Toronto/Mississauga Detached Homes Regression Analysis

Jack Smith, ID 1005034400

December, 4, 2020

I. Data Wrangling

Data Sample and Collection

The data was obtained from the Toronto Real Estate Board (TREB) on detached houses in Toronto and Mississauga. The data set contains information on sold homes such as the sale price, list price, number of bedrooms/bathrooms/parking spots, max square footage, property taxes, lot width/length and location. I took a random sample of 150 data points from the data set to analyze the data.

```
##   [1]  62 138 145 102 156 182 120 168 133 136 184  53  11  59  85 151  28 147
##  [19]  69  25  29 163  79 146 160  21 190  86  16 174  83  42 132  63  56  22
##  [37] 125  15  76 187 158  99 105  77  54 118  88 131 111  39 186  41 141 115
##  [55] 122  35 173  60 142  80  89 114  64  93 119  91  68  46 171  48 117 124
##  [73]  45  44 121  66  13   3  30 134 181 107 113  90  17 103 104 188  27  23
##  [91]  52  14  55   5   1  94 175  78 116 177  38 178  82 150 108 167  96 109
## [109] 155 176  71 139 152  73 130   4 154  72  58 129  61 148  43  97 165 162
## [127]   8 169  67  74 106 179 153 126  18  24 157  19  37  87  40   2  10  51
## [145]   9 101  32  92  57   6
```

Predictor Variable Removal

The predictor variable that was removed was max square footage of the property. The reason that max square footage was removed is that over half of the variables did not have answers for this predictor and had NA values instead. All the other predictors had a small number of NA values compared to max square footage. Having a very large number of NA values makes max square footage have less data to work with and could hurt the results, this is why I chose to remove it.

Data Removal

I choose to remove data points from my 150 point sample as there was multiple points that had existing NA values. After removing the the max square foot of the property variable, there were 10 data points that had NA values. I decided to only remove points that had NA values in them. When looking at the scatterplot there looks to be some points that would be influential. Of course I can not just remove the point without reason as just having large cooks distance or influence does not mean the variable has to be removed. The rest of the variable were not removed as looking at the values of the independent variables there did not seem to be reason that the variables did not fit with the rest of the data set (i.e. none of the values were unreasonable).

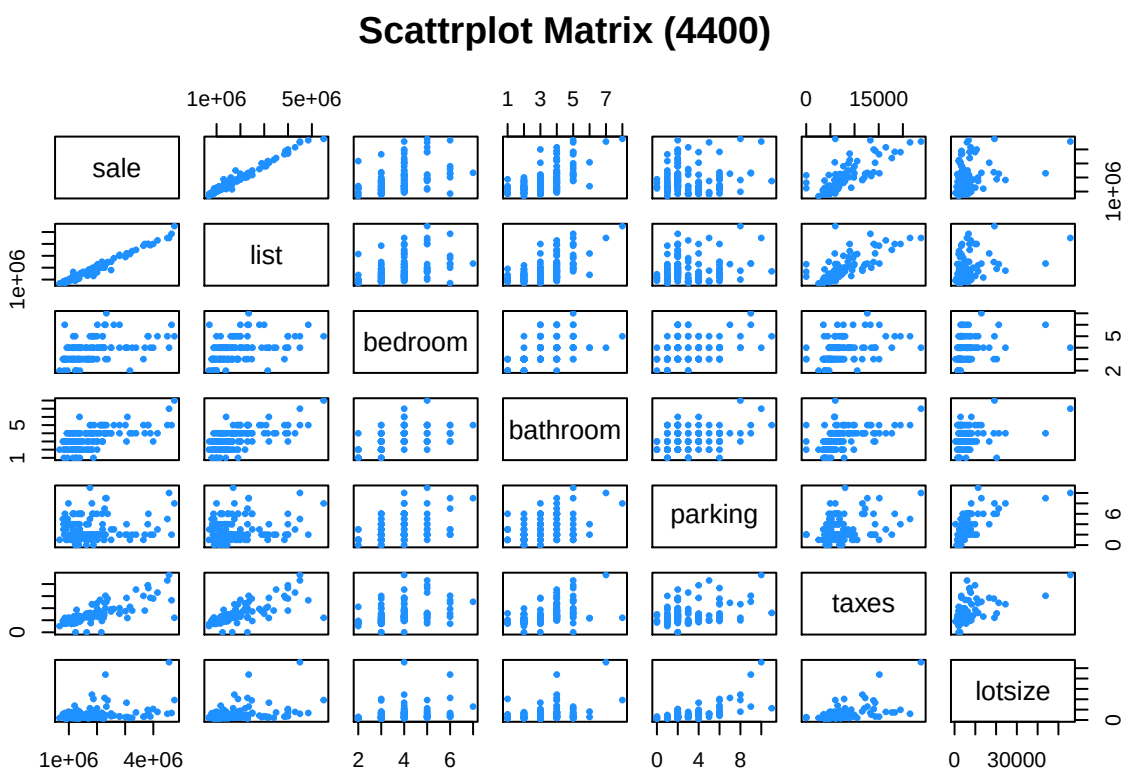
I. Exploratory Data Analysis

Variable Classification

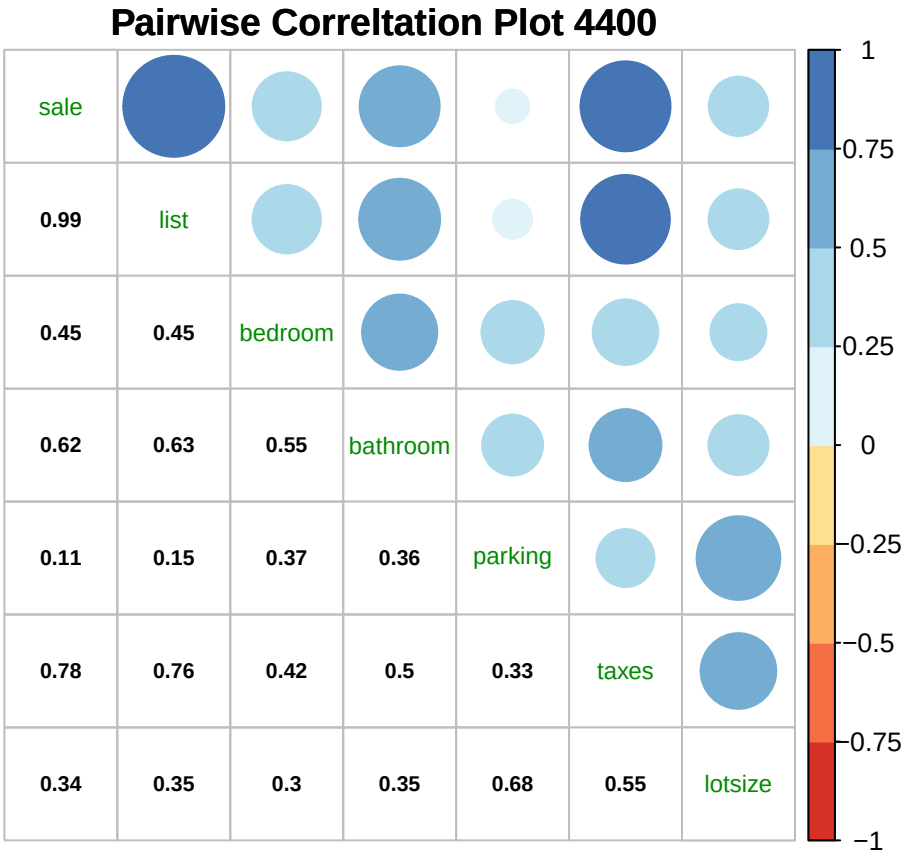
Each variable included in the data set can be classified as categorical, discrete or continuous.

Variable	Description
`Sale`	Discrete
`List`	Discrete
`Bedroom`	Discrete
`Bathroom`	Discrete
`Parking`	Discrete
`Maxsqfoot`	Discrete
`taxes`	Continuous
`lotwidth`	Continuous
`lotlength`	Continuous
`lotsize`	Continuous
`location`	Categorical

Scatterplot Matrix

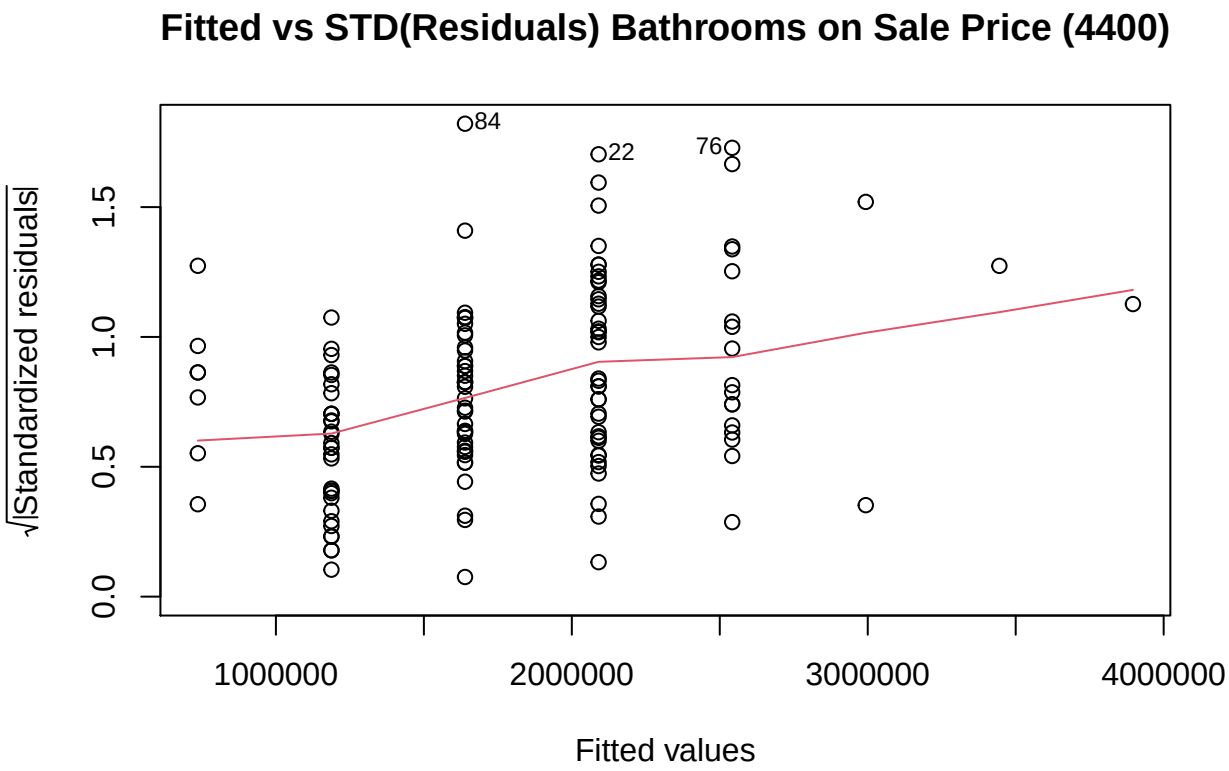


Pairwise Correlation Matrix



All of the variables have a positive correlation with sale price. The predictor with the highest correlation is list price with a 0.99 correlation coefficient which would make sense as the list price would be directly related to how much the home sells for. The next highest is property taxes paid at 0.78, property taxes are based on how much the property is worth which would have a strong effect on sale price. Bathroom and Bedroom count are next highest at 0.62 and 0.45 respectively, more bedrooms and bathrooms mean that larger families and more people can live their which would increase price. The bottom two are lot size and parking at 0.34 and 0.11 respectively, a larger lot means that there is more space and could increase price, while parking has a very small correlation at 0.11.

Variable with Non-Constant Variance



Based on the scatterplot matrix it looks like the bathroom predictor for sale price will have non-constant variance and the standardized residual plot is shown above for sale price and bathrooms. In the scatterplot matrix the variable bathroom looks to have increasing variance as the values near 1 bathroom have sale prices that are really close to each other and have less variation in sale price. Where as when the number of bathrooms gets larger there seems to be a wider range of sale prices and the variation has increased. As shown in the plot there is increasing variation as the fitted values get larger, this violates the MLR assumption that there is constant variance across all values.

II. Methods and Model

Additive Model

	Regression Coefficient	P-Value
Intercept	3.817099e+04	0.4764192
List Price	8.091612e-01	0.0000000
Bedrooms	9.588446e+03	0.5159529
Bathrooms	1.837606e+04	0.1702063
Parking	-1.945365e+04	0.0389223
Taxes	2.517904e+01	0.0000015
Toronto	1.045850e+05	0.0070500
Lot Size	3.536729e+00	0.1694753

The global F-test has a p-value of $< 2.2e\text{-}16$ which means that at least one of the predictors is significant. Looking at the actual p-values of the t-test result under the null hypothesis the values are zero, it shows

that 4 of the variables have a statistically different value than zero at the 95% confidence interval. Those 4 variables are List, Parking, Taxes and Toronto. The other variables can not be concluded to be statistically different from zero. Now I will Interpret only the values of that have significant t-tests, I get a model with list price, parking, taxes and the dummy location variable Toronto.

Looking only at the variables that have t-tests that are significant in the additive model presented above all of the predictors except parking are positive, meaning that as every variable but parking increases the expected sale price for a home goes up. Parking has a negative coefficient, the sale of price of the home is expected to decrease as parking spots increase. For a change in the number of parking spots, the expected sale price changes by 19450.

As its a linear model each variables effect on sale price for a change in x is given by $\Delta X_t \cdot B_t$ for each t predictor in the model. The Toronto variable is a dummy variable meaning it takes the values of 1 or 0, 1 if the house is in Toronto and 0 for Mississauga, this means that the expected sale price of a home being in Toronto is 104,600 higher than Mississauga. List price has a strong correlation of 0.99 from the correlation plot earlier and the effect of list price is 0.8092 with a very small p-value. Taxes have an effect of 25.18 for every change by 1 dollar.

AIC Model

The final AIC model is: $Sale = B_0 + B_1List + B_2Bathroom + B_3Parking + B_4Taxes + B_5Toronto$
or Sale = 4.60e+04 + 8.104e-01(List) + 2.000e+04(Bathroom) - 1.240e+04(Parking) + 2.777e+01(Taxes) + 1.043e+05(Toronto)

Therefore the AIC model dropped two predictor variables from the initial additive model, the two variables dropped were lot size and bedroom. Bedroom had the highest p-value and was dropped, lot size had a high p-value as well. But this is different from the model interpreted in additive model in the previous part as the predictor bathroom was not interpreted in the additive model section.

In terms of the signs of the coefficients they are still the same as the additive model. The intercept, list, bathroom, taxes and Toronto all have a positive value while the parking variable has a negative variable. All the variables except Toronto and parking have a larger coefficient in absolute value.

As the AIC model did not have the same predictors it is not consistent with the additive model.

BIC Model

The final BIC model is: $Sale = B_0 + B_1List + B_2Taxes + B_3Toronto$
or Sale = 4.394e+04 + 8.247e-01(List) + 2.616e+01(Taxes) +1.297e+05(Toronto)

The variables dropped from the initial full additive model are bedroom, bathroom, lot size and parking. The difference between the interpreted additive model and the BIC model is that additive model also had parking as a predictor. The variables in the BIC model take 3 of the 5 that the AIC did, the difference is that the BIC model also removed parking and bathroom from it's model.

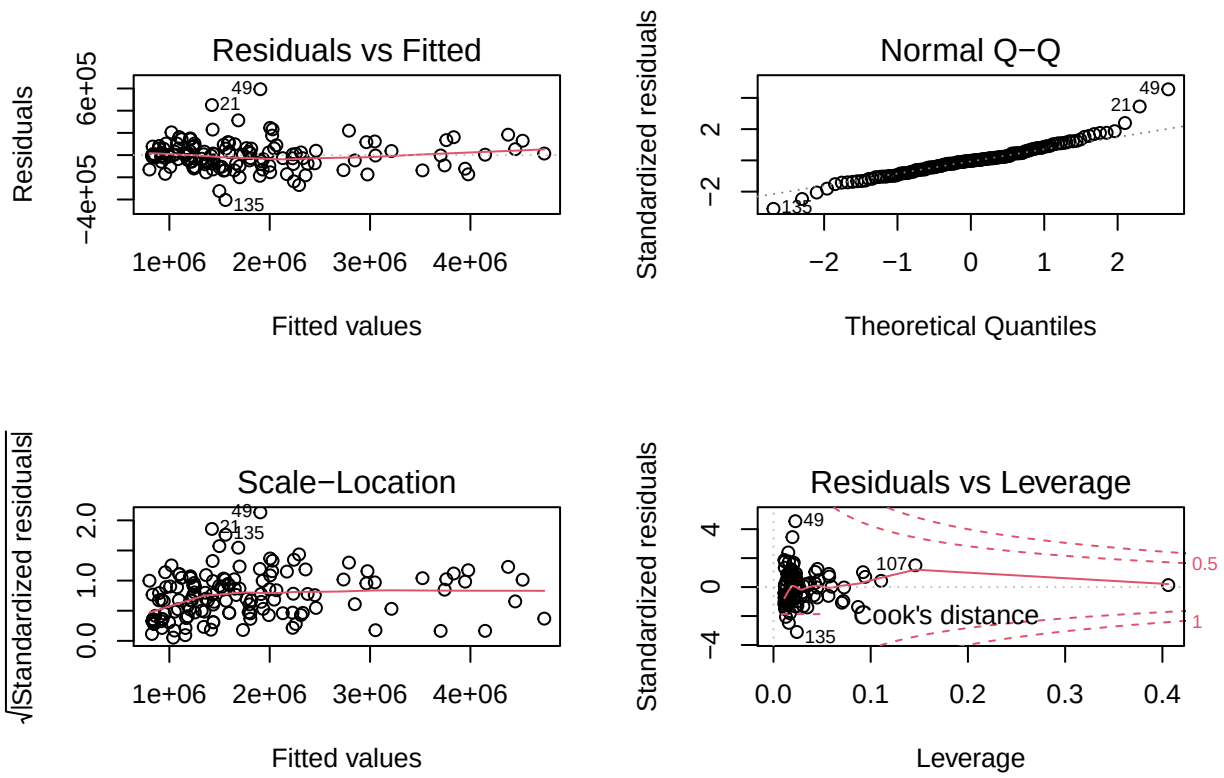
The three variables with the smallest p-values were the ones selected for the BIC model, only 1 variable that could be accepted at a 95% confidence interval was dropped from the model and that was parking.

In terms of the sign of the coefficient they were the same in both the AIC model and the BIC model as all the variables have a positive sign. The coefficient values were all larger in absolute value than in both the AIC and the additive model.

As the BIC model has a reduced equation with only 3 predictors it is not consistent with either the additive or AIC model.

III. Discussions and Limitations

Diagnostic Plots from obtained model



Plot Interpretation and Normal Error Assumption Disscussion

Examining the diagnostic plots for the BIC model, it can show if the model fits the linearity assumption and the normal error assumptions.

The Residuals vs Fitted graph shows that the model fits the linearity assumption. The graph has a few points that deviate but the linear fit works as there is no trend or pattern in the data.

The normal qq-plot shows that the errors follow a normal distribution, the majority of points are on the line and there is only a small amount of deviation at the tails of the distribution. Points 21/134/49 are seen to have larger deviation from the line but the error term still follows the normal distribution.

Looking at the standardized residual graph, it can be examined if the error term has constant variance. There is small deviation at the beginning and a few points 21/135/49 that are deviating but the thee does not seem to be a pattern. As no pattern exists and there is no trend, the error term should have constant variance.

The residuals vs leverage plot is helpful to find any influential cases if they exist. In the graph the purpose is to look for outlying values that could have an averse effect on the model and not necessarily for patterns. When looking at the graph it can be seen that there are no points that fall into the top left or bottom right of the graph. There are a few points of interest worth investigating by the graph such as 49, 135 and 107 that seem to have high leverage or standardized residuals.

Next Steps

In the next steps toward fitting a valid model, I would continue the steps toward fitting an optimal regression model.

The next step toward fitting a valid model would be doing an analysis of variance on the additive model and decide if there is a significant association between Sale price and any of the predictor variables. If there is a significant association check for redundancy in the model. If there is large amount of redundancy use a variable selection model like BIC or AIC calculated above. If there is not a great deal of redundancy then I would use a partial F-test to obtain the final model.

If I end up going with the BIC or AIC, I would look at the variables from the residuals vs leverage plot that could potentially have an averse effect on the model. I would decide if the points should be kept or be removed depending on if they fit in the model or not. If I removed any of the potential averse points above, I would re-fit they model and test it again, otherwise I would continue with the model.