# House Sales Price: A MLR Model

Afnan Rahman, Student number: 1004973031

December 5, 2020

First-time home buyers in Greater Toronto Area are facing extreme high rise in the prices for detached houses due to Covid impact. For analysis on the situation, we are establishing a multiple linear model, for buyers to predict the expected sale price of detached, single family homes in Toronto and Mississauga. For this model analysis we are provided with real203.csv data file that was obtained by the Toronto Real Estate Board (TREB) on detached houses in the two neighborhoods.The data frame has eleven parameters which are:

- ID: property identification
- sale: the actual sale price of the property in Canadian dollars
- list: the last list price of the property in Canadian dollars
- bedroom: the total number of bedrooms
- bathroom: the number of bathrooms
- parking: the total number of parking spots
- maxsqfoot: the maximum square footage of the property
- taxes: previous year's property tax
- lotwidth: the frontage in feet
- lotlength: the length in feet of one side of the property
- location: M - Mississauga Neighbourhood, T - Toronto Neighbourhood,

**I. Data Wrangling:**

**Part a. Creating Sample:**

In this part, a sample of 150 cases were randomly selected from the 192 observations of the given csv file.

The first 10 data IDs are shown below:

```
 [1] 175 160 153 190  55 134  63  10 111  42
```

**Part b. Addition of the new variable "Lotsize**

For this part, a new variable with the name 'lotsize' has been created by multiplying lotwidth by lotlength and lotwidth and lotlength were replaced with the newly created lotsize.

### Part c. Cleaned sample data

The final sample data was created by removing 10 rows containing at least one NA. Maxsqfoot, being the predictor with the highest number of NAs, was removed.

## II. Exploratory Data Analysis

### Part a. Classification of the variables:

ID: discrete

sale: continuous

list: continuous

bedroom: discrete

bathroom discrete

parking:discrete

taxes:continuous

location:categorical

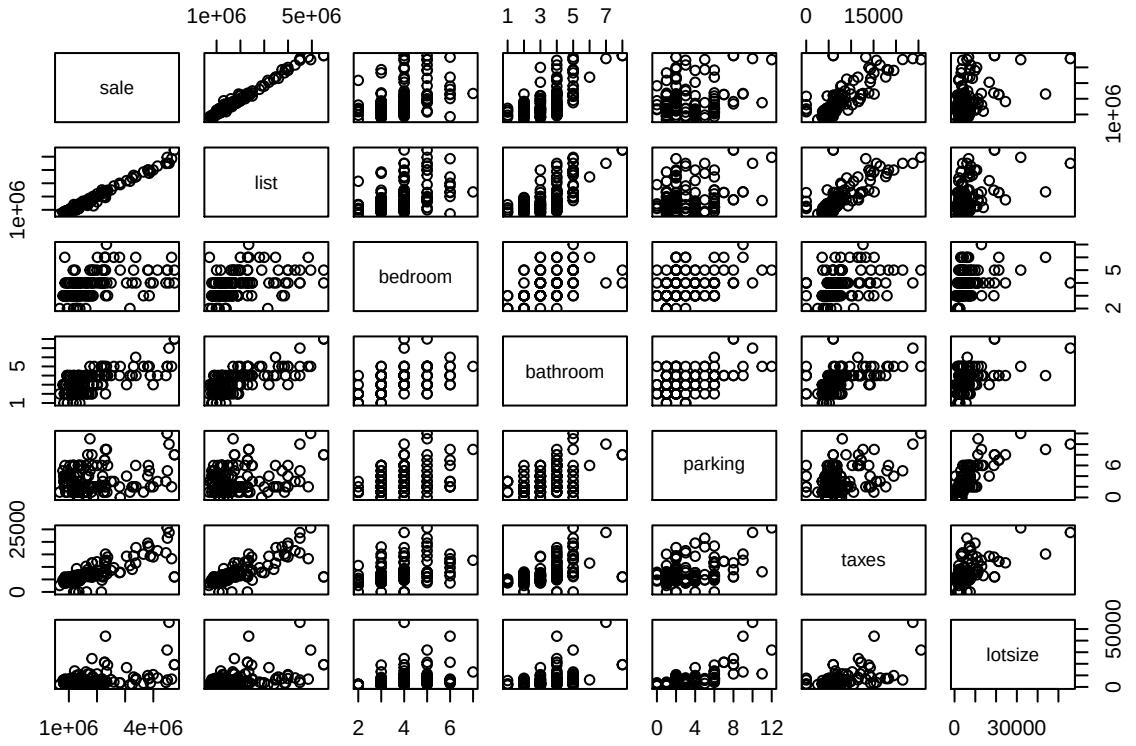
lotsize:continuous

### Part b. The Pairwise Correlations and Scatterplot Matrix:

*b.i) The pairwise correlations for all pairs of quantitative variables in the final sample data:*

| | sale | list | bedroom | bathroom | parking | taxes | lotsize |
|---|---|---|---|---|---|---|---|
| sale | 1.0000 | 0.9905 | 0.4498 | 0.6584 | 0.2419 | 0.7868 | 0.4205 |
| list | 0.9905 | 1.0000 | 0.4472 | 0.6824 | 0.2733 | 0.7645 | 0.4311 |
| bedroom | 0.4498 | 0.4472 | 1.0000 | 0.5370 | 0.3824 | 0.4028 | 0.3425 |
| bathroom | 0.6584 | 0.6824 | 0.5370 | 1.0000 | 0.4994 | 0.5017 | 0.4637 |
| parking | 0.2419 | 0.2733 | 0.3824 | 0.4994 | 1.0000 | 0.4033 | 0.7198 |
| taxes | 0.7868 | 0.7645 | 0.4028 | 0.5017 | 0.4033 | 1.0000 | 0.5854 |
| lotsize | 0.4205 | 0.4311 | 0.3425 | 0.4637 | 0.7198 | 0.5854 | 1.0000 |

*b.ii) Scatterplot matrix for all pairs of quantitative variables in the final sample data:*



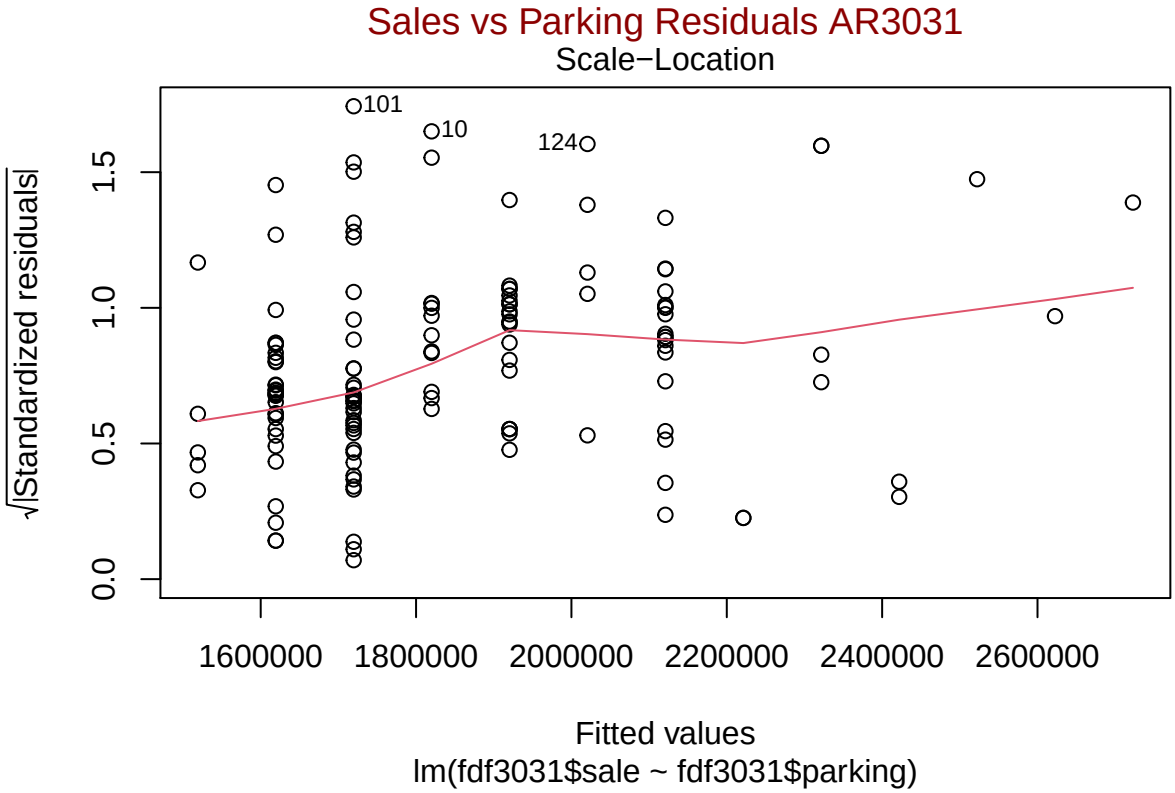

*b.iii) Rank of the quantitative predictors for sale price (in descending order):*

| Quantitative Predictors | Correlation Coefficients | Interpretation |
|---|---|---|
| list | 0.9905 | An almost perfect positive linear relationship with sale price. So when list price will increase it is almost certain that the sale price will increase. |
| taxes | 0.7868 | A strong positive linear relationship. So when taxes will increase the sale price will most likely increase |
| num of bathrooms | 0.6584 | A moderate positive linear relationship but significantly less than taxes. So when the # of bathrooms increases sale price will most likely increase but will not rise as high as taxes |
| num of bedroom | 0.4498 | A weak positive linear relationship. So when # of bedrooms will increase the sale price may/may not increase |
| lotsize | 0.4205 | A weak positive linear relationship. So when lotsize increases, the sale price may/may not increase |

| Quantitative Predictors | Correlation Coefficients | Interpretation |
|---|---|---|
| parking lot | 0.2419 | The predictor with the weakest positive linear relationship. So num of parking space may not have any affect in the sale price |

**Part c. Predictor that violates the assumption of constant variance:**

*Bedroom* has the strongest violation of the assumption of constant variance as the scatterplot shows the most fanning out of the values. The standardized residual plot is given below:



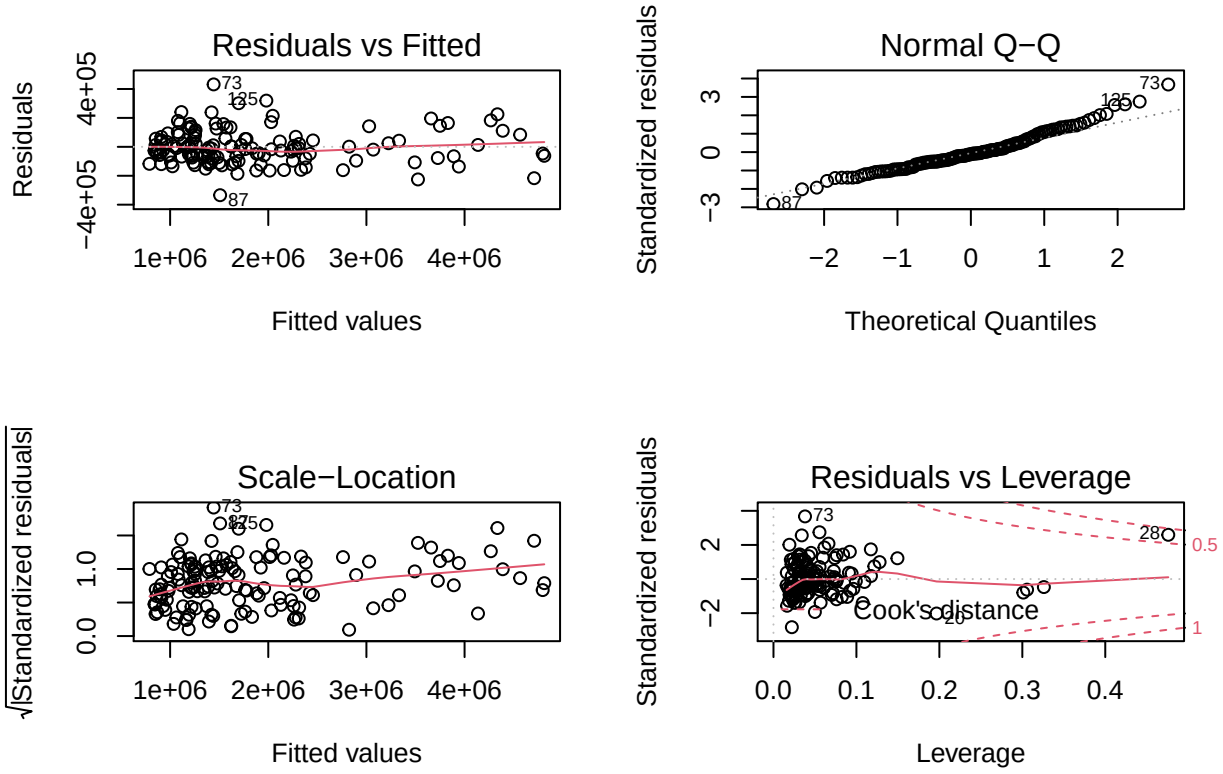

The increasing trend in the plot shows the *heteroskedasticity* for the parking predictor.Therefore, it can be certain that the assumption of constant variance has been violated here. A possible solution can be transforming the response variable, sale price through log transformation [log(Y)] or through weighted least square method.

## III. Methods and Model

### i. Analysis on the Additive Linear Regression model:

In this part, an additive linear regression model is fitted including all available predictors variables (list, taxes, bathroom, bedroom, lotsize) for sale price. As ID is not a useful predictor variable to make inference for the sale price analysis, therefore ID has not been considered in the additive model. Moreover, since the

location is a categorical variable, therefore, it has been taken as a dummy variable where 1 indicates Toronto and 0 indicates Mississauga.




To view the significant predictors along with the list the estimated regression coefficients and their p-values for the corresponding t-tests, summary of the additive model has been shown in the table below:

| | Est_Reg_Coefficients | t-value | p-value |
|---|---|---|---|
| Intercept | 57703.0472671214 | 1.12835994784421 | 0.261214579961051 |
| List | 0.836768790308036 | 41.7020827286279 | 7.18629472586476e-78 |
| Taxes | 20.6039435588065 | 5.1106345162982 | 1.10096085303572e-06 |
| Bedroon | 3525.01155223125 | 0.25787653381281 | 0.796903803895331 |
| Bathroom | 15904.9995649666 | 1.28853752541797 | 0.199812799350255 |
| lotsize | 1.01970490343411 | 0.441155190223318 | 0.659822386818839 |
| Parking | -12285.2850389614 | -1.50791178251192 | 0.133966785251982 |
| Location | 81190.9761432422 | 2.26305089973957 | 0.0252653049090106 |
| | significant? | | |
| Intercept | no | | |
| List | yes | | |
| Taxes | yes | | |
| Bedroon | no | | |
| Bathroom | no | | |
| lotsize | no | | |
| Parking | no | | |
| Location | yes | | |

From the above summary, it can be viewed that only list, taxes and location are significant in the additive model. The equation of the model is given below (where location is the dummy variable 1 when it is Toronto and 0 when it is Mississauga)

$\hat{sale} = 57703.04727 + 0.0.8368List + 20.6039Taxes + 81190.9761Location$

**List**: The model has *0.8368* as it's estimated regression coefficient for the predictor *list*. The positive value suggests that if all other variables are held constant then with one unit increase in the house's listed price, the sales price will increase 0.8368 units.

**Taxes**: For *taxes*, the estimated regression coefficient of *20.6039* suggests if all other variables are held constant then with one unit increase in the housing tax, the sales price will increase 20.6039 units. So in comparison with list, the change in taxes will increase the sales price more drastically. To be accurate, the taxes have (20.6039/0.8368 =) 24.62% more positive influence in the sale price than list mentioned above.

**Location**: The dummy variable *location* has the estimated regression coefficient of *81190.9761*, which indicates that house price will increase 81190.9761 units if it is in Toronto in comparison with same house (where all other predictors are held constant) being in Mississauga. Therefore, it can be inferred that location is a huge factor to determine buyers' affordability to buy a detached house in the GTA.

### ii. Backward elimination Method: AIC Model

The summary of the AIC model:

| | Est_Reg_Coefficients | t-value | p-value |
|---|---|---|---|
| Intercept | 6.145477e+04 | 1.316561e+00 | 1.902333e-01 |
| List | 8.403981e-01 | 5.330716e+01 | 4.299680e-92 |
| Taxes | 2.096147e+01 | 5.664309e+00 | 8.632368e-08 |
| Bedroom | 1.672518e+04 | 1.429590e+00 | 1.551621e-01 |
| Parking | -1.070574e+04 | -1.444214e+00 | 1.510129e-01 |
| Location | 7.616275e+04 | 2.256885e+00 | 2.563382e-02 |

**The final fitted AIC model:** As only three predictors have been listed in the BIC Model, therefore the final fitted BIC model is:

$\hat{sale}(AIC\ \ fitted\ \ model) = 61454.7676 + 0.8404\ \ List + 20.9615\ \ Taxes + 16725.18\ \ Bedroom - 10705.74\ Parking + 76162.7460\ Location$

In part i the additive linear regression model found was:

$\hat{sale}(Additive\ Model) = 57703.04727 + \ 0.8368\ List + 20.6039\ Taxes + 81190.9761\ Location$

The coefficients of list and taxes of AIC_model is quite consistent with the additive model. However, the AIC_model has 2 extra predictors (Bedroom and Parking) that the additive model did not have. The estimated coefficient for bedroom being *16725.18*, suggests that with a additional bedroom in the house the price will increase by 16725.18 CAD. This finding realistically makes sense because the sale price of house will most likely increase if there were more bedrooms in the house. However, it is a bit inconsistent with our findings from part II. There, a weak linear correlation were found between sales vs # of bedrooms (*0.4498*) . So we cannot be certain that the inference made for the Bedroom predictor in the AIC model is valid without finding further evidence.

The estimated coefficient for parking is *-10705.74*. This negative value suggests that with all other predictors held constant, if number of parking space increases then it will decrease the house's sale price. This is inconsistent with the sales vs parking relation found in the previous analysis. Even though the correlation

found in part I was the lowest amongst *(0.2419)* the all predictors, but still they had a positive relation. Even realistically, if a house has more parking lot then it is unlikely that it will reduce the price of the house.

In part a. as the the full model is being fitted, therefore the coefficients are weighted based on all covariates, including the insignificant ones. In this part, as the AIC method eliminates the variables that penalize the model without adding any information. Due to the reduction in the dimension, the coefficient estimates are not the same.Moreover, the inconsistency can be due to not removing enough outliers as the initial data set had 10 rows of NA and our data removal was set at 11 rows. Therefore, the influential points may have resulted the inconsistency in the findings in the AIC model.

Further analysis can be done based on the summary of the AIC model. From the summary(AIC_model3031) it has been found that only list, taxes and location are significant predictors in t-tests. This finding is consistence with our model in part i. To visualize the finding a table has been added below:

| Predictors | significance |
|---|---|
| list | yes |
| Taxes | yes |
| Bedroom | no |
| Parking | no |
| Location | yes |

As we have more predictors in the AIC model then expected, therefore, it can be inferred that, the AIC model is overfitting the data.

**BIC Model**

The summary of the BIC model:

```
          Est_Reg_Coefficients      t-value      p-value
Intercept         7.232227e+04 2.978087e+00 3.435470e-03
List              8.415976e-01 5.579939e+01 1.318216e-95
Taxes             2.010230e+01 5.543686e+00 1.485567e-07
Location          1.094614e+05 5.045456e+00 1.421986e-06
```

**The Final Fitted BIC model:** As only three predictors have been listed in the BIC Model, therefore the final fitted BIC model is:

$\hat{sale}(BIC\ Model3031) = 72320 + \ 0.8416\ List + 20.10\ Taxes + 109500\ Location$

In part i the additive linear regression model found was: $\hat{sale}(Additive\ Model3031) = 57703.04727 + 0.8368\ List + 20.6039\ Taxes + 81190.9761\ Location$

In part ii the final fitted AIC model was: $\hat{sale}(AIC\ fitted\ model) = 61454.7676 + 0.8404\ List + 20.9615\ Taxes + 16725.18\ Bedroom - 10705.74\ Parking + 76162.7460\ Location$

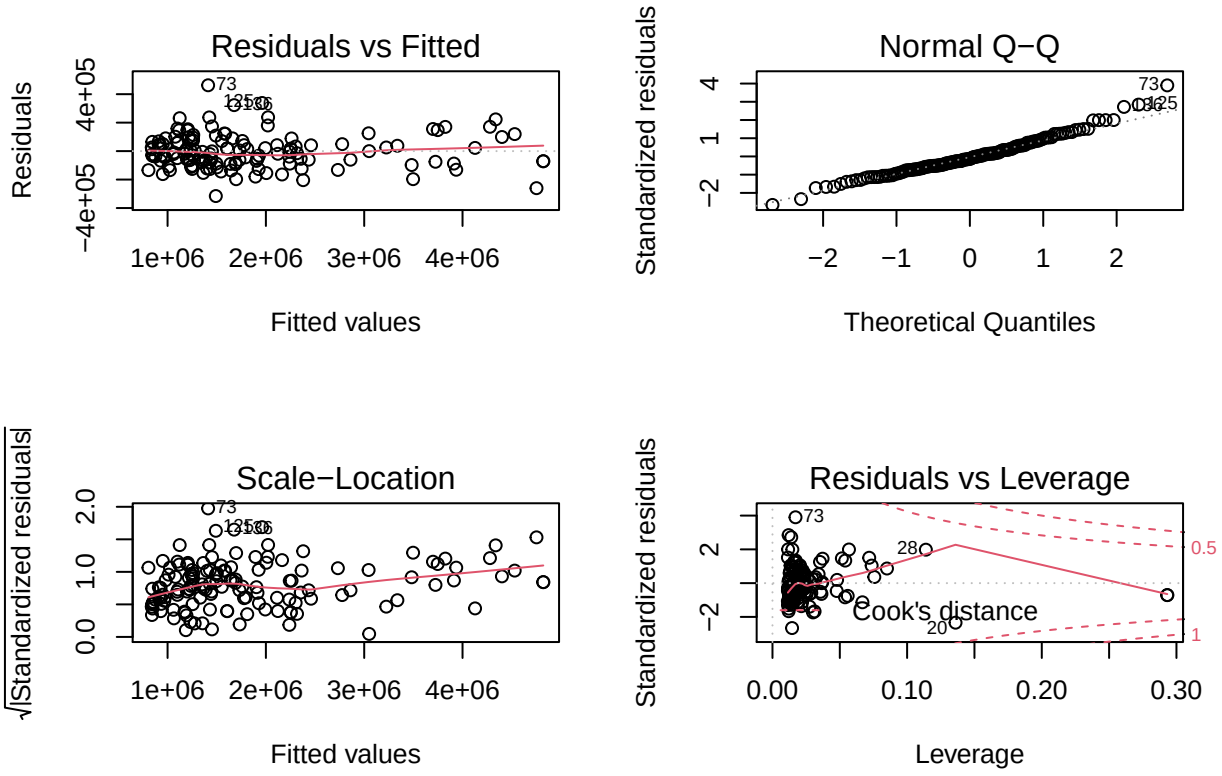
BIC is consistent with the additive model in terms of eliminating the insignificant predictors and identifying the significant ones (list, taxes and location). The coefficients of list and taxes of BIC_model is quite consistent with the additive and AIC_models'. However, the intercept for BIC model turns out to be the quite higher than additive and AIC models'.

To explain the inconsistency, similar explanation as AIC model can be provided here. In part a. full model's coefficients are weighted based on all covariates (both significant and insignificant ones) where the BIC method further eliminates the variables that penalize the model. Due to the reduction in the dimension,

the coefficient estimates are found to be different. Moreover, the inconsistency can be due to not removing enough outliers as the initial data set had 10 rows of NA and our data removal was set at 11 rows. Therefore, the influential points may have resulted the inconsistency in the findings in the AIC model.However, overall the BIC model showed a better consistency with the previous analysis than the AIC model.

## IV. Discussions and Limitations

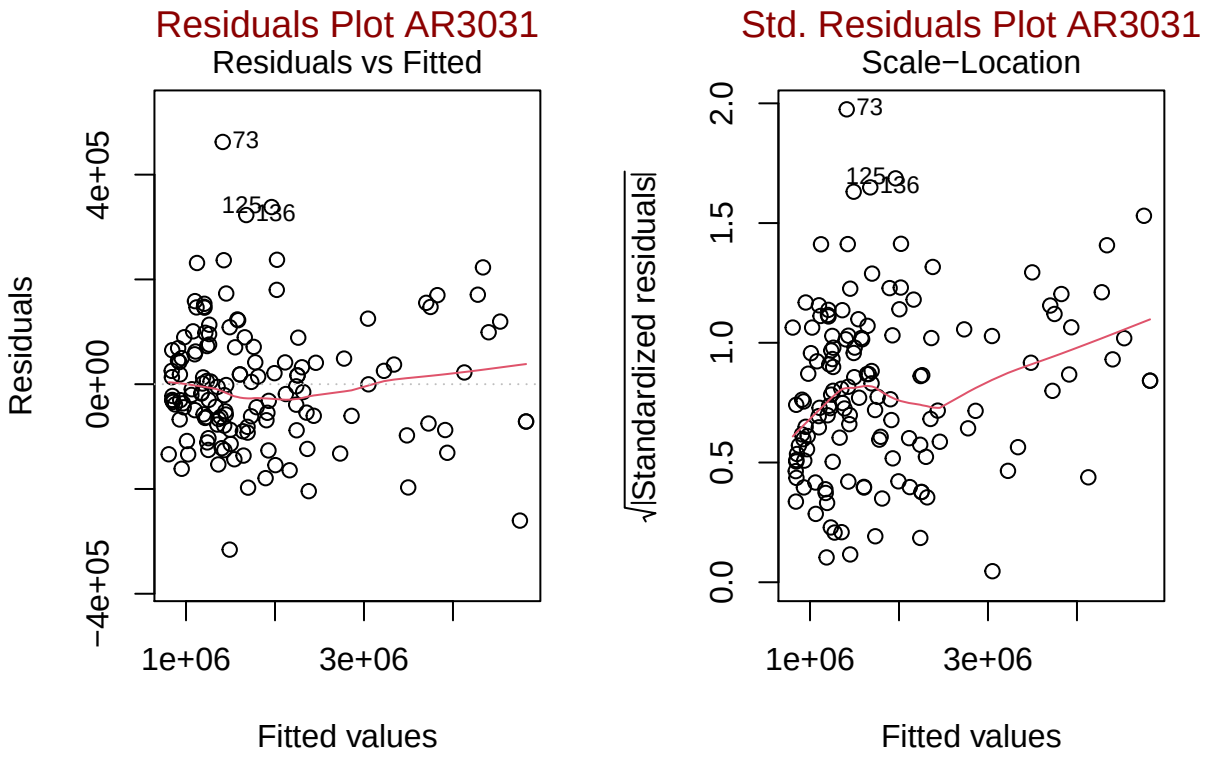
### a. 4 diagnostic plots of BIC Model:



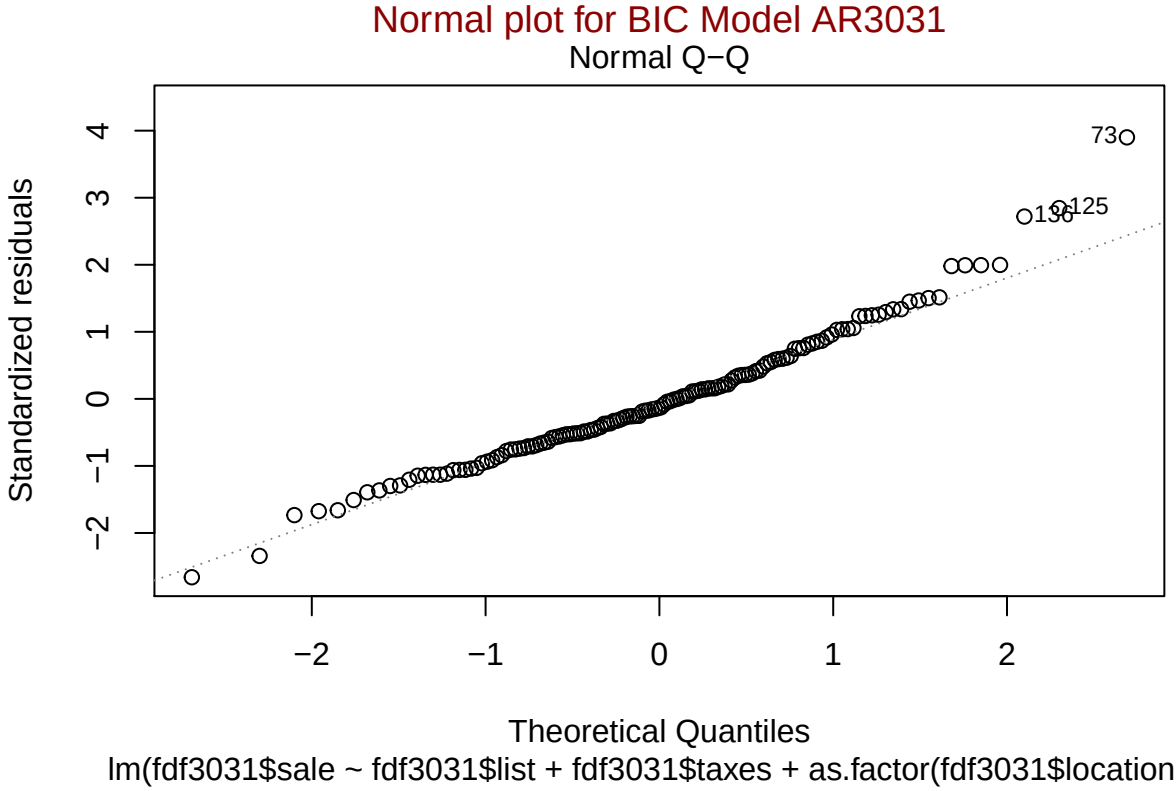

### b. Plot Interpretation:

Firstly, to identify the validity of the final sample (fdf: final data frame), the interpretation of the last plot (Residuals vs leverage plot) is being considered. From the plot in part a, it is evident that there is no outliers in the data. Therefore, our data frame creation shows evidence of validity.

Now, the first and third plots are being used to check for constant variance. The plots are given below:




The Residuals vs Fitted plot indicates the predictors values (list+taxes+location) on the x axis, and the residual on the y axis. And the Scale-Location plot also takes the predictors but only difference is, here, the x values are plotted against the square root of the standardized residuals.Both of these plots help identify heteroscedasticity of the residuals. The non linear relation in the 1st plot and the increasing trend in the 2nd plot indicates heteroscedasticity, therefore the assumption of constant variance has not been satisfied.




The normal Q-Q (quantile-quantile) plot here is illustrating the quantiles of the BIC model residuals. As the data fairly follows the 1:1 diagonal line. therefore it can be inferred that the residuals are normally-distributed.

**c. Possible next steps towards finding a valid 'final' model:**

The goal of this analysis is to provide a "House sales price prediction model" to a larger group of future house buyers. All three models in part III have shown some inconsistency. In the BIC final fitted model, the assumption of the constant variance has been violated. However, it is quite consistent with the additive model found in part III a. So we will be proceeding with the BIC model and make it more valid, there are few things we need to fix.

*Firstly,fixing non-constant variance:*

To solve the heteroscedasticity of our model, a weighted least squared regression is suggested.

*Secondly. checking for mutlicolinearity:*

We can utilize the variance inflation factors (VIF) testing for this. If VIF for the predictors are less than five then we can infer that there is no structural multicolinearity. Another way to check for multicollinearity is to do correlation among the predictors. If the correlation is higher than 0.5 then the relevant predictors show significant amount of multicollinearity. Below is shown the correlation among the predictors of the data:

```
           list bedroom bathroom parking  taxes lotsize
list     1.0000  0.4472   0.6824  0.2733 0.7645  0.4311
bedroom  0.4472  1.0000   0.5370  0.3824 0.4028  0.3425
bathroom 0.6824  0.5370   1.0000  0.4994 0.5017  0.4637
parking  0.2733  0.3824   0.4994  1.0000 0.4033  0.7198
```

```
taxes    0.7645  0.4028   0.5017  0.4033 1.0000  0.5854
lotsize  0.4311  0.3425   0.4637  0.7198 0.5854  1.0000
```

Here we can see that both of the numerical independent variable in our suggest model (BIC) has high collinearity *(0.7645)* which may lead to unwanted results in the model. To solve this, besides linearly combining the independent variables through additive modeling, we can also carry out analysis like partial least squares regression (PLS). It is based on covariance and widely used to do analysis where there are multiple explanatory variables and they are suspected to have correlation.

*Thirdly, taking a k-fold Cross-validation approach*:

This would help assess the predictive ability of our updated final model and its robustness by evaluating their performance on a new data set.

---