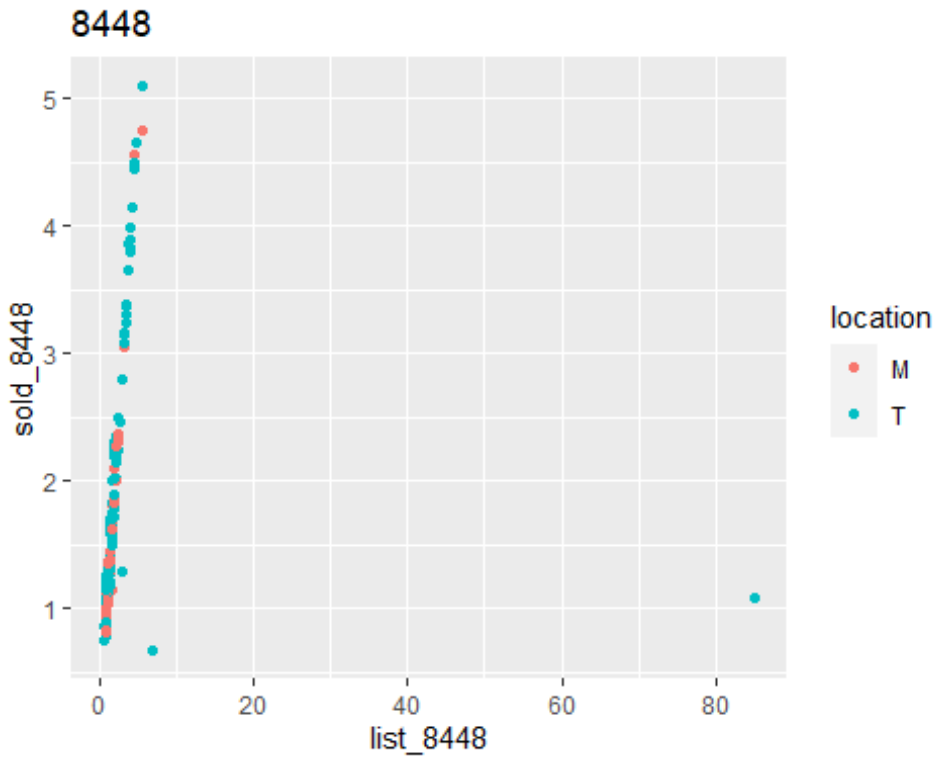
SLR in prediciting House Prices in Toronto and Missiauga

TONGFEI ZHOU

October 22, 2020

I. Exploratory Data Analysis



I will choose the listed price as the predictor in this scatterplot since we often want to know what the actual price would be after we know the listed price (I distinguished the locations as well here).

From the scatter plot we can see clearly that there is an influential point lied in the right most graph shown above. This graph is quite similar to the quiz4 5th question's graph and so we remove the influential point at the right most of the graph.

One can notice that there is a point with sold value quite small given the second largest list price, this is an outlier on the y-value, but it will not affect the line drastically. In fact, we can show that the linear models with or without this point is quite same. Data is shown as below.

```
## [1] 84.990  6.799
```

```
##
## Call:
## lm(formula = sold ~ list, data = data_1_TZ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9925 -0.1054 -0.0150  0.1183  0.6967
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.33775    0.05762   5.862  1.9e-08 ***
## list         0.78346    0.02766  28.328  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4137 on 197 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.8019
## F-statistic: 802.5 on 1 and 197 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = sold ~ list, data = data_2_TZ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57923 -0.06721 -0.02022  0.05908  0.40901
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13928    0.02412   5.776 2.97e-08 ***
## list         0.90998    0.01187  76.671  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1672 on 196 degrees of freedom
## Multiple R-squared:  0.9677, Adjusted R-squared:  0.9676
## F-statistic:  5878 on 1 and 196 DF,  p-value: < 2.2e-16

##            1            2            3            4            5
6
## 2.116599e-04 5.298814e-03 1.155868e-07 2.741912e-04 1.900660e-04
1.093297e+01
##            7            8            9           10           11
12
## 2.011184e-04 6.428785e-05 9.776454e-06 1.844324e-03 5.308063e-06
8.305911e-05
##           13           14           15           16           17
18
## 8.108002e-06 2.386576e-04 5.799690e-04 2.814835e-05 2.326665e-05
4.260922e-04
```

```
##           19           20           21           22           23
24
## 1.062684e-04 2.199955e-05 1.149834e-04 1.115079e-04 1.408660e-03
3.089962e-03
##           25           26           27           28           29
30
## 4.271242e-04 9.939386e-05 6.709471e-06 1.155868e-07 4.641761e-02
5.670039e-05
##           31           32           33           34           35
36
## 6.226911e-04 2.469093e-04 2.386576e-04 3.947998e-09 4.471606e-05
2.788259e-04
##           37           38           39           40           41
42
## 8.647991e-04 7.367216e-04 1.319560e-04 7.579350e-04 3.346473e-06
4.864579e-08
##           43           44           45           46           47
48
## 8.305911e-05 6.584605e-05 1.092630e-04 3.951752e-02 1.367435e-06
3.693277e-04
##           49           50           51           52           53
54
## 4.093156e-05 3.394120e-03 6.428170e-03 1.343950e-03 2.833867e-04
3.676586e-04
##           55           56           57           58           59
60
## 2.265942e-04 4.372880e-04 1.862914e-04 1.641686e-05 2.600588e-04
2.063112e-02
##           61           62           63           64           65
66
## 1.062684e-04 1.945927e-02 4.867276e-04 1.612836e-04 2.859958e-05
2.386576e-04
##           67           68           69           70           71
72
## 1.778043e-03 2.289518e-02 1.269615e-04 4.672905e-05 6.449246e-06
2.571462e-03
##           73           74           75           76           77
78
## 3.236492e-04 5.063460e-04 8.033892e-05 6.298950e-05 4.859077e-04
1.507138e-07
##           79           80           81           82           83
84
## 6.803393e-04 2.371394e-03 4.852522e-04 3.104384e-03 1.155160e-04
6.573772e-06
##           85           86           87           88           89
90
## 1.596865e-04 2.989832e-04 3.104384e-03 1.403605e-04 1.862914e-04
4.471546e-04
##           91           92           93           94           95
96
```

```
## 4.260922e-04 1.004439e-04 4.315468e-05 1.684197e-04 7.442818e-04
8.502381e-05
##           97           98           99          100          101
102
## 3.979321e-04 1.684197e-04 1.622778e-04 1.859817e-05 1.565436e-05
3.676586e-04
##          103          104          105          106          107
108
## 2.703340e-04 4.143739e-02 4.810633e-04 3.099847e-04 4.617798e-04
5.786013e-02
##          109          110          111          112          113
114
## 8.919379e-04 4.617798e-04 4.710109e-04 1.367435e-06 2.939522e-07
7.367216e-04
##          115          116          117          118          119
120
## 1.036812e-04 1.131618e-03 8.129673e-04 6.131456e-04 2.029237e-04
5.113296e-03
##          121          122          123          124          125
126
## 7.312284e-05 6.428170e-03 7.572662e-04 9.715462e-05 6.742934e-02
1.351286e-04
##          127          128          129          130          131
132
## 6.898362e-04 3.676586e-04 2.647683e-07 2.407701e-03 5.415278e-05
3.122835e-05
##          133          134          135          136          137
138
## 5.064734e-04 1.632658e-04 4.486208e-04 6.269082e-05 1.827915e-04
5.613391e-05
##          139          140          141          142          143
144
## 6.085956e-05 3.346473e-06 6.257912e-04 3.604248e-04 7.442818e-04
1.342829e-04
##          145          146          147          148          149
150
## 2.469093e-04 9.286838e-05 4.590271e-05 5.968953e-04 3.122835e-05
4.260922e-04
##          151          152          153          154          155
156
## 4.271242e-04 1.019159e-05 3.221978e-04 8.551009e-04 7.630694e-04
2.202534e-05
##          157          158          159          160          161
162
## 2.520639e-05 5.279710e-03 3.099847e-04 1.230453e-06 2.270183e-02
4.815813e-04
##          163          164          165          166          167
168
## 1.952439e-03 1.900660e-04 7.744633e-05 2.798454e-05 3.920068e-04
2.982360e-06
```

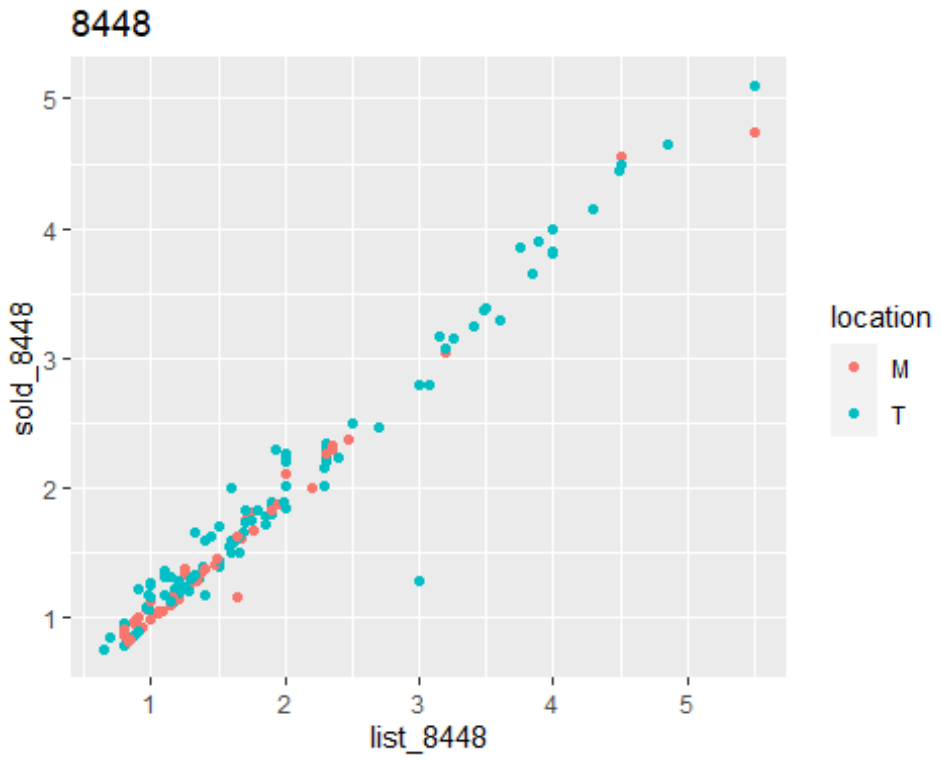
```
##          169          170          171          172          173
174
## 6.609159e-04 4.896257e-05 5.532280e-04 7.442818e-04 2.464277e-04
2.407701e-03
##          175          176          177          178          179
180
## 4.315468e-05 3.105927e-03 3.221978e-04 1.260631e-03 4.271242e-04
4.832301e-02
##          181          182          183          184          185
186
## 4.206527e-05 1.858133e-05 2.132597e-06 2.814835e-05 1.367435e-06
9.534264e-08
##          187          188          189          190          191
192
## 2.571100e-04 9.368681e-03 1.051258e-02 6.709471e-06 4.710109e-04
3.221978e-04
##          193          194          195          196          197
198
## 3.712326e-06 2.655825e-05 1.507138e-07 5.650988e-04 8.838387e-09
3.099847e-04
##          199
## 7.380056e-04
```

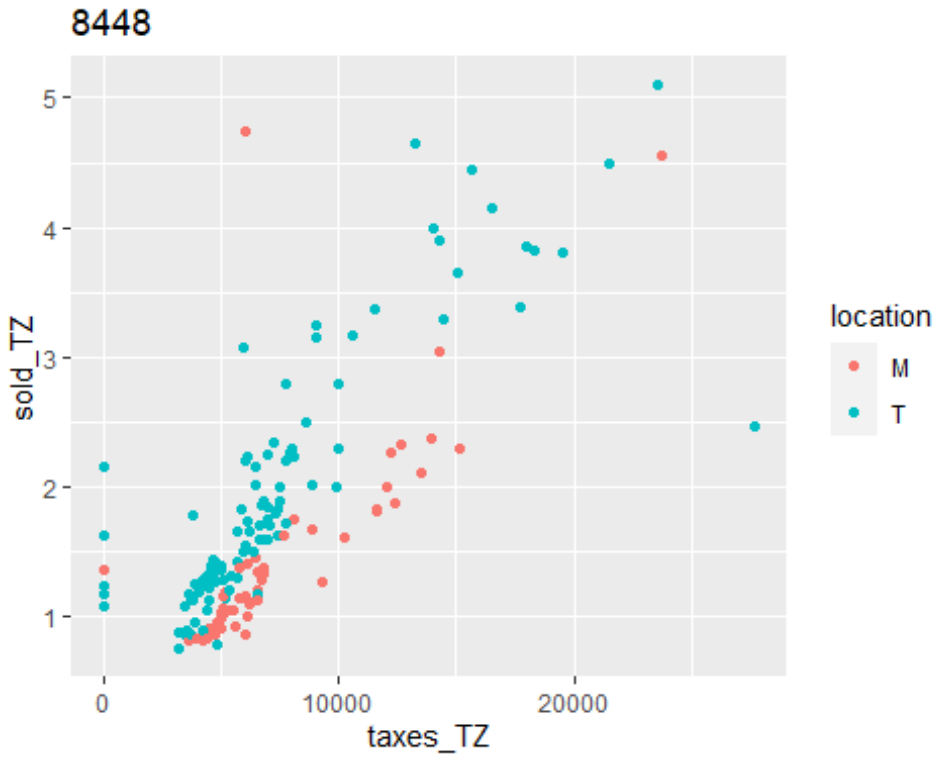
As one can see that, I used sort function first to get the influnetial point's x value and the possible outlier's x value, and then I run the linear model with or without the possible outiler but without the obvious influential point, and one can check that the linear regression line is pretty much the same.

However, after I calculate the Cook's distance without the obvious influential point ,there is a gap of the Cook's distance at the 6th point, which is this possible outlier, and the value of the Cook's distance has exceeded the threshold $\frac{4}{n-2} = 0.02$(Here n=199 since we have already removed the obvious influential point). Therefore, we should remove this point as well.

Therefore, I will remove the obvious influential point and the outlier point as well. So I will use the subset data named data_2_TZ for the remaining analysis.

Now we draw the rest two required scatterplots:





Interpretation of the three graphs: The first one shows that there is a strong linear relationship between the listed price and the acutal sold price in either Toronto or Missisauga area, with some influential and outlier points.

The second one shows that on average, given the same listed price, the actual sold price in Toronto is higher than it is in Missisauga, which reflects that the house price is more expensive in Toronto rather than in Missisauga. Here it shows once again that the linear relationship between the actual sold price and the listed price are quite strong.

The third one, however, shows that with some outliers, the taxes are roughly similar among two different regions. Besides, there is a moderate linear relationship between taxes and the actual sale price here.

II. Methods and Model

The following table shows the corresponded values in each linear regression model.

Summary Table among three data sets (continued below)

	R_squared_8448	Intercept_8448	Slope_8448	MSE_8448
lmod_alldata_TZ	0.9677	0.1393	0.91	0.02797
lmod_T_TZ	0.9577	0.1677	0.9081	0.04088
lmod_M_TZ	0.9829	0.1348	0.8923	0.01042

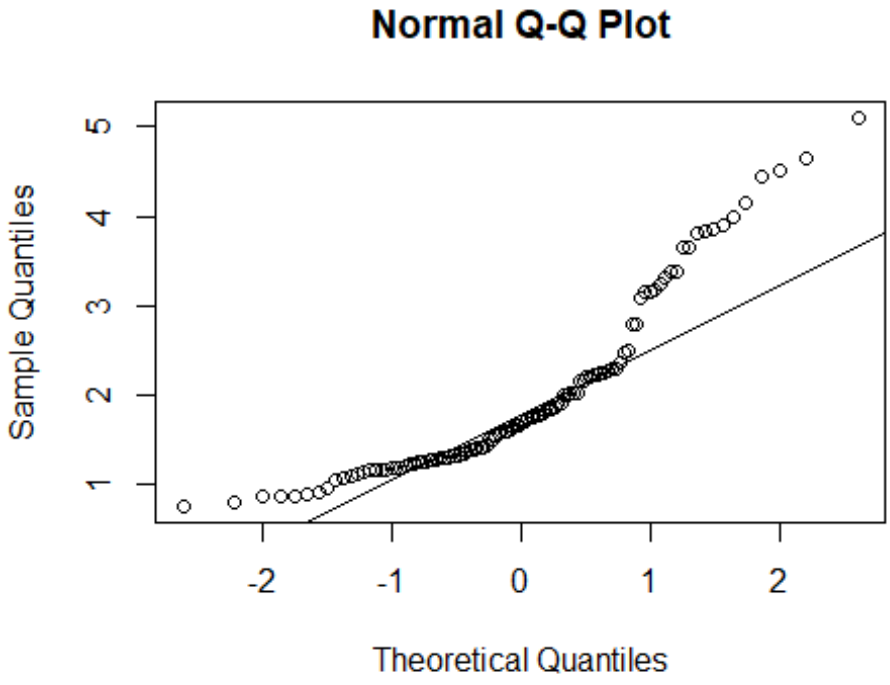
	P_value_8448	CI_left_8448	CI_right_8448
lmod_alldata_TZ	4.171e-148	0.8866	0.9334
lmod_T_TZ	4.952e-76	0.8717	0.9444
lmod_M_TZ	9.905e-78	0.867	0.9175

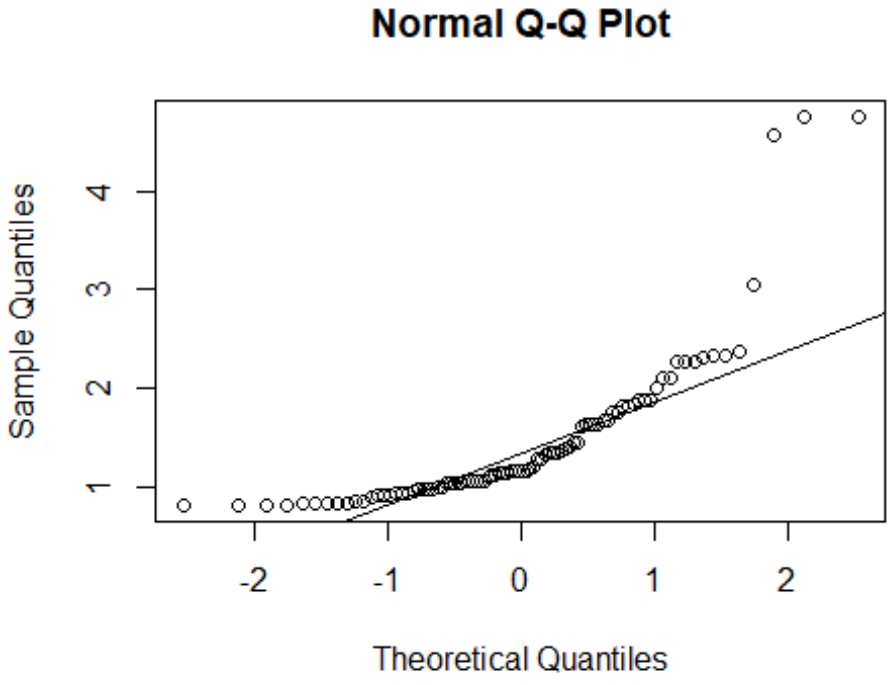
Now we interpret the R^2 in three models: Generally, R^2 stands for the proportion of the variance of y_i explained by the linear regression model. For the first model, the R^2 stands for the proportion of the variance of y_i explained by the model based on the whole data. For the second model, the R^2 stands for the proportion of the variance of y_i explained by the model based on the data where location is at Toronto. For the third model, the R^2 stands for the proportion of the variance of y_i explained by the model based on the data where location is at Missisauga.

Notice that all three R^2 results are quite similar, with highest in Missisauga dataset and lowest in Toronto dataset, and middle one overall. All these three results are quite high since from the scatterplot we plot above with sold~listed, one can see that the linear relationship is very strong, almost perfectly linear on either overall data, only Toronto, or only Missisauga. Therefore, most of the variance of y has been explained by the model successfully, and so leads to quite high R^2. This also explains why their results are similar since either datasets show a strongly positive linear relationship and either datasets' lmod has postive slope.

To conduct an appropriate pooled two-sample t-test, we need to be sure that the following assumptions are held well: 1. The two samples are independent 2. The two populations have the same variance. Here we do not have the populations so we check whether the sample variance are same or not and whether they have. Further, to conduct the test we need to check whether the two samples are normal or not.

```
## [1] 0.9583656 0.6008927
```





As one can check that the Normal Q-Q plot and conclude that the normality assumption is highly violated and the difference of variance is quite large (>0.3). Furthermore, since

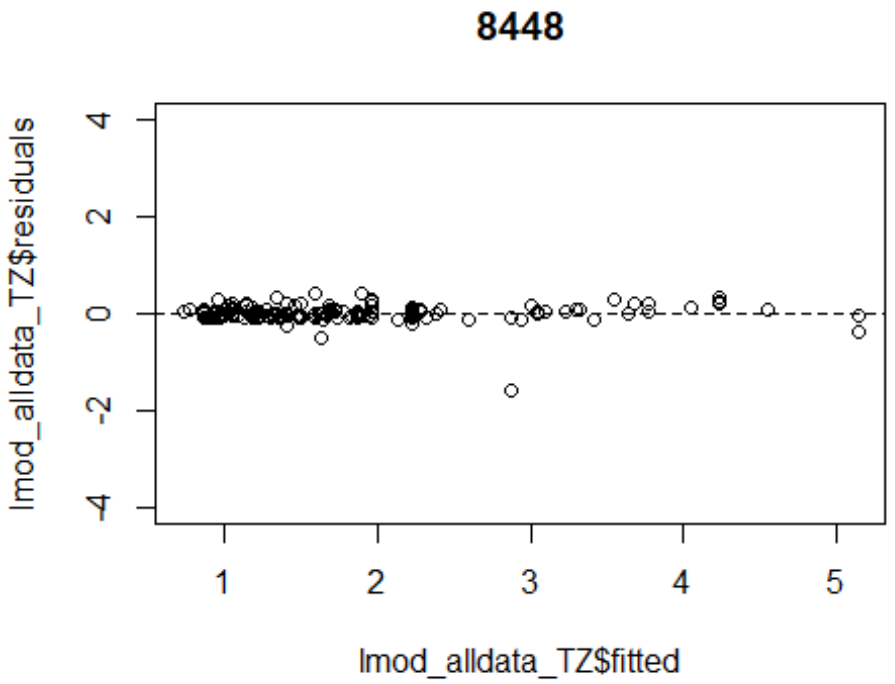
Toronto and Missisauga are quite near, their sample may not be independent since there should be some potential correlation, for example, as the house price for Toronto goes up, so should be the Missisauga since they stand for GTA. Therefore, the conditions to conduct a pooled two-sample t-test is not appropriate to be used here.

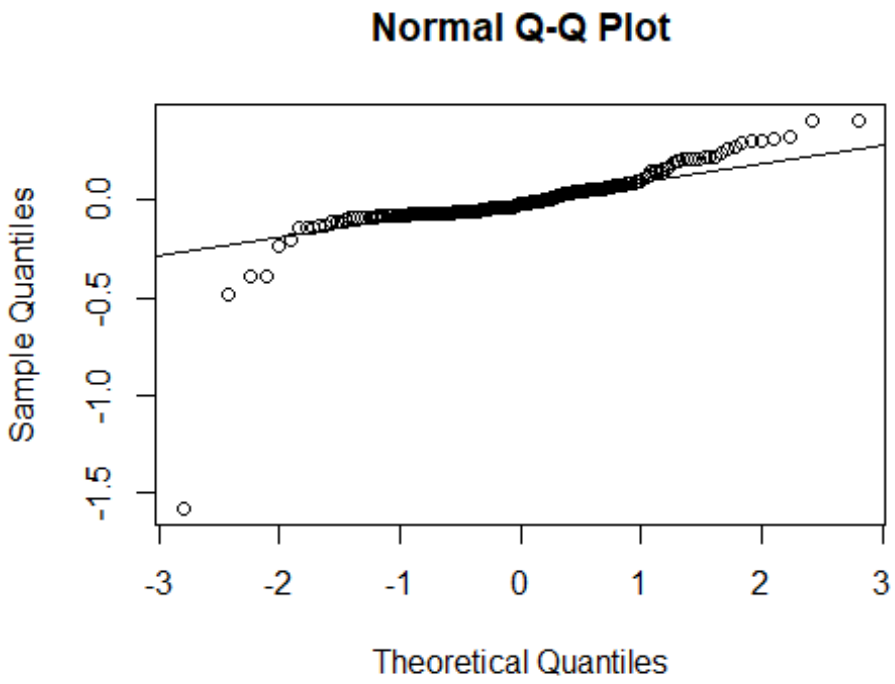
III. Discussions and Limitations

To select the best model, one can see above that all three models have quite high R^2 values, but one can notice that the sample size is smaller in Toronto and Missisauga datasets and so maybe the two corresponded models are not trained very well, and so with higher variance and wider predicting interval probably, which indicates that the two models may not perform well in predicting. What is more, there might have bias in the subsets as well since they only indicate certain area. Last, the overall dataset after removing the influential points and outliers, it shows very strongly positive linear relationship and so it is a overall great model.

Therefore, I will choose the overall dataset model among the three models.

We analyze the normal error SLR assumptions for the overall dataset model as below:





One can see above that the residuals vs. fitted value plot shows no pattern at all, the dots did not fanning out to the right or from the left, and vary randomly above or below 0, which means that the assumptions of linearity, independence and constant variance all hold very well. However,

by checking the Normal Q-Q plot we see that the normality assumption is not hold very well since we have tails dragged quite far away from the qqline.

Finally, for more predictors to add to fit a multiple linear regression model, we can choose GDP and income as the predictors.