Predicting sale price of detached houses in GTA by MLR method

Jiayang Wang, Id 1003790935

December 5, 2020

The increasing price of detached houses in GTA due to COVID-19 leads us to investigate the sale price of detached houses. Using the dataset from the Toronto Real Estate Board (TREB), we build MLR models to predict the sale price of properties by some predictors to help buyers predict the expected detached houses' sale price.

I. Data Wrangling

Our data is obtained by random selection from the dataset of TREB, including 150 observations and 11 variables: id, sale, list, bedroom, bathroom, parking, maxsqfoot, taxes, lotwidth, lotlength, location.

IDs of sample data:

```
##   [1]   1   2   4   5   7   8   9  10  11  12  14  15  17  19  20  21  22  23
##  [19]  25  26  28  30  31  32  33  34  35  36  37  38  39  40  41  42  44  45
##  [37]  46  47  48  50  53  54  55  56  60  61  62  63  64  65  66  67  68  69
##  [55]  70  71  72  74  76  77  78  79  81  82  83  84  85  87  88  89  90  91
##  [73]  93  94  95  96  97  98  99 102 103 105 106 108 109 110 112 113 114 116
##  [91] 118 119 125 126 131 132 133 134 135 136 138 139 140 141 143 144 146 149
## [109] 150 151 153 154 157 158 159 160 161 162 163 164 165 166 167 168 169 170
## [127] 171 172 173 174 176 179 180 181 183 185 186 187 188 191 194 196 201 204
## [145] 205 207 212 218 227 229
```

We firstly create a new variable lotsize, size of the property, by multiplying lotwidth by lotlength, and use lotsize as a new predictor instead of lotwidth and lotlength.

Then we observe that 87 observations have missing values in the predictor maxsqfoot; therefore, we remove maxsqfoot because it contains too many missing values of observations which has less statistical power when modeling. Similarly, there are still 11 observations that have missing values in other predictors, so that these 11 observations are also removed.

II. Exploratory Data Analysis

Our cleaned data includes 139 observations and 8 variables:

sale: Continuous.

list: Continuous.

badroom: Discrete.

bathroom: Discrete.

parking: Discrete.

taxes: Continuous.

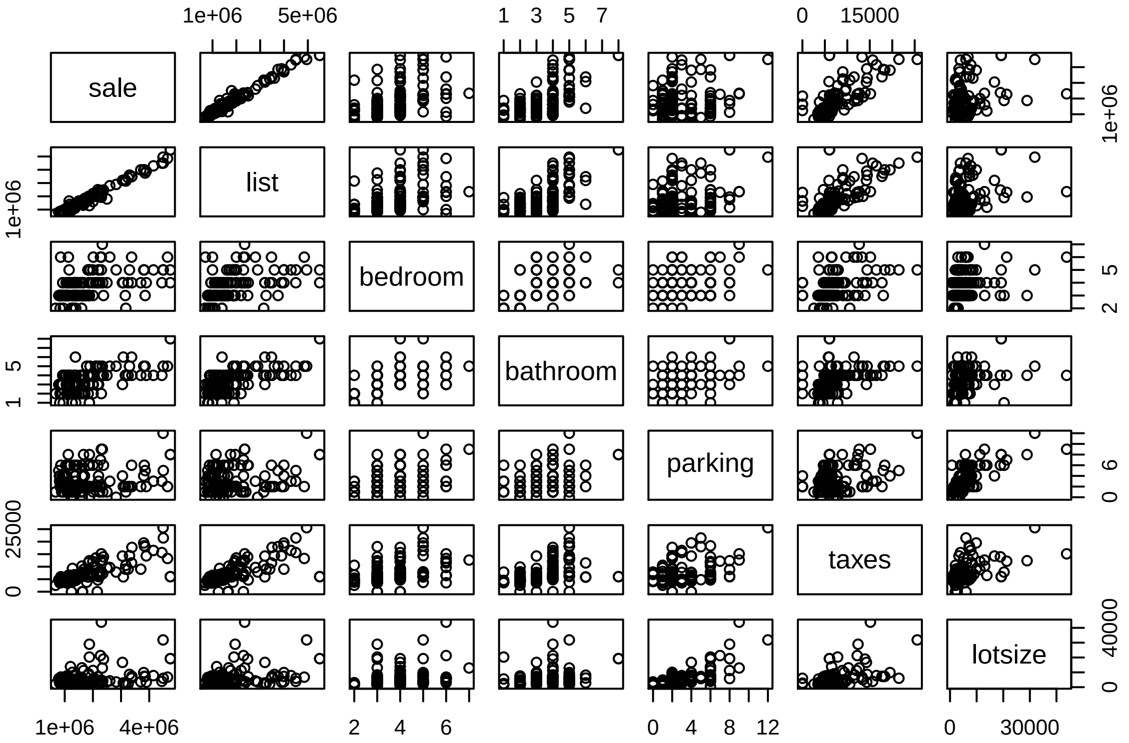
lotsize: Continuous.

location: Categorical.

Pairwise correlations:

```
##            sale   list bedroom bathroom parking  taxes lotsize
## sale     1.0000 0.9869  0.4508   0.6202  0.1796 0.7378  0.3353
## list     0.9869 1.0000  0.4497   0.6503  0.2380 0.7143  0.3723
## bedroom  0.4508 0.4497  1.0000   0.5450  0.3492 0.4182  0.3260
## bathroom 0.6202 0.6503  0.5450   1.0000  0.3868 0.4325  0.3301
## parking  0.1796 0.2380  0.3492   0.3868  1.0000 0.3612  0.7113
## taxes    0.7378 0.7143  0.4182   0.4325  0.3612 1.0000  0.5116
## lotsize  0.3353 0.3723  0.3260   0.3301  0.7113 0.5116  1.0000
```

Scatterplot matrix 0935

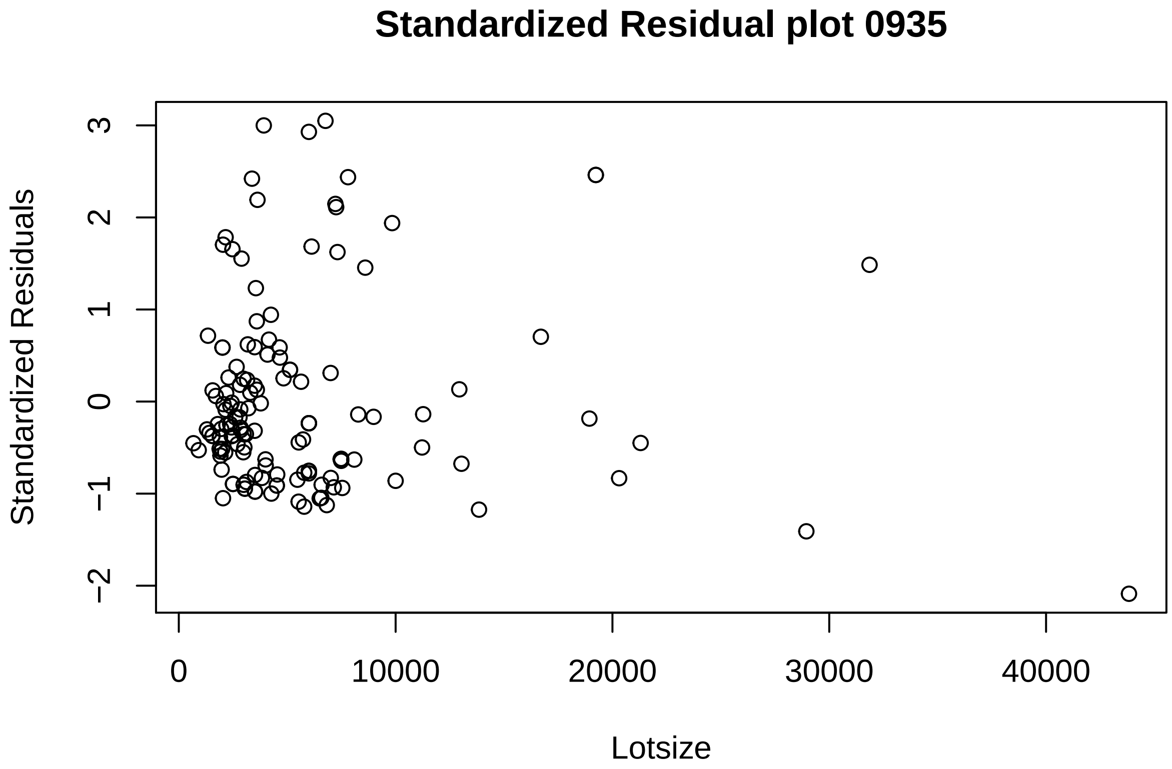


Rank of predictors in terms their correlation coefficient with sale price in descending order:

```
##     list    taxes bathroom  bedroom  lotsize  parking
##   0.9869   0.7378   0.6202   0.4508   0.3353   0.1796
```

From the scatterplot matrix, we see that the relationship between lotsize and sale price may not have constant variance because it seems to fan out and has an increasing variance.

To verify this we build an SLR model for lotsize and sale price, and plot its standardized residual plot:



We see that it indeed has an increasing variance, which violates the assumption of homoscedasticity.

III. Methods and Model

We firstly build an MLR model for sale price using all predictors (full model)

Table of results for the full model (All rounded to 4 decimal places):

predictors	coefficients	p-values
Intercept	80847.1754	0.1338
list	0.8351	<0.0001
bedroom	17373.5390	0.2148
bathroom	1747.6661	0.9017
parking	-19627.5221	0.0301
taxes	23.1965	<0.0001
lotsize	-1.7149	0.5624
locationT	80226.8010	0.0395

We have four significant predictors (i.e. p-value for t-test less than significance level of 5%):

list: When list price increase by 1 dollar, the expected sale price will increase by 0.8351 dollars when other predictors are unchanged.

parking: When parking increases by 1 unit, the expected sale price will decrease by 19627.5221 dollars when other predictors are unchanged.

taxes: When taxes increase by 1 dollar, the expected sale price will increase by 23.1965 dollars when other predictors are unchanged.

locationT: When other predictors are unchanged and the location of the detached house changed from Mississauga to Toronto, the expected sale price will increase by 80226.8010.

Backward elimination with AIC

By backward elimination method of stepwise regression with AIC, we have the fitted model:

$Sale\hat{P}rice$ = 85171.0307 + 0.8358$ListPrice$ + 18109.4496$BedRoom$ - 21936.3952$Parking$ + 22.4420$Taxes$ + 81678.6633$LocationToronto$

Where $Sale\hat{P}rice$ is the expected sale price of properties; $ListPrice$ is the list price of properties; $BedRoom$ is the number of bedrooms; $Parking$ is the number of parking spots; $Taxes$ is the previous year's property tax; $LocationToronto$ is Toronto Neighbourhood.

We find that in our AIC model, predictors are not the same with significant predictors in our full model, where $BedRoom$ is significant in the AIC model but not in the full model. This is acceptable because these two models are based on different test and criteria which may have different results.

Backward elimination with BIC

By backward elimination of stepwise regression with BIC, we have the fitted model:

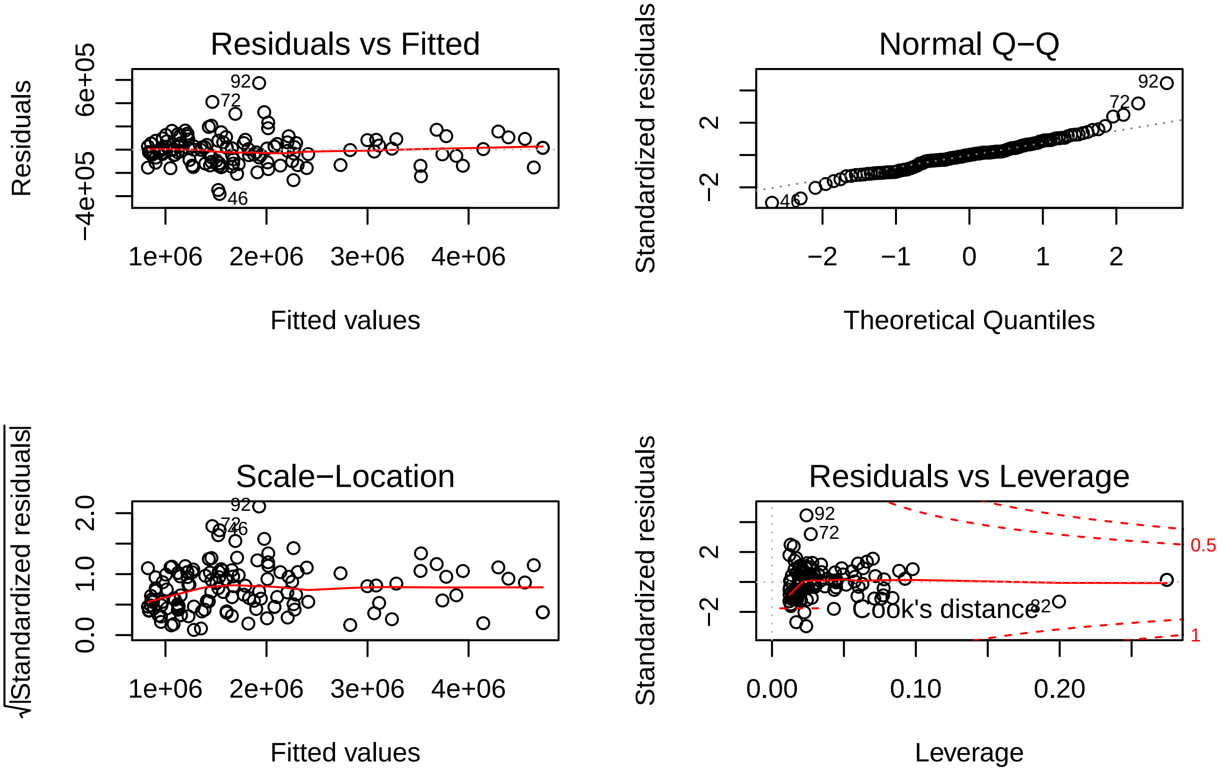
$Sale\hat{P}rice$ = 128443.5137 + 0.8403$ListPrice$ - 19110.5591$Parking$ + 22.9345$Taxes$ + 87157.1780$LocationToronto$

We find that our BIC model is not consistent with the AIC model because BIC has a larger penalty than AIC such that the predictor has to be more significant to be added to the BIC model. However, BIC predictors are the same as significant predictors in the full model. We can say that these four predictors are good predictors because we have same results by two different test and criteria, so we can consider this model to be our final model.

IV. Discussions and Limitations

When testing MLR assumptions, we assume errors have normal distribution with mean 0 and variance σ^2, which is in the true model and we cannot observe, so we use residuals to estimate errors, and make inference of our model by observing residuals. We assume the errors are uncorrelated because our sample items are drawn randomly and independently.

Diagnostic plots 0935



The residuals v.s. fitted plot shows the relationship between fitted values and residuals. We can see residuals have mean 0, and there is no pattern between fitted values and residuals, so our MLR model is appropriate. From the standardized residual plot, we see that most of the points are randomly distributed along the line, except for there is a tiny trend of increasing variance at the start, but there is a constant variance in general, so the assumption of homoscedasticity basically holds. From the normal Q-Q plot we see that there are some deviations of points at start and end but the most of points are evenly distributed along the line, so the model assumption of normality roughly holds but there is still room for improvement. The residuals v.s. leverage plot shows that there are some influential points that may influence the parameter estimates of our MLR model, such as number 92 and 72.

Next step

Although our MLR model assumptions basically hold, there still are some works that can be done to improve our MLR model. For example, we can apply some transformation on $Sale\hat{P}rice$ or on predictors to improve the model assumptions of linearity, homoscedasticity and normality. Besides, we can delete some influential points before fitting the model so that we could have more accurate parameter estimates and a better-performed model. Additionally, we can apply model validation. We can divide 70% of our data as training data, and 30% as testing data; then, we use training data to fit an MLR model and use testing data to test the performance of the fitted model. By this method, the probability of model overfitting can be reduced and make our model more predictive.