# A multiple linear model to predict the sale price of single-family, detached homes in the two neighbourhoods in the Greater Toronto Area

Linxia Li (1005715488)

December 5, 2020

## I. Data Wrangling

### Report the IDs of the sample selected

```
##   [1] 106 144 126  38  89  72  31  46 170  68 182 162  43 174 115  78 169  81
##  [19] 138  64 189  49  85 125 229 150 139  56 207 166 159 212  60 132 108  11
##  [37]  62  16 119 201 140   8 158 146 107  39 104  69 134  32  37 187 172  15
##  [55]  52  76  84 118  65 148  51  93  33  79 177 218 167 157  92 195 113  73
##  [73] 191  12 154  47  17  88 112  96  45 194  83  44 179  70  75 135 186  23
##  [91]  14  87  91  24 185 188 116 133 161  42 101  28  22 160   6  99 152 153
## [109]   3  40  13 110  55  36   7 137 183   4 227   2  53 204  61  82 145  34
## [127]   1 143 103 109   5 155 131 122  77  90  63  94  27 193 100  95  20  80
## [145] 163 117  48  19  71  26
```

### Create a new variable called "lotsize" by multiplying "lotwidth" by "lotlength" and use it to replace them

```
## Rows: 150
## Columns: 10
## $ ID        <int> 106, 144, 126, 38, 89, 72, 31, 46, 170, 68, 182, 162, 43,...
## $ sale      <int> 950000, 2270000, 1259000, 4500000, 1200000, 2350000, 3700...
## $ list      <int> 799000, 2300000, 999000, 4500000, 1149000, 2299000, 39950...
## $ bedroom   <int> 2, 6, 2, 5, 3, 4, 4, 5, 3, 3, 4, 3, 4, 3, 3, 4, 3, 1, 5, ...
## $ bathroom  <int> 1, 4, 1, 5, 2, 4, 4, 3, 1, 2, 3, 4, 4, 3, 4, 4, 4, 2, 5, ...
## $ parking   <int> 1, 7, 1, 5, NA, 1, 2, 2, 3, 1, 4, 2, 2, 4, 2, 2, 6, NA, 1...
## $ maxsqfoot <int> NA, 3500, 1500, 5000, 1500, 3000, NA, NA, NA, 1500, 2000,...
## $ taxes     <dbl> 3841.000, 12200.000, 3885.000, 21486.000, 4114.000, 7191....
## $ location  <chr> "T", "M", "T", "T", "T", "T", "T", "T", "M", "T", "M", "M...
## $ lotsize   <dbl> 1975.410, 21300.000, 1553.750, 6000.000, 2825.000, 4158.0...
```

The above data called data1 replace two variables "lotwidth" and "lotlength" by "lotsize" which is calculated by "lotwidth" times "lotlength".

### Clean the data by removing at most eleven cases and one predictor

```
## Rows: 141
## Columns: 9
## $ ID       <int> 106, 144, 126, 38, 72, 31, 46, 170, 68, 182, 162, 43, 174,...
## $ sale     <int> 950000, 2270000, 1259000, 4500000, 2350000, 3700000, 20000...
## $ list     <int> 799000, 2300000, 999000, 4500000, 2299000, 3995000, 159900...
## $ bedroom  <int> 2, 6, 2, 5, 4, 4, 5, 3, 3, 4, 3, 4, 3, 3, 4, 3, 5, 5, 4, 3...
## $ bathroom <int> 1, 4, 1, 5, 4, 4, 3, 1, 2, 3, 4, 4, 3, 4, 4, 4, 5, 4, 3, 3...
## $ parking  <int> 1, 7, 1, 5, 1, 2, 2, 3, 1, 4, 2, 2, 4, 2, 2, 6, 12, 2, 6, ...
## $ taxes    <dbl> 3841.000, 12200.000, 3885.000, 21486.000, 7191.000, 8995.0...
## $ location <chr> "T", "M", "T", "T", "T", "T", "T", "M", "T", "M", "M", "T"...
## $ lotsize  <dbl> 1975.410, 21300.000, 1553.750, 6000.000, 4158.000, 3630.00...
```

After creating a new variable "lotsize" and using it to replace "lotwidth" and "lotlength", we obtain a new data called data1 containing 150 observations and 10 variables. From data1, we can find out that the variable "maxsqfoot" contains a number of NA which means that there are some data unavailable. We have known that NA can have an influence on the multiple linear model I would like to find. Therefore, I would like to remove one of the predictors of sale price "maxsqfoot".

Besides, the unavailable data also appear in other variables other than "maxsqfoot" so that I should remove those cases. I used "na.omit" to remove those cases and obtain a new data called real containing 141 observations and 9 variables. I removed 9 cases from the original data and the data "real" would be used in the remaining parts.

## II. Exploratory Data Analysis

**Classify each variable included in this assignment as categorical or discrete or continuous**

Categorical variable: location

Discrete variables: ID, sale, list, bedroom, bathroom, parking, maxsqfoot

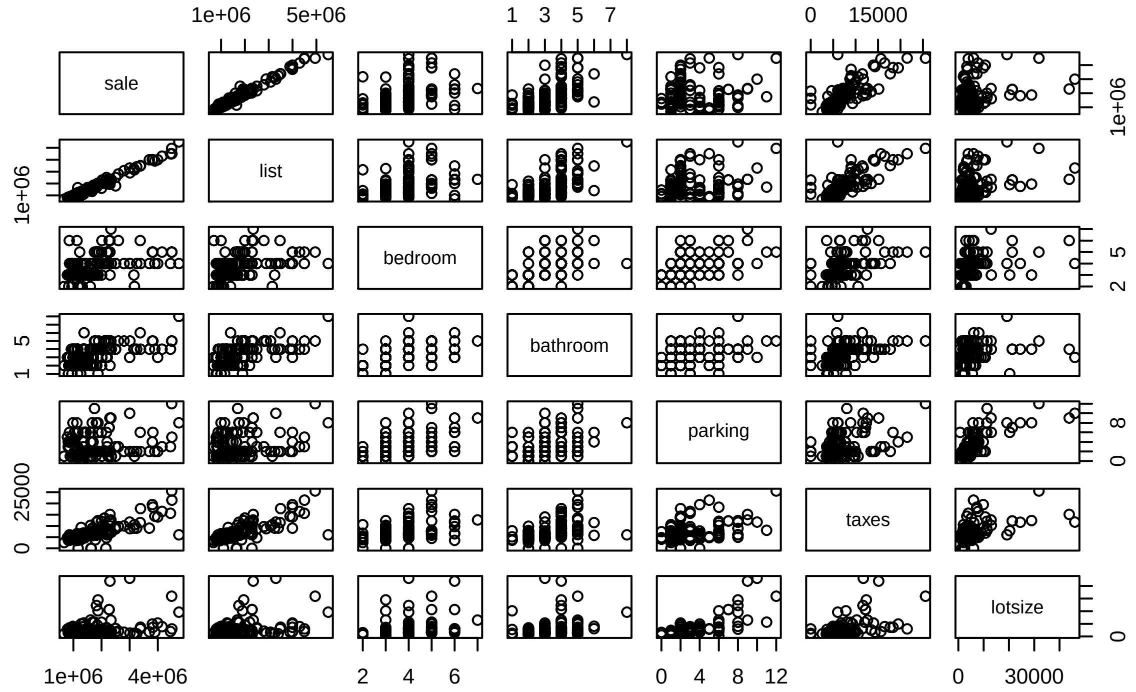
Continuous variables: taxes, lotwidth, lotlength, lotsize

**Pairwise correlations for all pairs of quantitative variables in the data (LL5488)**

```
##            sale   list bedroom bathroom parking  taxes lotsize
## sale     1.0000 0.9860  0.4134   0.5669  0.1569 0.7751  0.3104
## list     0.9860 1.0000  0.4129   0.5903  0.2141 0.7608  0.3457
## bedroom  0.4134 0.4129  1.0000   0.5327  0.4541 0.4624  0.3408
## bathroom 0.5669 0.5903  0.5327   1.0000  0.3616 0.4561  0.2490
## parking  0.1569 0.2141  0.4541   0.3616  1.0000 0.3730  0.7203
## taxes    0.7751 0.7608  0.4624   0.4561  0.3730 1.0000  0.5023
## lotsize  0.3104 0.3457  0.3408   0.2490  0.7203 0.5023  1.0000
```

**Scatterplot matrix for all pairs of quantitative variables in the data**

## :atterplot matrix for all pairs of quantitative variables in the data (LL548



**Ranking of the quantitative predictors for sale price in terms of correlation coefficients (LL5488)**

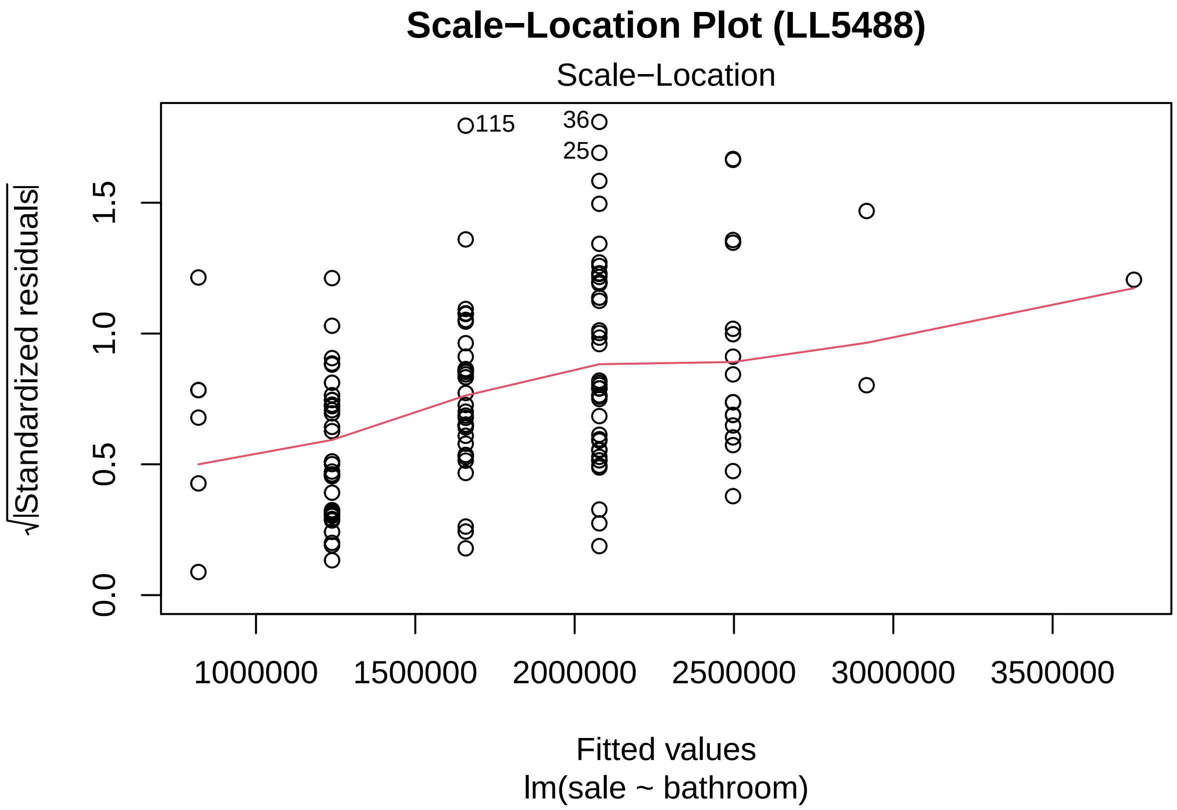
| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| list | taxes | bathroom | bedroom | lotsize | parking |
| 0.9860 | 0.7751 | 0.5669 | 0.4134 | 0.3104 | 0.1569 |

**Description:** The above table shows the ranking of the quantitative predictors for sale price in terms of correlation coefficients from highest to lowest. The correlation coefficients are from the above pairwise correlations for all pairs of quantitative variables in the data. From the above table, we can find out that the correlation coefficients of list price for sale price is 0.9860 which is the largest among those predictors so that there is a strong linear relationship between sale price and list price. The correlation coefficients of parking for sale price is 0.1569 which is the smallest among those predictors so that there is a weak linear relationship between sale price and parking.

**A single predictor of sale price causing the assumption of constant variance be strongly violated**

Based on the scatterplot matrix, a single predictor bathroom can cause the assumption of constant variance be strongly violated.

**A plot of the (standardized) residuals from SLR of sale price and bathroom**



From the above Scale-Location plot of the linear regression between sale price and bathroom, we can find out that the red line is not quite approximately horizontal and the spread around the red line varies with the fitted values which means that the residuals do not appear randomly spread. Therefore, this plot confirms that the predictor bathroom can cause the assumption of constant variance be strongly violated.

## III. Methods and Model

**Fit an additive linear regression model with all available predictors variables for sale price**

```
##
## Call:
## lm(formula = sale ~ list + bedroom + bathroom + parking + taxes +
##     location + lotsize, data = real)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -394393  -77490    2274   64803  564846
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.392e+04  5.392e+04   1.185   0.2380
## list         8.317e-01  2.278e-02  36.508  < 2e-16 ***
## bedroom      2.141e+04  1.440e+04   1.487   0.1394
## bathroom     4.152e+03  1.346e+04   0.308   0.7582
## parking     -1.898e+04  8.918e+03  -2.128   0.0352 *
## taxes        2.191e+01  4.650e+00   4.711 6.11e-06 ***
## locationT    8.301e+04  3.954e+04   2.099   0.0377 *
## lotsize     -7.245e-02  2.467e+00  -0.029   0.9766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 127000 on 133 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9792
## F-statistic: 943.5 on 7 and 133 DF,  p-value: < 2.2e-16
```

**List the estimated regression coefficients and the p-values for the corresponding t-tests for these coefficients (LL5488)**

| Term | Estimate | p-value |
|---|---|---|
| (intercept) | 63920 | 0.2380 |
| list | 0.8317 | < 0.0001 |
| bedroom | 21410 | 0.1394 |
| bathroom | 4152 | 0.7582 |
| parking | -18980 | 0.0352 |
| taxes | 21.91 | < 0.0001 |
| locationT | 83010 | 0.0377 |
| lotsize | -0.07245 | 0.9766 |

**Interpret the estimated model coefficient if the t-test was significant**

**List:** Holding all other explanatory variables in the model fixed, for every 1 dollar increase in the list price, on average sale price increases by 0.8317 dollars.

**Parking:** Holding all other explanatory variables in the model fixed, for every 1 dollar increase in parking, on average sale price decreases by 18980 dollars.

**Taxes:** Holding all other explanatory variables in the model fixed, for every 1 dollar increase in taxes, on average sale price increases by 21.91 dollars.

**Location:** Holding all other explanatory variables in the model fixed, the sale price for the location Toronto is 83010 larger than that for the location Mississauga. In other words, holding all other variables in the model fixed, when the location is Toronto, on average sale price increases by 83010 dollars. When the location is Mississauga, on average sale price will not increase.

**Start with the full model fitted above and use backward elimination with AIC**

```
## Start:  AIC=3321.84
## sale ~ list + bedroom + bathroom + parking + taxes + location +
##     lotsize
##
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 1.3915e+07 2.1456e+12 3319.8
## - bathroom  1 1.5347e+09 2.1471e+12 3319.9
## <none>                   2.1456e+12 3321.8
## - bedroom   1 3.5674e+10 2.1813e+12 3322.2
## - location  1 7.1102e+10 2.2167e+12 3324.4
## - parking   1 7.3075e+10 2.2187e+12 3324.6
## - taxes     1 3.5809e+11 2.5037e+12 3341.6
## - list      1 2.1502e+13 2.3648e+13 3658.2
##
## Step:  AIC=3319.84
## sale ~ list + bedroom + bathroom + parking + taxes + location
##
```

```
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 1.6614e+09 2.1473e+12 3318.0
## <none>                   2.1456e+12 3319.8
## - bedroom   1 3.5774e+10 2.1814e+12 3320.2
## - location  1 7.2083e+10 2.2177e+12 3322.5
## - parking   1 9.4501e+10 2.2401e+12 3323.9
## - taxes     1 3.7403e+11 2.5196e+12 3340.5
## - list      1 2.1978e+13 2.4123e+13 3659.0
##
## Step:  AIC=3317.95
## sale ~ list + bedroom + parking + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   2.1473e+12 3318.0
## - bedroom   1 4.8905e+10 2.1962e+12 3319.1
## - location  1 7.3058e+10 2.2203e+12 3320.7
## - parking   1 9.9352e+10 2.2466e+12 3322.3
## - taxes     1 3.7828e+11 2.5255e+12 3338.8
## - list      1 3.0658e+13 3.2806e+13 3700.4
```

```
##
## Call:
## lm(formula = sale ~ list + bedroom + parking + taxes + location,
##     data = real)
##
## Coefficients:
## (Intercept)         list      bedroom      parking        taxes    locationT
##   7.033e+04    8.354e-01    2.318e+04   -1.943e+04    2.160e+01    7.893e+04
```

The fitted model with AIC:

$\hat{sale} = 70330 + 0.8354list + 23180bedroom - 19430parking + 21.60taxes + 78930location$

The results are not consistent with those in the fullmodel we fitted above. List price, the number of bedrooms, the total number of parking spots, taxes and location are repeated from the fullmodel as relevant predictors for sale price. However, the number of bathrooms and lotsize which were highlighted by the fullmodel were not highlighted by the model selected by backward elimination using AIC here.

### Use BIC instead of AIC

```
## Start:  AIC=3345.43
## sale ~ list + bedroom + bathroom + parking + taxes + location +
##     lotsize
##
##            Df  Sum of Sq        RSS    AIC
## - lotsize   1 1.3915e+07 2.1456e+12 3340.5
## - bathroom  1 1.5347e+09 2.1471e+12 3340.6
## - bedroom   1 3.5674e+10 2.1813e+12 3342.8
## - location  1 7.1102e+10 2.2167e+12 3345.1
## - parking   1 7.3075e+10 2.2187e+12 3345.2
## <none>                   2.1456e+12 3345.4
## - taxes     1 3.5809e+11 2.5037e+12 3362.2
## - list      1 2.1502e+13 2.3648e+13 3678.9
##
## Step:  AIC=3340.48
## sale ~ list + bedroom + bathroom + parking + taxes + location
##
```

```
##            Df  Sum of Sq        RSS    AIC
## - bathroom  1 1.6614e+09 2.1473e+12 3335.6
## - bedroom   1 3.5774e+10 2.1814e+12 3337.9
## - location  1 7.2083e+10 2.2177e+12 3340.2
## <none>                   2.1456e+12 3340.5
## - parking   1 9.4501e+10 2.2401e+12 3341.6
## - taxes     1 3.7403e+11 2.5196e+12 3358.2
## - list      1 2.1978e+13 2.4123e+13 3676.7
##
## Step:  AIC=3335.64
## sale ~ list + bedroom + parking + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## - bedroom   1 4.8905e+10 2.1962e+12 3333.9
## - location  1 7.3058e+10 2.2203e+12 3335.4
## <none>                   2.1473e+12 3335.6
## - parking   1 9.9352e+10 2.2466e+12 3337.1
## - taxes     1 3.7828e+11 2.5255e+12 3353.6
## - list      1 3.0658e+13 3.2806e+13 3715.1
##
## Step:  AIC=3333.87
## sale ~ list + parking + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## - parking   1 6.5862e+10 2.2620e+12 3333.1
## <none>                   2.1962e+12 3333.9
## - location  1 9.0719e+10 2.2869e+12 3334.6
## - taxes     1 4.2807e+11 2.6242e+12 3354.0
## - list      1 3.1157e+13 3.3353e+13 3712.5
##
## Step:  AIC=3333.09
## sale ~ list + taxes + location
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   2.2620e+12 3333.1
## - taxes     1 3.8997e+11 2.6520e+12 3350.6
## - location  1 5.9327e+11 2.8553e+12 3361.0
## - list      1 3.2344e+13 3.4606e+13 3712.8
```

```
##
## Call:
## lm(formula = sale ~ list + taxes + location, data = real)
##
## Coefficients:
## (Intercept)         list        taxes    locationT
##   6.350e+04    8.292e-01    2.149e+01    1.440e+05
```

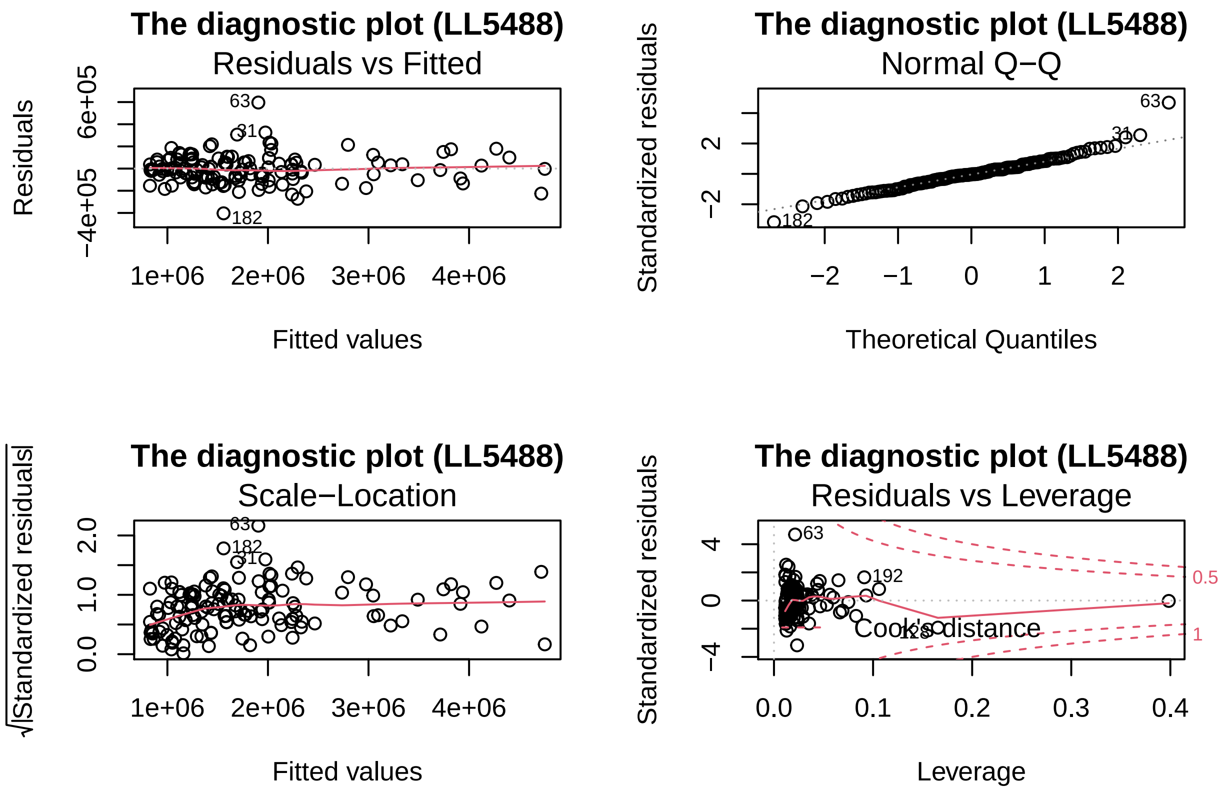
The fitted model with BIC:

$\hat{sale} = 63500 + 0.8292list + 21.49taxes + 144000location$

The results are not consistent with those in the fullmodel and the model fitted with backward elimination using AIC we fitted above. List price, taxes and location are repeated from the fullmodel and the model fitted with backward elimination using AIC as relevant predictors for sale price. However, the number of bedrooms and the total number of parking spots which were highlighted by the fullmodel and the model fitted with backward elimination using AIC were not highlighted by backward elimination using BIC here.

# IV. Discussions and Limitations

### The four diagnostic plots



### Interpret each of the four residual plots shown above

### Residuals vs Fitted plot

From the residual plot we made above, we can find out that the residuals are equally spread around a horizontal line which indicates that the model provides an adequate summary of the data. Besides, those residuals are correlated non-linearly with the fitted values which suggests that there is a linear trend between sale price and those three predictors: list, taxes and location.

### Normal Q-Q plot

The normal QQ plot shows if residuals are normally distributed. We have known that if the residuals are lined well on the straight dashed line, the model we can fit is pretty good. From the normal QQ plot we made above, we can find out that most of the standardized residuals are lined well on the straight dashed line which means that the model we fitted is good and the residuals are normally distributed. Besides, we can also find out that the distribution of the response is symmetric.

### Scale-Location plot

The scale-location plot shows if residuals are spread equally along the ranges of predictors and can check the assumption of constant variance (homoscedasticity). From the Scale-Location plot we made above, we can find out that the red line is approximately horizontal. Besides, the spread around the red line does not vary with the fitted values which means that the residuals appear randomly spread. Therefore, it is good to see that a horizontal line with equally (randomly) spread points so that the assumption of constant variance holds.

**Residuals vs Leverage plot**

The residuals vs leverage plot can help us find the influential cases. We have known that not all outliers are influential in the linear regression analysis. Even though they have extreme values , they might not be important to determine a regression line which means that the results (model we would like to fit) would not be much different if we include or exclude them from analysis. However, some outliers can have an influence on the results. If there are some outliers at the upper right corner or at the lower right corner (outside of the dashed line, Cook's distance), the results can be influenced. However, in the residuals vs leverage plot we made, we did not find any points which are at the upper right corner or at the lower right corner (outside of the dashed line, Cook's distance) so that there are no outliers and the model we fit is good.

**Discussion about whether the normal error MLR assumptions are satisfied**

From the Residuals vs Fitted plot, the residuals are equally spread around a horizontal line without distinct patterns so that the errors are uncorrelated and there is a linear relationship between the response variable sale price and those predictors (independent variables).

From the Normal Q-Q plot, most of the standardized residuals are lined well on the straight dashed line which means that the errors are normally distributed.

From the Scale-Location Plot, the red line is approximately horizontal and the residuals appear randomly spread which means that the errors have constant variance.

Overall, the normal error MLR assumptions are satisfied.

**Next steps to find a valid final model**

In the Methods and Models section, we have used backward elimination using AIC and BIC to find a multiple linear model with the sale price as response variable and list price, taxes and location as three predictors. Besides, we also produce four diagnostic plots to show whether the model we found is good or not. However, we have not assess the predictive ability of the models we found above. Therefore, we would like to use a method called cross validation and take some steps to evaluate the predictive ability of them to find a valid final model.

We have known that k-fold Cross-validation is a standard approach to assess the predictive ability of models by evaluating their performance on a new data set.

The steps of the Cross Validation:

First, we should randomly divide the data into roughly k equal sets.

Second, we should establish the model by using all but one of the k folds and this set is called the training data set.

Third, we should use the remaining data set (the fold that was left out) called test data to evaluate the model.

Then, we should repeat the second step and the third step k times by changing the kth fold.

Eventually, we should calculate cross validation error by finding the average of the squared differences between the response and fitted values for the test set. A good candidate (model) will have small cross validation error. Therefore, the model with the smallest cross validation error would be the final model.