Detached House Prices in Downtown Toronto and Mississauga: Multiple Linear Regression

IS3109

December 7, 2020

I. Data Wrangling

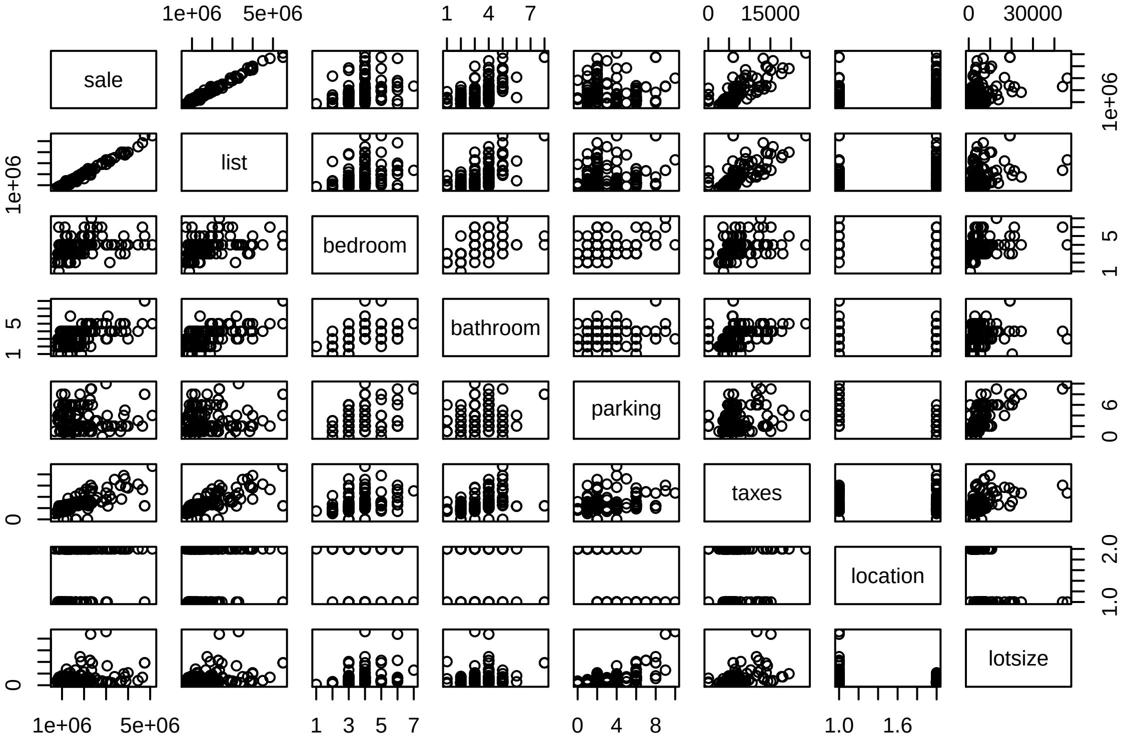
In this analysis, previously introduced, simple linear regression for detached houses prices in Toronto is expanded with additional variables. In this paper multiple linear regression (MLR) is going to be used. There are additional auxiliary variables in this paper such as number of bedrooms, parking spots and the size of the property. We expect multiple variables to account for more variation for the outcome variable than the variable in the previous paper, hence, improve our predictions for the sale price of a detached house.

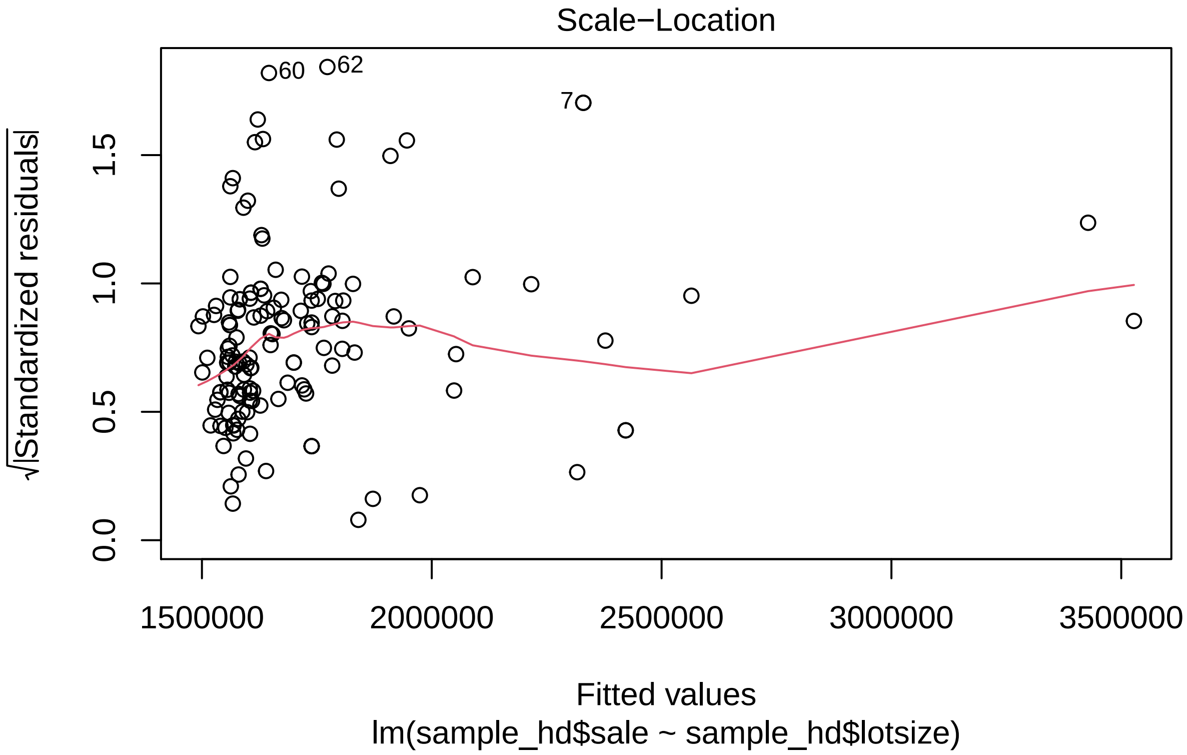
We have sampled 150 random points from the given data set. Further, we have created the variable lotsize which is obtained by the multiplication of variables lotwidth and lotlenght, this variable is considered to be the property size in feet since it is calculated by multiplying the frontage with one other side of the property. This is useful because there are numerous missing values for the variable maxsqfoot which is the maximum square footage of the property. After doing a variance inflation factor (VIF) test, we observe that variable maxsqfoot has the highest value, in addition to having far too many missing values, so it is removed from the data set due number of NA values and possible multicollinearity. Moreover, the diagnostic plots for this MLR analysis were viewed and four points were removed from the sample. The corresponding IDs for these points are 25, 65, 156, 194. These points were removed because points 65, 156 and 194 were high leverage points disturbing the fit of the model. Besides, points 25 and 194 was violating the normality assumption. After removing a variable and four data points from the model we have seen that the R^2 increased from 0.982 to 0.986 and the R^2_{adj} increased from 0.979 to 0.985. Hence, the final form of the sample with 146 observations and 9 variables is used for the rest of this paper.

II. Exploratory Data Analysis

From the 9 variables that are used in this analysis 4 of them are continuous, 4 are discrete and 1 is categorical. The categorical variable is the location of the house: if the house is located in downtown then the category is T (for Toronto), otherwise the category is M (for Mississauga). The continuous variables are sale, list, taxes and lotsize whereas the discrete variables are ID, bedroom, bathroom and parking.

The scatterplot matrix below shows the pairwise correlations between all variables in the sample. The most important correlations between variables are the ones with the outcome variable (i.e., sale). Hence, the Pearson correlation coefficients between the independent variables and the outcome variable are reported. Highest correlation is between the actual sale price and the last list price (0.99062), which, is most reasonably expected. This is followed by the correlation with previous year's property tax (0.73041), then surprisingly by number of bathrooms (0.6197). We didn't, in particular, expect to number of bathrooms to highly correlate with the sale price. This correlation coefficient is larger correlation between the sale price and of number of bedrooms (0.39829), property size (0.32474), number of parking spots (0.091753).





After looking at the scatterplots above and the diagnostic plots for the variables, we observe that the variable lotsize violates the homoscedasticity assumption for the linear regressions. We observe, from the scatterplot, points are much denser on the bottom left quadrant of the plot. Further from the square root of standardized residuals vs. fitted values plot located above, we observe a S shaped trend for the data points. We would like to observe no trend here, however, the residuals for the values of the property size are not distributed with a constant variance. Thus, the variable lotsize violates the assumption of homoscedasticity.

III. Methods and Model

The table below displays the coefficients, and the p-values for the corresponding t-tests of the coefficients.

```
##                                Estimate     P-value
## Intercept                    7.0170e+04  1.4540e-01
## Last List Price              8.4018e-01  2.8967e-82
## Number of Bedrooms           8.4632e+03  4.7736e-01
## Number of Bathrooms          3.2598e+03  7.8814e-01
## Number of Parking Spots     -1.0146e+04  2.1846e-01
## Previous Year's Property Tax  2.2253e+01  3.3234e-08
## Located in Downtown          9.0717e+04  9.4134e-03
## Property Size               -9.0440e-01  6.9007e-01
```

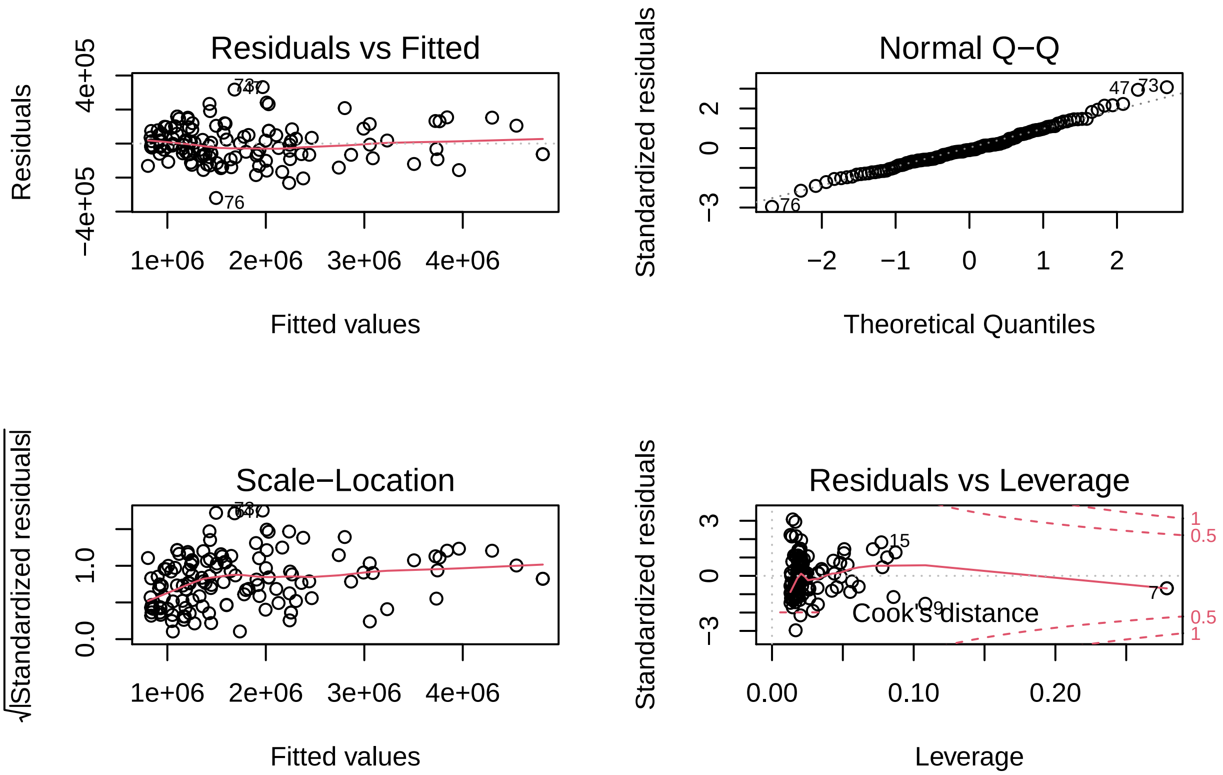
There are three statistically significant coefficients in this MLR. These are list, taxes and location. Considering the previous list price of the property we expected that it would strongly predict the outcome variable, since it also had a significant prediction in the previous paper. The p-value for the t-test was extremely small with $(2.8967 * 10^{-82})$. This is statistically significant in all confidence intervals. The estimate was approximately 0.84018. We can confidently say the estimate is not zero and has a most of the variance in the outcome variable can be explained for this variable. Following the previous listing price, estimate for taxes is also statistically significant for all confidence intervals. The p-value associated with the t-test is $(3.3234 * 10^{-8})$. Again, from the previous paper we have expected taxes to have a statistically significant relationship with the outcome variable. The corresponding coefficient estimate is roughly 22.253. The third significant predictor was the location so which neighborhood the house belonged to. The p-value for this estimate was 0.0094 making this variable statistically significant for $\alpha = 0.001$. This is still a really high significance considering the conventional 5% level. The coefficient estimate is approximately 90717. This estimate implies if all else equals house in downtown is approximately 90717 dollars more expensive than the one in Mississauga.

After conducting stepwise regression with AIC, we obtained the fitted model:

$\widehat{SalePrice} = 55000 + 0.841(List) + 21.8(Taxes) + 120000(Location = Downtown)$

From this model we see that with the backward elimination same three variables that were found statistically significant remain in the model. After eliminating for redundant variables we are left with only three. Although, the variables are the same and estimates are similar, the estimate for the variable location is considerably different from the previous table. The estimate for the location variable has increased approximately by 30000. This suggests a further difference in the prices of the houses between two neighborhoods: downtown and Mississauga. After doing BIC and AIC regressions we obtained the same regression from above. Hence, the final model is the above. After doing the BIC instead of AIC we get almost no change between the two regressions. There are no observed differences for estimates, p-values, R^2, nor the F-test.

IV. Discussions and Limitations



In this analysis the first assumption of the linear models that is a linear relationship between variables is satisfied considering the correlation and scatterplot between the sale price and the list price alone. In addition, we need to consider the diagnostic plots to make sure the assumptions are met. Looking at the residuals vs. leverage we can say there are not many leverage points besides one point (ID = 136) influencing the fit of the model. In addition, with randomization we would expect that the errors are uncorrelated. Homoscedasticity, however, can be questioned. We obsrve a somewhat linear pattern with a slight increase. We may need to expand our sample size to further analyze if the constant variance assumption is met. Lastly, we can say that normality of the errors assumption is somewhat satisfied. Out of the 134 observations from the AIC and BIC regression (8 additional data points were lost due to missing values for lotsize and parking) only 3 lay outside the normal line. We can say that there is an observed linear trend here with only few values not laying on the line. Again, since your sample is considerable small, expanding the sample size may benefit this assumption greatly. Normality is more challenging with smaller samples but considering the size of the sample for this paper we can say that the assumption is met.

The steps that should be taken for strengthening this research would be expanding the sample size, addition of different neighborhoods from the Greater Toronto Area, and some more additional variables. There may still be additional predictors for the actual sale price of the house but list and taxes associated with the property, as it is, may still explain most of the variation in the outcome variable.