STA 302 A3: MLR Model for Toronto and Mississauga House Prices

Yichen Ji ID:1004728967

2020/11/28

Introduction

In this report, I will use TREB data and construct a MLR model to help predict the sale price of single-family, detached houses in two neighborhoods in the GTA area.

I. Data Wrangling

(a) We first randomly select a sample of 150 cases and report their IDs:

```
##   [1]  45  39  77 160 178 136  58 122  87 132  75 117  46 172 113  83  36  23
##  [19] 171  72 145 154  80   3  64 157 139  21  38  55 179 118  11 141 111  84
##  [37]   5 189  60 110  56  12 153 123  49 109 173  95  17  25  57 100 181   6
##  [55] 163 170  26 105 140   7  90  22  91 159  89 131 102 147  20 107 186   1
##  [73] 142 162 164   8  24  94 166  97 156  54  35  37 128 108   9  85   4 127
##  [91] 175 169 183 152 182 112 101 151  28  86  74 134 137  16  27  61  48  19
## [109] 133  79  59  41  73 146 124  51 191  32 129   2  52  93 150  18 106 155
## [127] 176  76 187 116 126  47  34  70  44  40 103 188  14 168 121 119 120  66
## [145] 192 185 148  99  88 143
```

(b) Then we use a new variable called 'lotsize' to replace 'lotwidth' and 'lotlength':

```
lotsize = sample8967$lotlength * sample8967$lotwidth
sampleYJ = select (sample8967,-c(lotwidth, lotlength))
sampleYJ$lotsize = lotsize # replace lotwidth, lotlength with lotsize
```

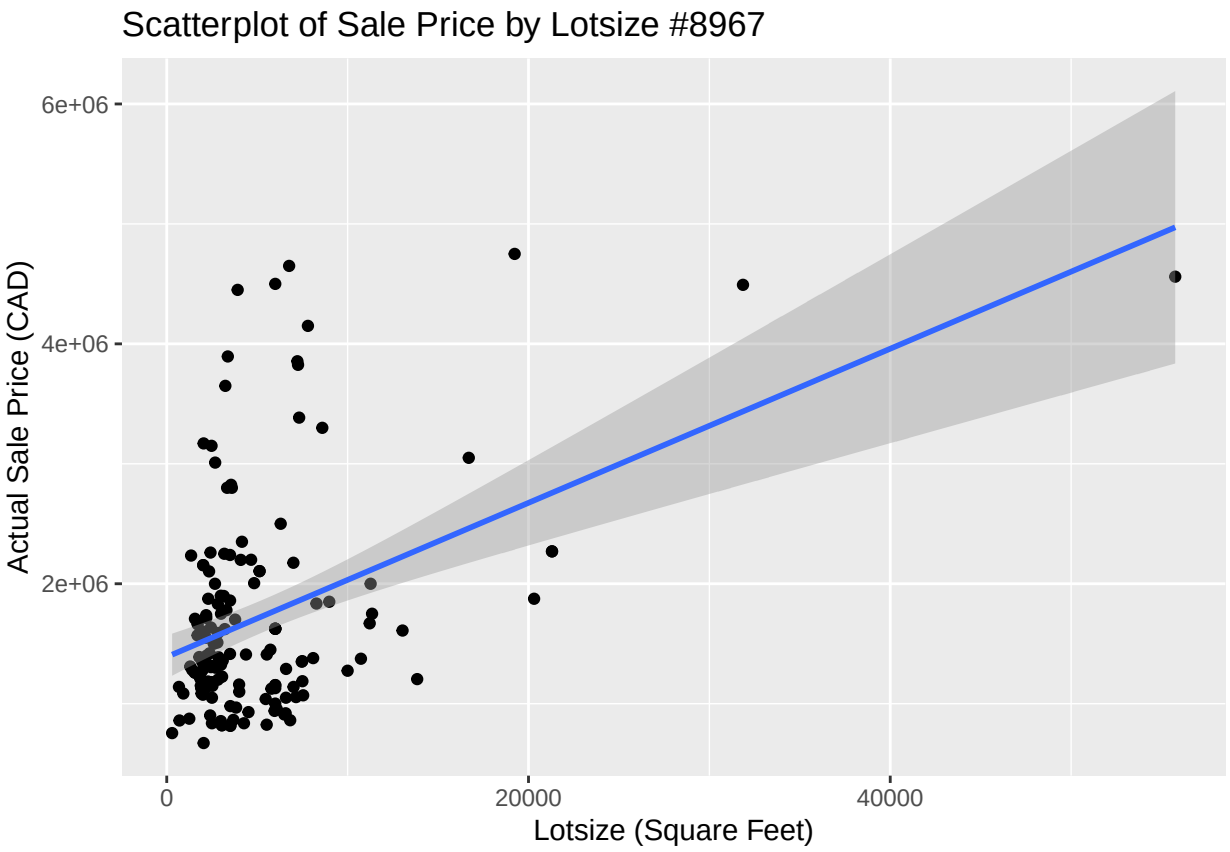
(c) Now we clean the data by first looking at the summary of the sample:

```
##        ID              sale             list            bedroom
##  Min.   :  1.00   Min.   : 672000   Min.   : 649000   Min.   :1.00
##  1st Qu.: 48.25   1st Qu.:1142250   1st Qu.:1091425   1st Qu.:3.00
##  Median :101.50   Median :1410000   Median :1424500   Median :4.00
##  Mean   :101.93   Mean   :1750075   Mean   :1750255   Mean   :3.64
##  3rd Qu.:154.50   3rd Qu.:2003750   3rd Qu.:1999000   3rd Qu.:4.00
##  Max.   :229.00   Max.   :5100000   Max.   :5499000   Max.   :7.00
##
##     bathroom        parking         maxsqfoot        taxes
##  Min.   :1.000   Min.   : 0.000   Min.   :1500   Min.   :    4.375
##  1st Qu.:2.000   1st Qu.: 2.000   1st Qu.:2375   1st Qu.: 4508.000
```

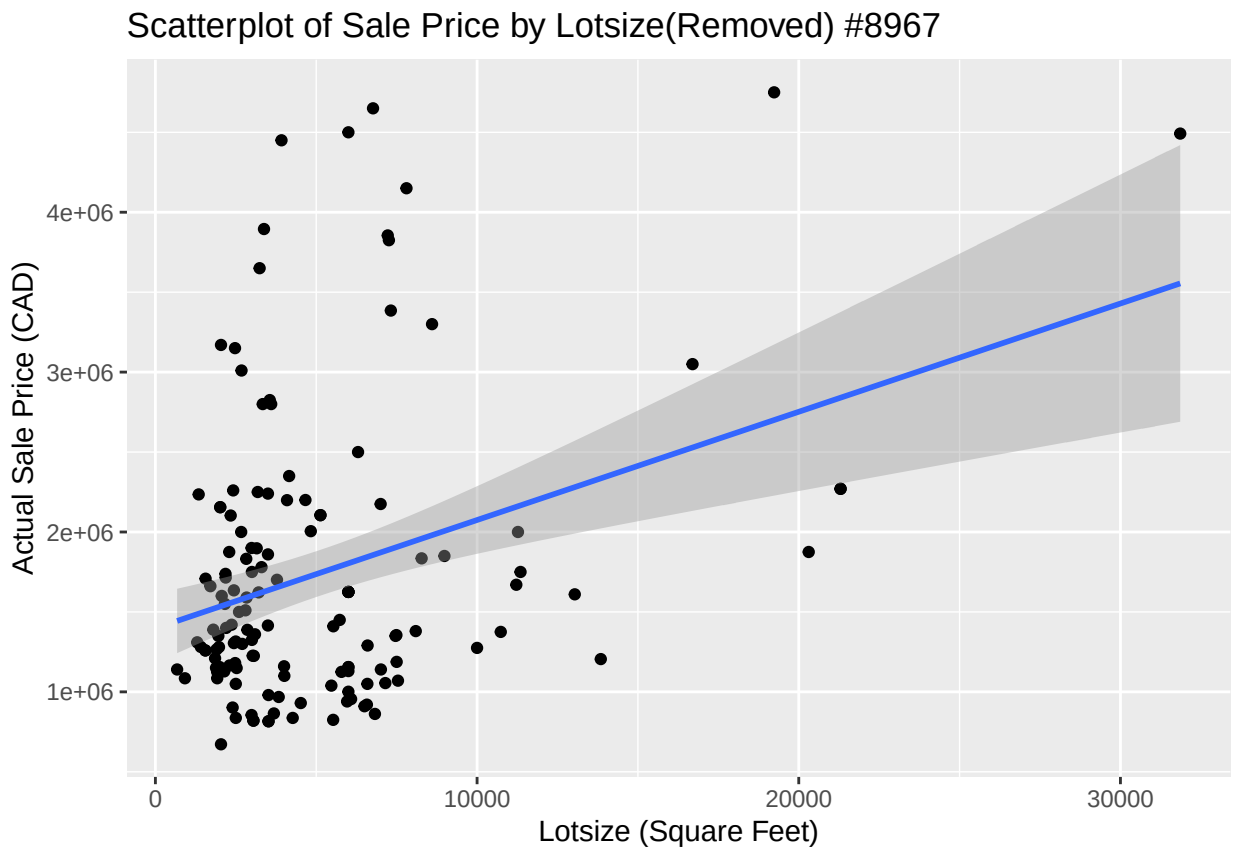
```
##  Median :3.000   Median : 2.000   Median :3000   Median : 6017.000
##  Mean   :3.313   Mean   : 3.147   Mean   :2875   Mean   : 7038.523
##  3rd Qu.:4.000   3rd Qu.: 4.000   3rd Qu.:3500   3rd Qu.: 7698.000
##  Max.   :8.000   Max.   :12.000   Max.   :5000   Max.   :25575.000
##                  NA's   :7        NA's   :90     NA's   :1
##    location            lotsize
##  Length:150         Min.   :  297.4
##  Class :character   1st Qu.: 2362.6
##  Mode  :character   Median : 3498.5
##                     Mean   : 5289.4
##                     3rd Qu.: 6021.0
##                     Max.   :55756.0
##                     NA's   :2
```

We can see that there are several missing values in 'parking', 'taxes' and 'lotsize' (7, 1 and 2 respectively). Also, 'maxsqfoot' has 90 missing values, which would give us a big hurdle to interpret the result when we include this variable and run MLR. Therefore, we remove 'maxsqfoot' as well as those 10 cases containing missing values.

Then, if we have a glimpse at the scatter plot of sale price by lotsize:



There are two high leverage points (lotsize>30000) lying off from the pattern of the bulk of data. Since we are only allowed to remove at most 11 cases, we only remove the point with highest leverage(lotsize>40000):



II. Exploratory Data Analysis

(a) Here's the classification of variables:

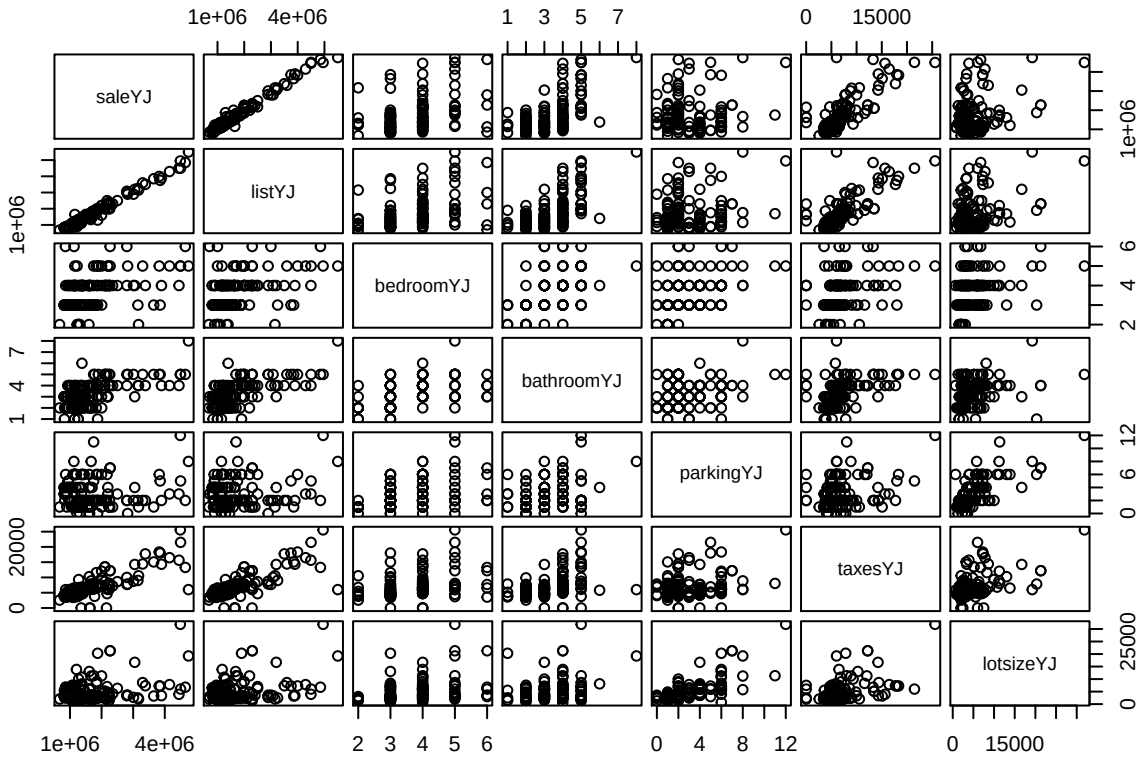
Categorical Variables: location

Discrete Variables: ID, bedroom, bathroom, parking

Continuous Variables: sale, list, taxes, lotsize

(b) Here are the pairwise correlation matrix and scatter plot matrix for all pairs of quantitative variables i.e. without 'location' and 'ID':

```
##            saleYJ listYJ bedroomYJ bathroomYJ parkingYJ taxesYJ lotsizeYJ
## saleYJ     1.0000 0.9884    0.4210     0.5960    0.1035  0.7916    0.3374
## listYJ     0.9884 1.0000    0.4301     0.6174    0.1538  0.7780    0.3785
## bedroomYJ  0.4210 0.4301    1.0000     0.5338    0.3749  0.3675    0.3632
## bathroomYJ 0.5960 0.6174    0.5338     1.0000    0.3151  0.4479    0.3152
## parkingYJ  0.1035 0.1538    0.3749     0.3151    1.0000  0.2796    0.7170
## taxesYJ    0.7916 0.7780    0.3675     0.4479    0.2796  1.0000    0.4899
## lotsizeYJ  0.3374 0.3785    0.3632     0.3152    0.7170  0.4899    1.0000
```

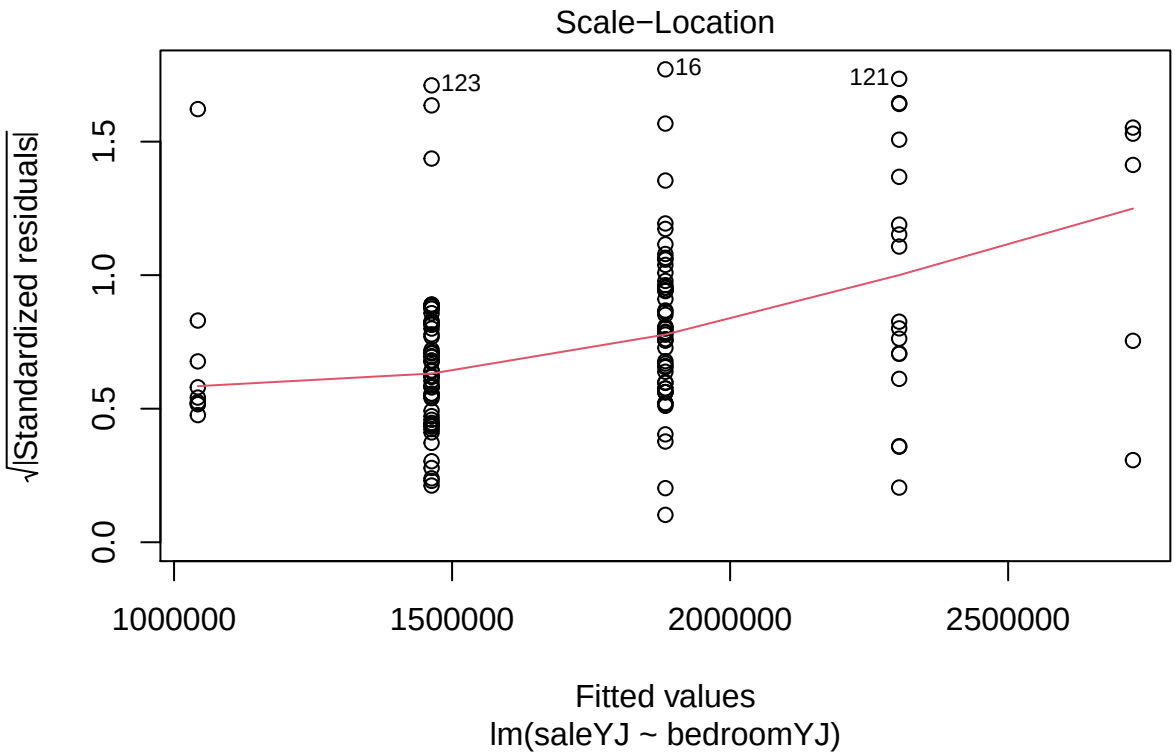


Then we rank the predictors in terms of their correlation coefficient for sale price (from the highest to lowest):

```
##     saleYJ     listYJ  bedroomYJ bathroomYJ  parkingYJ    taxesYJ  lotsizeYJ
##          1          2          5          4          7          3          6
```

We get: list > taxes > bathroom > bedroom > lotsize > parking.

(c) If we check the diagnostic plot of sale price on bedroom:



Based on the scale-location plot, the red smooth line is not horizontal but upward sloping and the standardized residuals spread wider and wider, indicating a strong violation against the assumption of equal variance (homoscedasticity).

III. Methods and Model

(i) First, fit an additive regression model:

```
full.lm = lm(saleYJ ~ listYJ + taxesYJ + lotsizeYJ + bedroomYJ + bathroomYJ + parkingYJ + locationYJ)
```

Then we list their estimated coefficients and p-values:

```
##             coefficient p_value
## (Intercept)  64891.2882  0.2360
## listYJ           0.8321  0.0000
## taxesYJ         21.6510  0.0000
## lotsizeYJ       -2.6520  0.5013
## bedroomYJ    13616.6252  0.3420
## bathroomYJ    8212.1678  0.5419
## parkingYJ   -11869.7377  0.1488
## locationYJT  84928.0526  0.0261
```

As we can see, there are 3 significant t-test results (list, taxes and location) by the 5% significance level. The interpretation of each coefficient is:

1. Holding other factors fixed, an additional dollar increase in the last list price is expected to increase \$0.8321 in the mean actual sale price.
2. Holding other factors fixed, an additional dollar increase in the previous year's taxes is expected to increase the mean actual sale price by \$21.651.
3. Holding other factors fixed, the properties in Toronto Neighborhood are expected to have the mean actual sale price about \$84928 higher than those in Mississauga Neighborhood.

(ii) If we perform backward elimination with AIC:

Leaving out the steps, the final model using backward elimination with AIC is

$$sale = 1.189 * 10^5 + 0.8432\ list + 20.36\ taxes - 1.223 * 10^4\ parking + 8.583 * 10^4\ location$$

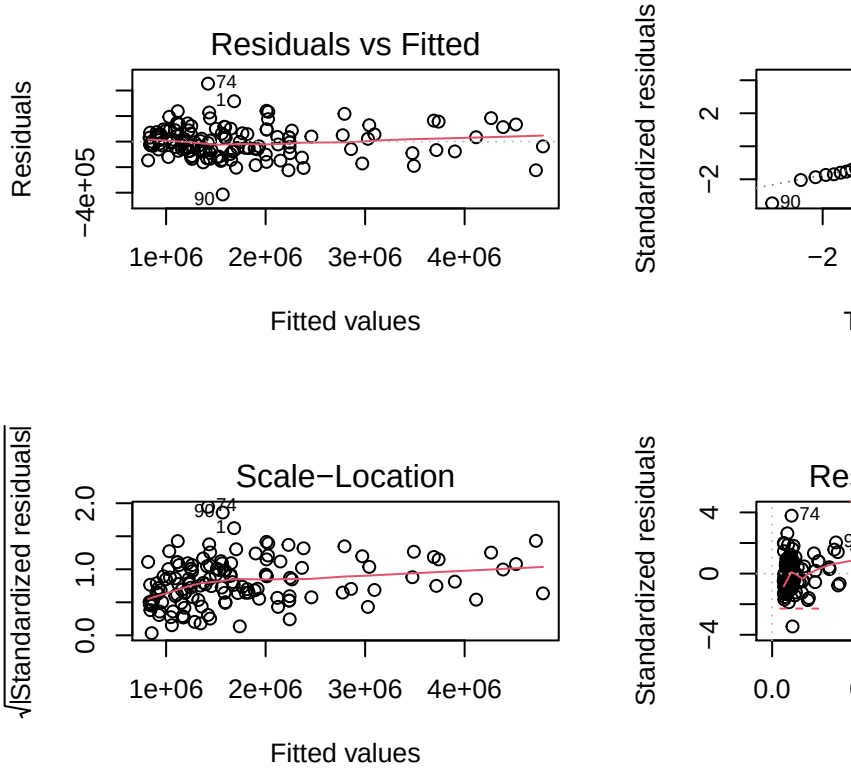
This result is inconsistent with those in (i) since one insignificant predictor 'parking' whose p-values > 0.05 is still in the final model. That's because the penalty for model complexity is not strong enough, so AIC overfits the sample.

(iii) The final model using backward elimination with BIC is

$$sale = 7.325 * 10^4 + 0.8356\ list + 19.84\ taxes + 1.27 * 10^5\ location$$

This result is consistent with our t-test and p-value output and inconsistent with backward AIC tautologically. BIC penalizes complex model more heavily than AIC, thus favors simpler models than AIC.

IV. Discussion and Limitations



(a) Diagnostic plots for our MLR additive model:

(b) Interpretation of residual plots:

- Residuals v.s. Fitted plot: The residual points are randomly scattered and the red smooth line is horizontally lying around 0, which is a good sign that residuals are uncorrelated with the fitted values and there is no non-linear relationship.
- Normal Q-Q plot: Except for few outliers e.g. #1, #74, #90, most points follow along the straight line, so we can say residuals are normally distributed.
- Scale-Location plot: The distribution of standardized residuals has no distinct trend and data-points are randomly spread, similar to the Residual v.s. Fitted plot.
- Residuals v.s. Leverage plot: There is no noteworthy point and all points are inside of the Cook's distance line, but there are few points with high leverage.

As for MLR assumptions, I think all of them are well established(linearity, errors being uncorrelated with 0 mean, homoscedasticity, normality).

(c) Next steps towards finding a valid model:

1. Since Our goal is to predict the sale price, we can apply k-fold cross-validation to assess its predictive ability.
2. We can also use the added variable plot to show the relationship between the response variable and one of the predictors after controlling for the presence of the other predictors.
3. There are few leverage points that weren't considered since we are only allowed to remove 11 cases, so I'd like to identify and remove the outliers, then refit the model.