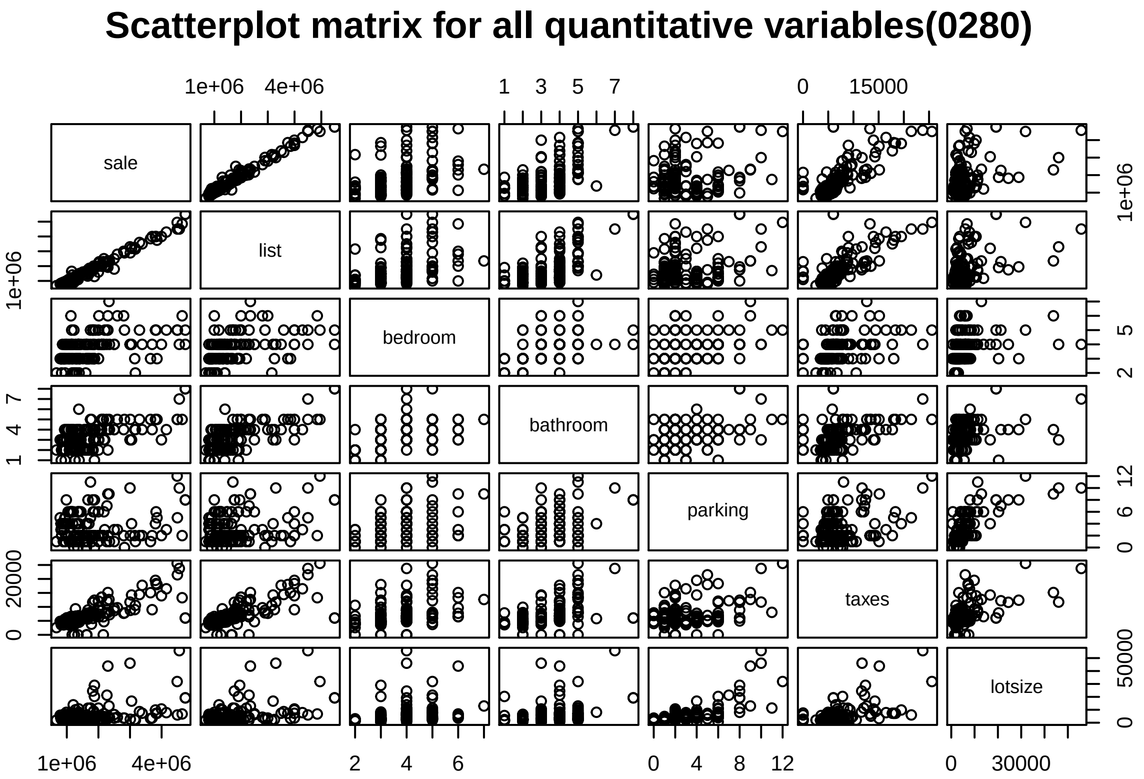
II. Exploratory Data Analysis

a)

- categorical variable: Location
- discrete variable: bedroom, bathroom, parking
- continuous variable: sale, list, taxes, lotsize

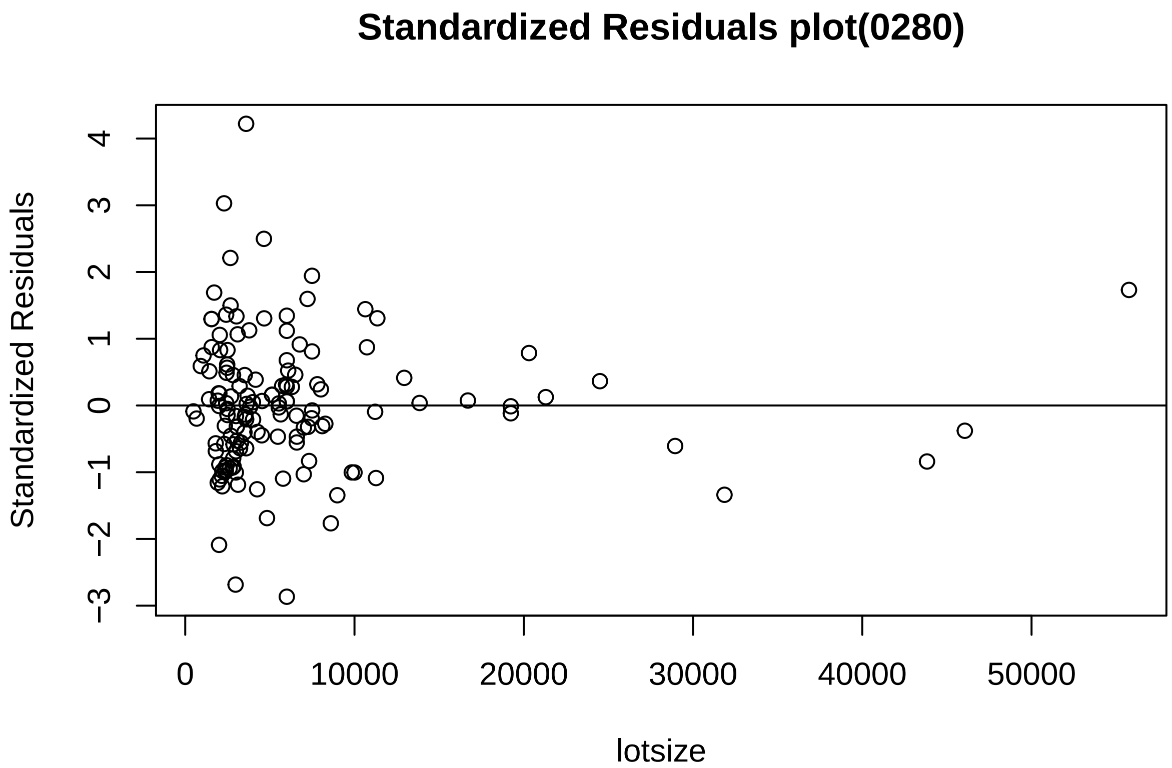
b)

```
##             sale   list bedroom bathroom parking  taxes lotsize
## sale      1.0000 0.9867  0.4439   0.6230  0.2182 0.7622  0.4244
## list     0.9867 1.0000  0.4448   0.6435  0.2646 0.7363  0.4405
## bedroom  0.4439 0.4448  1.0000   0.5249  0.3704 0.4215  0.2977
## bathroom 0.6230 0.6435  0.5249   1.0000  0.4180 0.4733  0.3645
## parking  0.2182 0.2646  0.3704   0.4180  1.0000 0.3702  0.7142
## taxes    0.7622 0.7363  0.4215   0.4733  0.3702 1.0000  0.5526
## lotsize  0.4244 0.4405  0.2977   0.3645  0.7142 0.5526  1.0000
```



- The quantitative predictor for sale price rank(in terms of correlation coefficient from highest to lowest): list, taxes, bathroom, bedroom, lotsize, parking.

c)



Based on the scatterplot, the predictor lotsize strongly violated the assumption of constant variance. - By checking the standardized residual plot of lotsize, it turns out that there's a cone pattern at y=o and thus demonstrates that the constant variance assumption is not satisfied.

III. Methods and Model

i)

```
##
## Call:
## lm(formula = sale ~ list + bedroom + bathroom + taxes + parking +
##     lotsize + location, data = datafinalZ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -378905  -77776   -5360   63462  558810
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.759e+04  5.649e+04   1.019   0.3098
## list         8.131e-01  2.153e-02  37.758  < 2e-16 ***
## bedroom      1.202e+04  1.435e+04   0.838   0.4037
## bathroom     1.670e+04  1.378e+04   1.212   0.2277
## taxes        2.166e+01  4.148e+00   5.222 6.58e-07 ***
## parking     -1.812e+04  8.643e+03  -2.097   0.0379 *
## lotsize      2.885e+00  2.299e+00   1.255   0.2118
## locationT    1.071e+05  3.826e+04   2.798   0.0059 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 134800 on 134 degrees of freedom
## Multiple R-squared:  0.9814, Adjusted R-squared:  0.9804
## F-statistic:  1011 on 7 and 134 DF,  p-value: < 2.2e-16
```

- list,taxes,parking and location are significant since there p-value are all smaller than the cut off 5%
- the coefficient of the list price means for every 1 dollar increase in the list price, the sale price of detached house increase by 8.131e-01 on average.
- the coefficient of the taxes means for every 1 dollar increase in the taxes, the sale price of detached house increase by 2.166e+01 on average.
- the coefficient of the number of parking means for every 1 parking slot increase in the number of parking, the sale price of detached house decrease by 1.812e+04 on average.
- the coefficient of the locationT means holding every other predictors constant, the average sale price of detached house in Toronto is 1.071e+05 higher than the sale price in Mississauga.

ii)

```
## Start:  AIC=3362.16
## sale ~ list + bedroom + bathroom + taxes + parking + lotsize + 
##     location
## 
##            Df  Sum of Sq        RSS    AIC
## - bedroom   1 1.2743e+10 2.4462e+12 3360.9
## - bathroom  1 2.6666e+10 2.4601e+12 3361.7
## - lotsize   1 2.8589e+10 2.4621e+12 3361.8
## <none>                   2.4335e+12 3362.2
## - parking   1 7.9844e+10 2.5133e+12 3364.7
## - location  1 1.4220e+11 2.5757e+12 3368.2
## - taxes     1 4.9529e+11 2.9288e+12 3386.5
## - list      1 2.5890e+13 2.8323e+13 3708.7
## 
## Step:  AIC=3360.9
## sale ~ list + bathroom + taxes + parking + lotsize + location
## 
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 2.5601e+10 2.4718e+12 3360.4
## <none>                   2.4462e+12 3360.9
## - bathroom  1 4.2336e+10 2.4886e+12 3361.3
## - parking   1 6.9857e+10 2.5161e+12 3362.9
## - location  1 1.5434e+11 2.6006e+12 3367.6
## - taxes     1 5.2899e+11 2.9752e+12 3386.7
## - list      1 2.5896e+13 2.8342e+13 3706.8
## 
## Step:  AIC=3360.38
## sale ~ list + bathroom + taxes + parking + location
## 
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 3.2443e+10 2.5043e+12 3360.2
## <none>                   2.4718e+12 3360.4
## - parking   1 4.5799e+10 2.5176e+12 3361.0
## - location  1 1.4088e+11 2.6127e+12 3366.3
## - taxes     1 6.3098e+11 3.1028e+12 3390.7
## - list      1 2.7253e+13 2.9724e+13 3711.5
##
## Step:  AIC=3360.23
## sale ~ list + taxes + parking + location
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   2.5043e+12 3360.2
## - parking   1 4.4298e+10 2.5486e+12 3360.7
## - location  1 1.1321e+11 2.6175e+12 3364.5
## - taxes     1 6.0642e+11 3.1107e+12 3389.0
## - list      1 4.2883e+13 4.5388e+13 3769.6
```

The final model is

$$\hat{saleprice} = 5.759 * 10^4 + 0.813 * listprice + 21.666 * taxes - 1.812 * 10^4 * parking + 1.071 * 10^5 * locationT$$

- locationT=1 for location is Toronto otherwise locationT=0 for location is Mississauga - The results are consistent with those in part i

iii)

```
## Start:  AIC=3385.81
## sale ~ list + bedroom + bathroom + taxes + parking + lotsize +
##     location
##
##            Df  Sum of Sq        RSS    AIC
## - bedroom   1 1.2743e+10 2.4462e+12 3381.6
## - bathroom  1 2.6666e+10 2.4601e+12 3382.4
## - lotsize   1 2.8589e+10 2.4621e+12 3382.5
## - parking   1 7.9844e+10 2.5133e+12 3385.4
## <none>                   2.4335e+12 3385.8
## - location  1 1.4220e+11 2.5757e+12 3388.9
## - taxes     1 4.9529e+11 2.9288e+12 3407.2
## - list      1 2.5890e+13 2.8323e+13 3729.4
##
## Step:  AIC=3381.59
## sale ~ list + bathroom + taxes + parking + lotsize + location
##
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 2.5601e+10 2.4718e+12 3378.1
## - bathroom  1 4.2336e+10 2.4886e+12 3379.1
## - parking   1 6.9857e+10 2.5161e+12 3380.6
## <none>                   2.4462e+12 3381.6
## - location  1 1.5434e+11 2.6006e+12 3385.3
## - taxes     1 5.2899e+11 2.9752e+12 3404.4
## - list      1 2.5896e+13 2.8342e+13 3724.5
##
## Step:  AIC=3378.12
## sale ~ list + bathroom + taxes + parking + location
##
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 3.2443e+10 2.5043e+12 3375.0
## - parking   1 4.5799e+10 2.5176e+12 3375.8
## <none>                   2.4718e+12 3378.1
## - location  1 1.4088e+11 2.6127e+12 3381.0
## - taxes     1 6.3098e+11 3.1028e+12 3405.4
## - list      1 2.7253e+13 2.9724e+13 3726.3
##
## Step:  AIC=3375.01
## sale ~ list + taxes + parking + location
##
##            Df  Sum of Sq        RSS    AIC
## - parking   1 4.4298e+10 2.5486e+12 3372.5
## <none>                   2.5043e+12 3375.0
## - location  1 1.1321e+11 2.6175e+12 3376.3
## - taxes     1 6.0642e+11 3.1107e+12 3400.8
## - list      1 4.2883e+13 4.5388e+13 3781.5
##
## Step:  AIC=3372.55
## sale ~ list + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   2.5486e+12 3372.5
## - location  1 5.5425e+11 3.1028e+12 3395.5
## - taxes     1 5.7057e+11 3.1191e+12 3396.3
## - list      1 4.5446e+13 4.7995e+13 3784.4
```

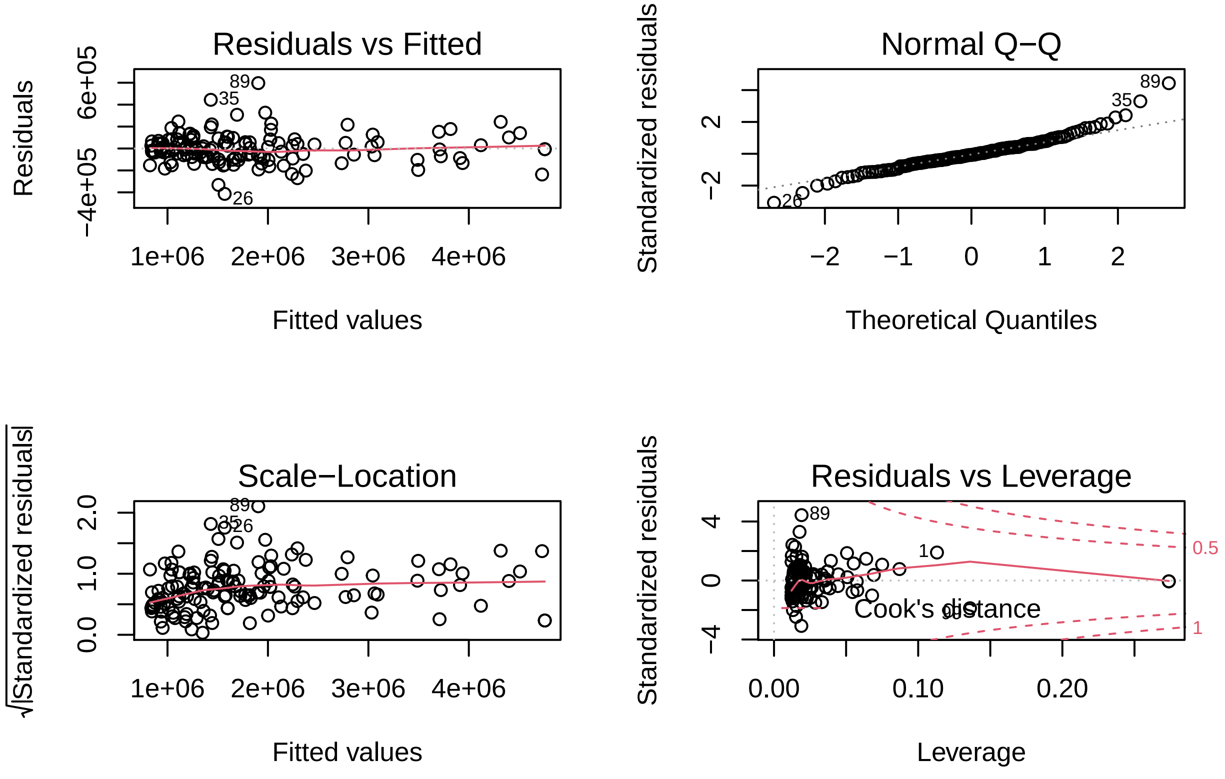
- The final model is

$$\hat{saleprice} = 5.759 * 10^4 + 0.813 * listprice + 21.666 * taxes + 1.071 * 10^5 * locationT$$

- The results are not consistent with those in part i
- The reason that the results are different from the difference in the evaluation of different method. Since BiC penalize complex model more heavily than AIC, thus favors simpler models than AIC and this explains why there are fewer predictors in ii than i.

IV. Discussions and Limitations

a)



b)

- residuals vs fitted plot: there is no pattern around the 0 horizontal line and points spread randomly
- normal QQplot: a majority of points fall on the 45 degree line and thus the normal error assumption are satisfied
- scale-location plot: a random scatter of points around the horizontal axis, no pattern or trend are found
- residuals vs leverage:there are no points that lies outside of the red line region indicating that there are no outliers or influential points

c)

- Discuss whether there are other predictors and if we add them in, will our model becomes better or not. For example, the age of the detached house may also affect the sale price. Added variable plot are helpful when considering the introduction of an additional predictor variable.
- Use the variance inflation factors method to decide whether the multicollinearity exist between existing model.
- We could use the global F-test and Individual t-test to help confirm the correctness of AIC and BIC
- Other methods, for example, penalized regression,cross validation and the data of adjusted R square can be used to confirm the accuracy of model. Check the difference between the model generated by different model and see which one is. better