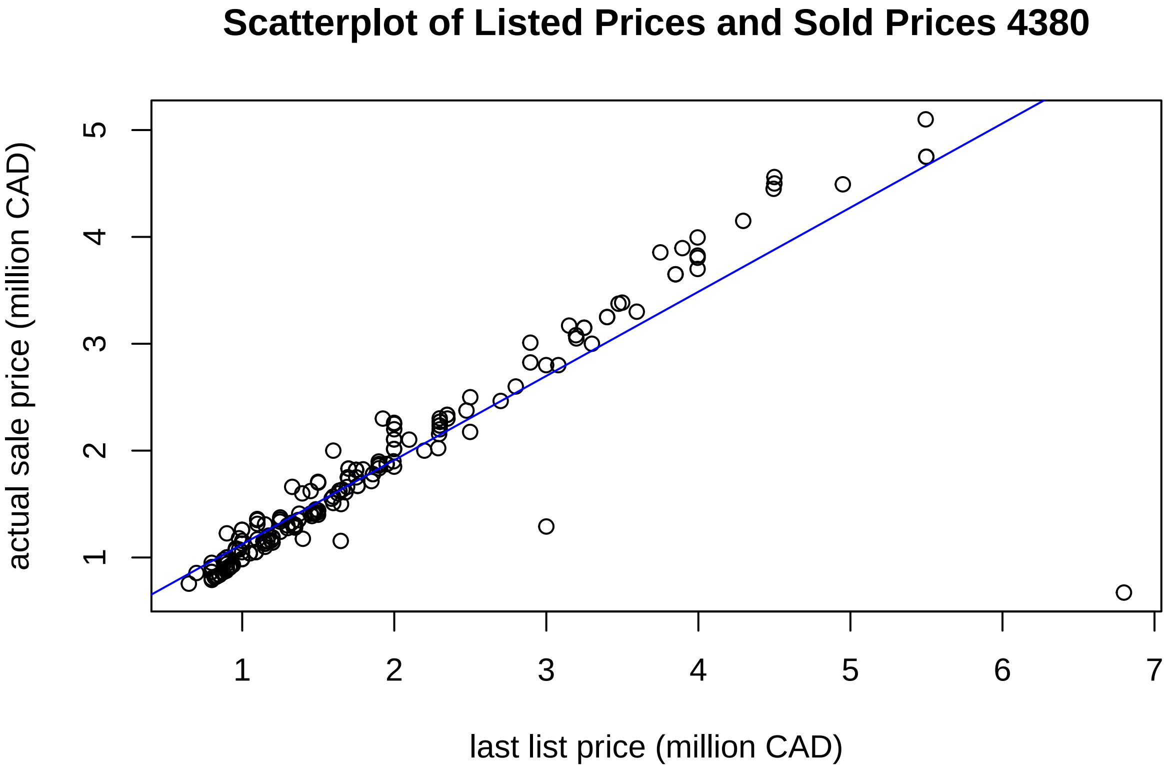
Data Analysis for the First-Time Home Buying Data

Y. C.

October 24, 2020

I. Exploratory Data Analysis

Scatterplot

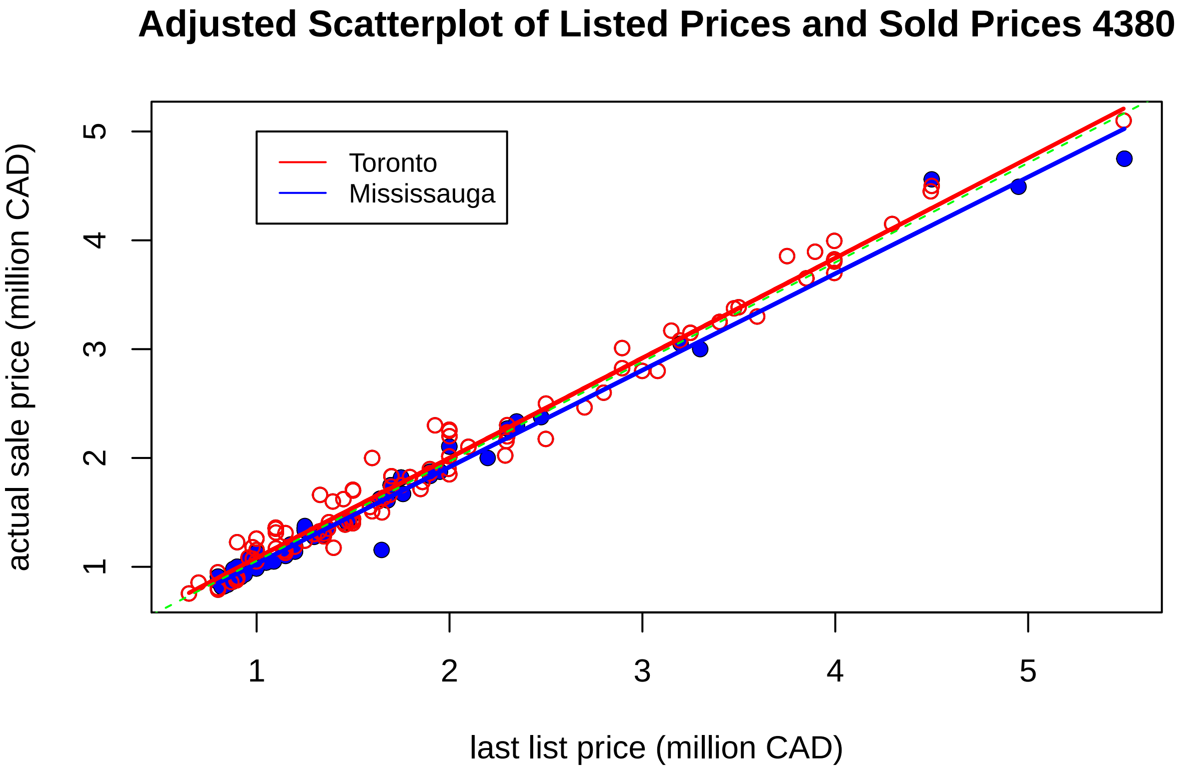


As you can see in the scatterplot, there are two outliers that influence the regression line. Without the two points, the slope estimate (β_1) of the regression will be a bit larger. After inspection, the two points have ID 59 (3.0000, 1.290) and 95 (6.7990, 0.672). Also, looking at the cook's distance for each point and sort it from biggest to smallest, the two points also stands out (have the biggest cook's distance, with ID 59: 6.345e-02, ID 95: 1.008e+01). According to both the scatterplot and the cook's distance, by removing the two outliers, we will be able to get a better regression line that fits the majority of the data.

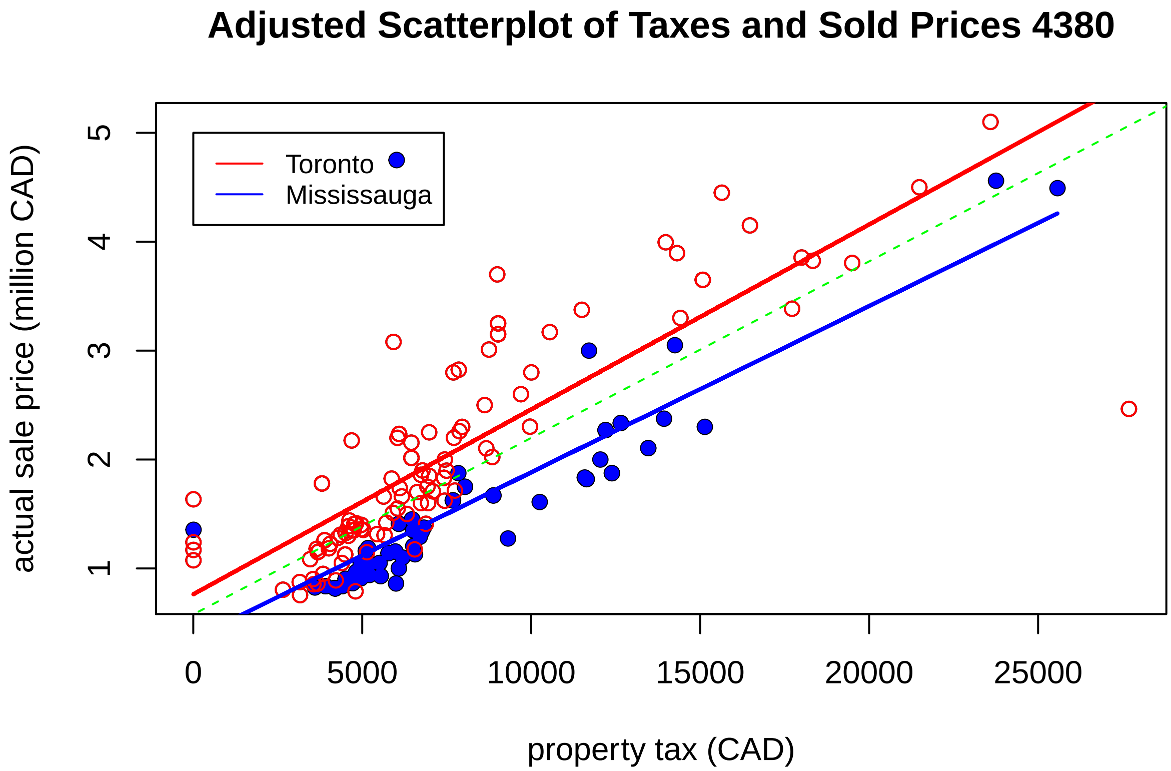
This is the scatterplot of the sample data with 200 cases drawn from the original data. It shows the relationship between listed price and sold price, with listed price being the independent variable and the sold price being the dependent variable. As the plot shows, the regression line is highly influenced by the two outliers: point (6.7990, 0.672) and (3.0000, 1.290). Other than the two outliers, the majority of the data are strongly positively correlated; higher listed price predicts higher sold price.

By removing the two outliers, we plot two scatterplots with the response variable being the sale price and the independent variable being the list price and the taxes. Properties in neighbourhood M (Mississauga) and in neighbourhood T (Toronto) is distinguished by different color (M: blue, T: red) and seperate simple regression models are fitted.

Adjusted Scatterplot



In the second plot, relationship between listed price and sold price is depicted, with listed price being the independent variable and the sold price being the dependent variable. For the purpose of fitting a better regression line, two points that are considered outliers ((6.7990, 0.672) and (3.0000, 1.290)) are removed. After fitting a seperate simple linear regression model, it is shown in the plot that there is nearly no difference between properties in neighbourhood M and neighbourhood T; both SLR show a strong positive linear relationship and have similar regression lines. Also, the regression lines of the subsets (neighbourhood M and T) are similar to that of the whole data.



The third scatterplot depicts the relationship between taxes and the sold price, with taxes being the independent variable and the sold price being the dependent variable. The data used in this plot is the same data used in the second plot. After fitting a seperate simple linear regression model to each neighbourhood, we can observe that both data have positive linear relationship and the slope estimate (β_1) of the two regression lines are the similar. However, for the same amount of property tax, the average of the sold price is higher in Toronto.

II. Methods and Model

Table

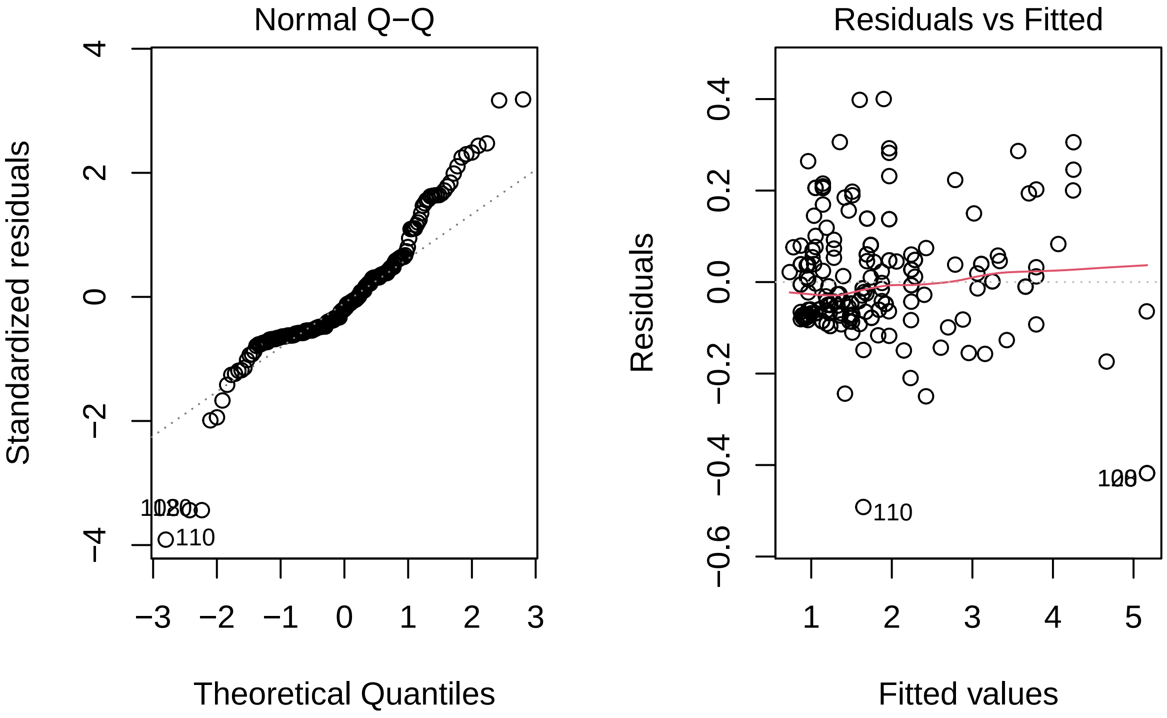
```
##       R-squared Estimated Intercept Estimated Slope
## M + T "0.983"   "0.1398"            "0.9143"
## M     "0.986"   "0.1416"            "0.8881"
## T     "0.982"   "0.1638"            "0.9183"
##       Estimate of the Variance of the Error Term p-value  95% CI for slope
## M + T "0.0159"                                   "0.0000" "(0.8972, 0.9314)"
## M     "0.0104"                                   "0.0000" "(0.8655, 0.9107)"
## T     "0.0177"                                   "0.0000" "(0.8941, 0.9425)"
```

According to the R square values, 98.30% of the variation in the dependent variable (sold price) in the first SLR (neighbourhood M + T combined) is explained by the regression line, 98.60% of the variation in the dependent variable (sold price) in the second SLR (neighbourhood M) is explained by the regression line, and 98.20% of the variation in the dependent variable (sold price) in the third SLR (neighbourhood T) is explained by the regression line. The R square values of the three data are similar. Since most of the points in the scatterplot fit the regression line after two outliers have been removed for the three datasets and it is previously shown that the regression line and SLR for the three datasets are similar, they have similar R square values. Namely, most of the variance for the three datasets are explained by the regression line.

Four conditions need to be met in order to apply the pool two sample t-test. The first and second is that both samples have to be normally distributed. Here, since the data size is big enough to assume normality, it is likely that both the subset of neighbourhood M and neighbourhood T follows the normal distribution. Third, the two samples have to be independent. This condition is met: the price in neighbourhood M is independent of the price in neighbourhood T; the house price in neighbourhood M will not influence the house price in neighbourhood T. Lastly, the two population need to have the same variance. Here, the variances of the two slope estimates need to be checked. Since the two SLR model is very similar, it is expected that the variance of the two slope estimates will be roughly the same. From the first to the forth condition, it is likely that every conditions are met. Therefore, a pooled two-sample t-test can be used here.

III. Discussions and Limitations

According to the fitted models in part II along with the second scatterplot in part I, I would choose the first fitted model in part II. By observing the difference between the values of the three models depicted in the table in section II, we can see that there are little to no differences. Since the estimated slope and intercept are similar, the three datasets have similar regression lines. Also, the second scatterplot in part I and the values in part II (variance of the error term) shows that it is neat to say that the difference is small enough to ignore and there is no need to have two seperate models for each neighbourhood.



According to the plot of the Normal QQ plot of the standardized residuals r_i, although there are some violations of normality in both tails, the majority of the residuals are normally distributed. Furthermore, the residual vs. fitted values plot gives a null plot. Namely, there is no pattern in the plot. Therefore, we can conclude homoscedasticity of the residuals. In sum, by checking the two plots, we can see that there are little to no violations of the normal error SLR assumptions for the selected model (neighbourhood M + T).

Another two potential numeric predictors that could be used to fit a multiple linear regression for sale price are the distance from public transportation (ex. TTC) and the age of the house. As the distance from public transportation decreases, the price is predicted to increase, as this means less walking time to the public transportation. Also, as the age of the house increases, the house price is predicted to decrease.