Analysis on SLR Models for Sale Price of Detached Houses in the GTA (.)

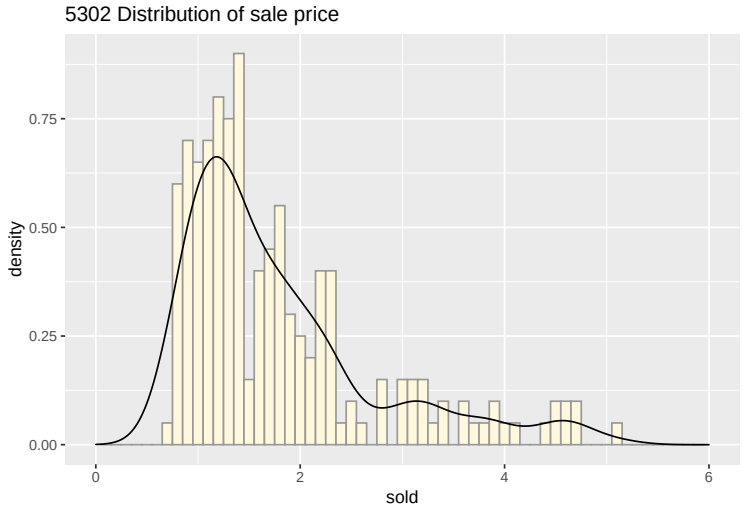
Chaerin Song

October 24, 2020

I. Exploratory Data Analysis

Brief explanation of the data: House price data was obtained from the Toronto Real Estate Board (TREB) on detached houses in the city of Toronto and the city of Mississauga.

Data distribution overview



The response variable (sale price) is an unimodal right skewed distribution. With its mean value of 1.793 million CAD, minimum value of 0.672 million CAD, and maximum value of 5.1 million CAD, the distribution is generally normal. There is no significant outlier, although the 5.1 million CAD data point could possibly be an outlier.

Removing influential points

Here are the data points with the top 5 highest Cook's distances for each of the sale price models, by list price and by taxes.

```
##         56        157        165         84        198
## 1.19983867 0.06890073 0.06890073 0.06373509 0.05902358
```

```
##          117           22          157          165           84
## 2.331380e+03 3.784363e-02 3.014401e-02 3.014401e-02 2.598430e-02
```

The Cook's distances for two of the above points are over 1, and their Cook's distance values are significantly big compared to other points.
Now that we have two points, [56] and [117], to be potentially removed, we can briefly investigate on these points before we actually remove them from our models.

```
## Leverage of the 117th point in the sold~list model: 0.966768768280612
```

```
## Standard residual of the 117th point in the sold~list model: -12.6599925582602
```

We usually conclude an i^{th} point as a leverage point if $h_{ii} < \frac{4}{n}$, where n is the total number of observations. Also, if the absolute value of i^{th} point's Standard residual, $|r_i|$, is substantially bigger than 2, we consider the point to be an outlier in the y-direction.
So the 117^{th} point in the sold~list SLR model is an outlier in both the x-direction and the y-direction, also called a *bad leverage point*.

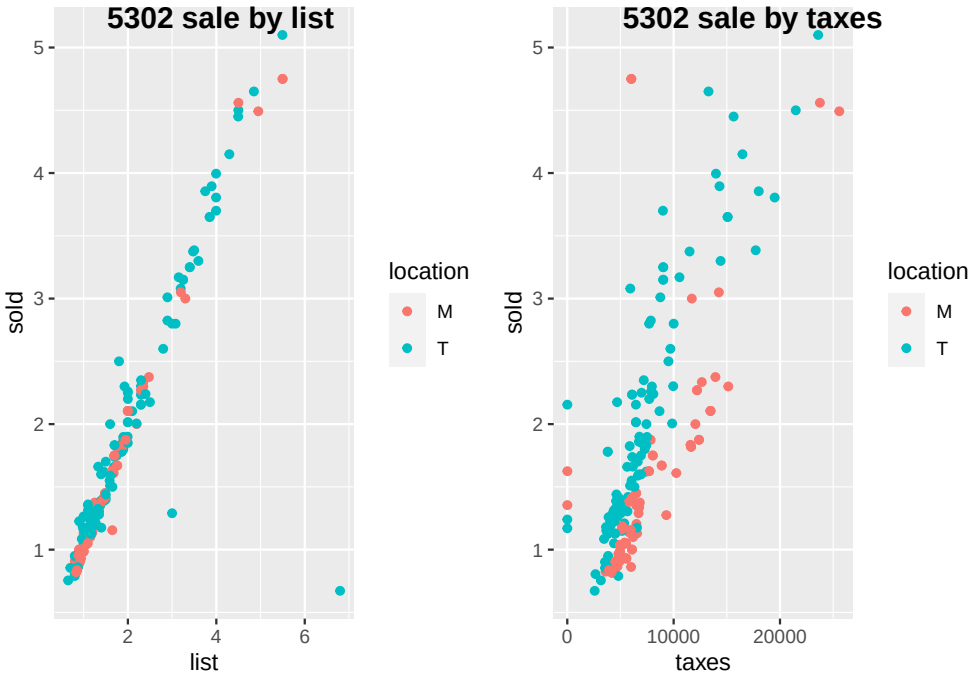
```
## Leverage of the 56th point in the sold~taxes model: 0.11100796221518
```

```
## Standard residual of the 56th point in the sold~taxes model: -4.38377540872113
```

Similarly, because the 56^{th} point in the sold~taxes model has an h_{ii} bigger than $\frac{4}{n=200} = 0.02$ and $|r_i|$ substantially bigger than 2, this point is also considered as a *bad leverage point* that should me removed.

Now that we have shown the validity of removing these two points, we will exclude them from the remaining parts of the analysis.

Scatterplots of the sale price



Interpretation of 3 different plots

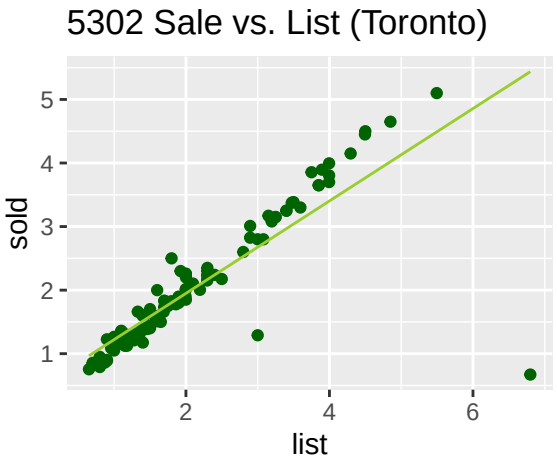
The first density plot shows the distribution of sale price data, which is what we are interested in the most. Here, our takeaway is that the house sale price in two Toronto neighborhoods is close to normally distributed, with some of very high priced houses.

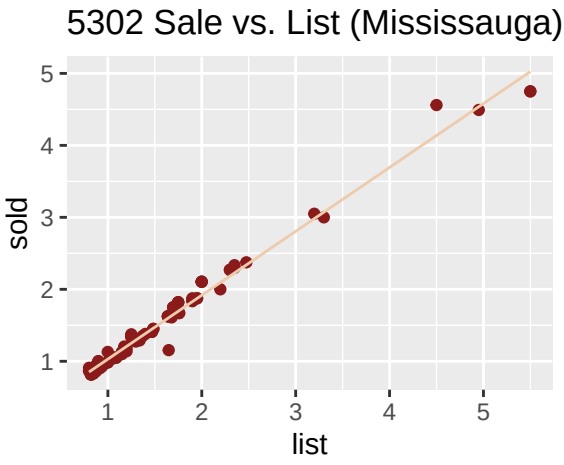
Next, from the scatterplots of sale price by list price and taxes, we can see the relationship between the two different factors.

The scatterplot of sale price by list price demonstrates a positive relationship between two factors, and is seemingly linear other than a couple of points.
Similarly, by looking at the scatterplot of sale price by taxes, the two factors seem to have a positive linear relationship, although the data plots are not as well aligned as the first scatterplot.

II. Methods and Model

3 Simple linear regression models of sale price by list price





Results

	all data	T	M
R^2	0.8109	0.7093	0.9852
b_0	0.3356	0.4956	0.1420
b_1	0.7897	0.7268	0.8878
Estimated variance of error term	0.4193^2	0.5367^2	0.1049^2
p-value for test of $H_0 : \beta_1 = 0$	0.0000	0.0000	0.0000
95% C.I for b_1	$0.7360, 0.8434$	$0.6393, 0.8144$	$0.8640, 0.9115$

P-value outputs are very small, so for convenience, we use 0.0000 here.

Interpretation of R^2 values

Note: R^2 gives the percentage of variation in y's explained by regression line. R^2 is not resistant to outliers, and is affected by the spacing of X. A high R^2 does not indicate that the estimated regression line is a good

fit.
General observation: R^2 for the model of T neighborhood is smaller than the R^2 for the model of all data, and R^2 for the model of M neighborhood is bigger than the R^2 for the model of all data.

As the scatterplot of T neighborhood shows, there are two data points that are placed far away from the rest of the data. These points would be considered as outliers and would have affected T neighborhood SLR model to have a noticeably smaller R^2 value than the all data model.
Meanwhile, if we look at the scatterplot of M neighborhood, it is easy to tell that M neighborhood has a high variation in X. This would have resulted in its SLR model having a bigger R^2 value than the all data model.

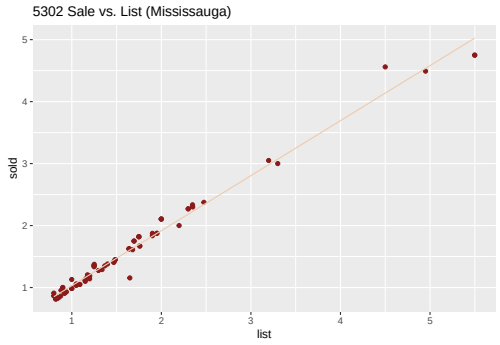
Can we use a pooled two-sample t-test for comparing the slopes?

Among all the assumptions we are making when conducting a pooled two-sample t-test, we can assume that both samples are simple random samples that are normally distributed, as that is our assumption for the all data in the previous SLR models. Also, we can say that the two samples are independent since there is no relationship between the individuals in different samples.

However, we cannot assume that the y_i's from the two populations have the same variance. If we want to use a pooled two sample t-test to determine if there is a statistically significant difference between the slopes of the two SLR models, we would first have to check for the equal variances.

III. Discussions and Limitations

The best fitted model



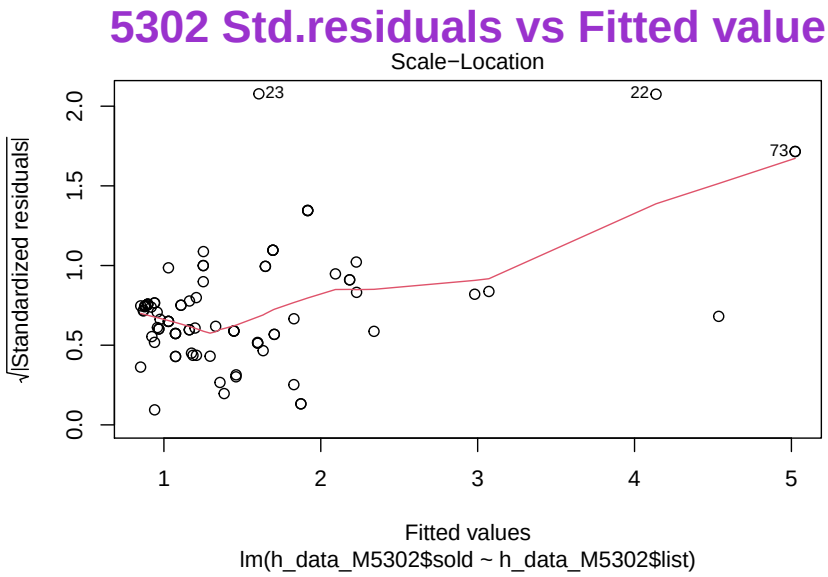
I chose the sales price model of the Mississauga neighborhood as the best among our three SLR models. Just by looking at the regression plot, there is no visibly extreme outlier that would have affected the slope significantly. Also, the slope looks very similar to the trend of most of the points. Although there are several big leverage points, none of them seem to be too influential on the model. To statistically show this, we can simply check the top values of Cook's distance:

```
##         22         73         76         23         35
## 1.30830956 1.16953716 1.16953716 0.11181576 0.02374695
```

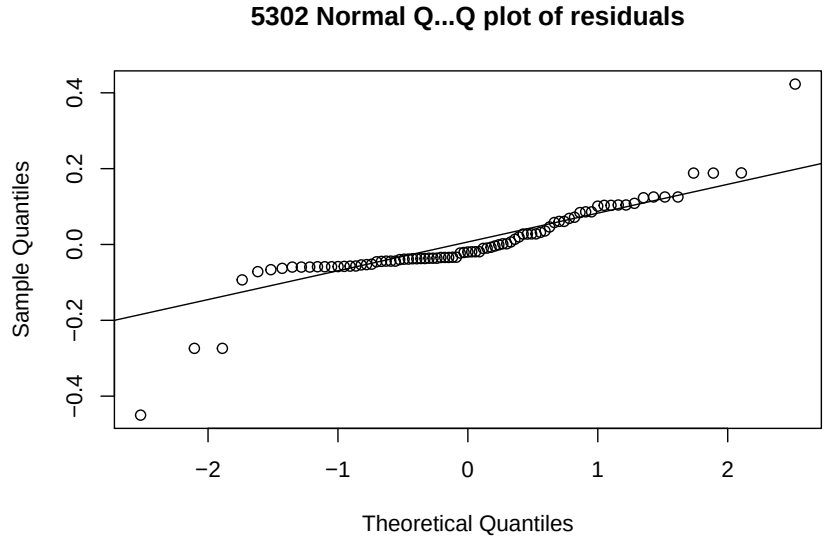
It turns out that there are 3 points whose Cook's distances are bigger than 1, but it is not substantial. So compared to the other two models where there are some noticeably influential points, this model seems to be the most valid fit. More validity will be checked in the following section.

Checking normal error SLR assumptions

Normal error SLR assumption is very important in determining the validity of a model. Here are our two residual plots:



If there is a homoscedasticity in the variance of errors, we should see a horizontal line with equally spread points. However, this is not the case in the above plot. Our plot shows a generally inconsistent slope with a general upward trend of the variances in the residual errors as fitted values increase, which suggests that our model violates the homoscedasticity part of the SLR assumptions.



In addition, the plot above suggests that our residuals might not be so normal. This plot is a normal QQ plot that visualizes the normality of data. Regardless of some inconsistent values at each end of the plot,

the general trend of the points does not align well with the straight qqline.

In short, there are some violations on the normal error SLR assumptions.
Although our models, including this Mississauga neighborhood model, are not perfect as of now, we hope to improve our models in future studies.

potential numeric predictors

There could be many different potential predictors of the actual price at which a house is sold.

First, the age of a house could be an important factor in its price. Older homes not only have outdated features, they are more likely to be in need of renovations. Generally, since homes that are newer appraise at a higher value, there could be a meaningful correlation between house sale price and age.

Secondly, it is known that the size of a house is a critical factor in house price. Depending on how many rooms there are and how many residents the house can accommodate, the price could drastically change. In order to further investigate on how much influence the size of a house has on its price, we are looking forward to conducting a statistical analysis.