Multiple Regression Lab

Gauri Patil

10/18/2020

**Description:**

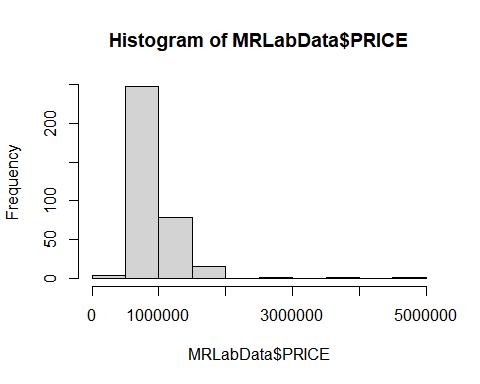
The housing market of Kirkland, WA is very competitive. In this lab, we perform regression modeling on a housing dataset of Kirkland. We have collected this data from Redfin website for the houses, townhouses and condos sold in Kirkland zip code 98034 over the last one year. The dataset consists of 350 rows and 27 variables from which we choose five namely Beds, Baths, Square Feet, Lot size and Year Built for our liner model.

**Summary Statistics:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** | **NA’s** |
| **Price** | 471000 | 705000 | 795500 | 921413 | 1053750 | 4925000 |  |
| **Beds** | 0.000 | 3.000 | 4.000 | 3.737 | 4.000 | 7.000 |  |
| **Baths** | 0.750 | 1.750 | 2.500 | 2.401 | 2.750 | 4.500 | 3 |
| **Lot Size** | 2347 | 7200 | 8030 | 9754 | 9875 | 90604 |  |
| **Year Built** | 1920 | 1969 | 1977 | 1983 | 1996 | 2020 |  |
| **Square feet** | 790 | 1600 | 2075 | 2233 | 2745 | 6070 |  |
| **$/Square feet** | 116.0 | 364.2 | 417.0 | 428.3 | 469.2 | 1653.0 |  |
| **Days on market** | 3.0 | 46.0 | 109.0 | 140.9 | 223.2 | 364.0 | 18 |
| **Zip or postal code** | 98011 | 98034 | 98034 | 98034 | 98034 | 98034 |  |
| **HOA/Month** | 9.00 | 37.50 | 51.50 | 71.41 | 80.00 | 425.00 | 286 |
| **MLS#** | 1345182 | 1550792 | 1594644 | 1588315 | 1633014 | 1672651 | 18 |
| **Latitude** | 4.7.70 | 47.71 | 47.72 | 47.72 | 47.73 | 47.74 |  |
| **Longitude** | -122.3 | -122.2 | -122.2 | -122.2 | -122.2 | -122.2 |  |

# Plots:

#### Price:



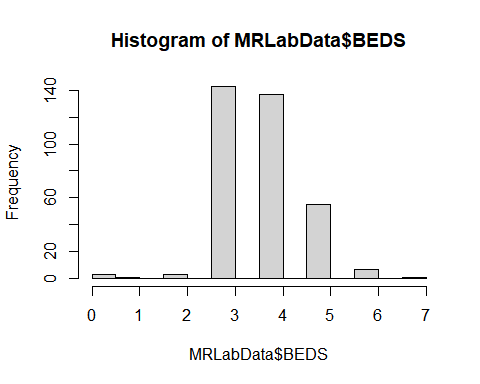
The distribution for price is slightly normal and the histogram for price is right skewed. Highest bar is for houses in the range of 50k to 1million indicating most houses are in this range. It would be useful to see data transformation like log on this variable.

#### Beds vs Price Scatterplots:

|  |  |
| --- | --- |
|  |  |

We see most houses have 2 – 5 beds. There are some outliers with 0 beds and 7 beds. A few outliers with 3 – 4 beds but very high prices 3 to 5 million. Beds being discrete variable shows lines of points. The correlation doesn’t seem linear. We see a wide confidence interval on both ends of the regression line due to very few data points.

#### Histogram for Beds:



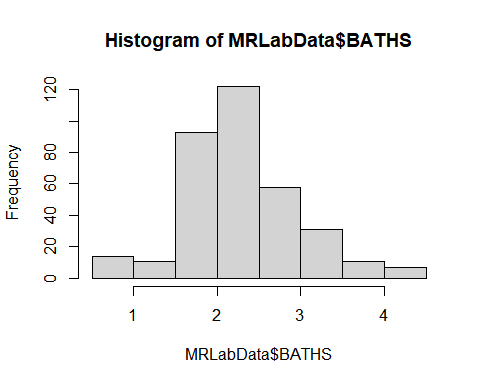
The histogram for beds seems to be somewhat normal with less frequency of houses towards the left side which are with less beds.

#### Baths vs Price Scatterplots:

|  |  |
| --- | --- |
|  |  |

The correlation between baths and price seems somewhat linear. We see a smaller number of baths have lower prices and as the number of baths increases the price range goes higher too. There are some outliers towards the top of the graph with exceptionally high prices. We see an almost straight regression line with narrow confidence interval signifying the error range.

#### Histogram for Baths:



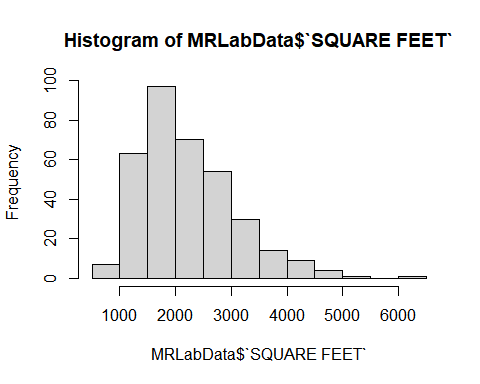
The distribution of baths is very normal. The peak is at around 2.5 which is the mean and median.

#### Square feet vs Price:

|  |  |
| --- | --- |
|  |  |

We see most of the houses have square footage between 1000 to 4000. The relationship between square feet and price is very linear. We see few outliers with very high prices. The regression line looks more exponential than straight line but it has low error range.

#### Histogram for Square Feet:



The histogram for square feet area of the house is normal with some skewness towards the right. We can consider removing outliers with high square footage to correct the skewness.

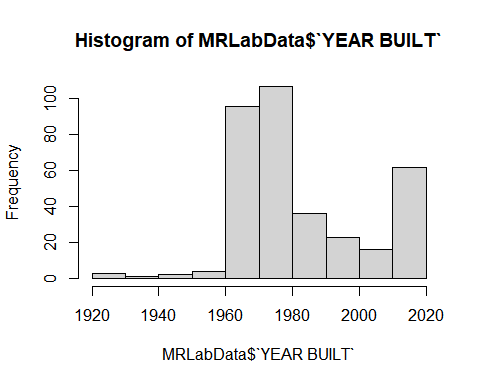
#### Year Built vs Price Scatterplots:

|  |  |
| --- | --- |
|  |  |

Kirkland houses were mostly built after 1960 and there are very few houses built before that time. We can say that Kirkland started populating in 1960. And then we can again see after 2000 less houses were built. During the recession in 2009 we also see dip in number of houses which increases significantly towards the current 2020 year.

We don’t see any linear correlation between year built and price of houses. The regression line is kind of slightly u-shaped flat line.

#### Histogram of Year Built:



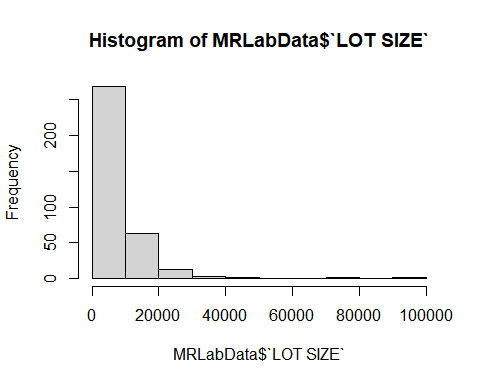
The histogram shows similar pattern with more frequency of houses during certain range of years and dip between 2000 to 2010.

#### Lot size vs Price Scatterplots:

|  |  |
| --- | --- |
|  |  |

We see a lot sizes mostly concentrated at one area and there are few outliers with very high prices and very high lot sizes. The relationship does not look linear. The regression line is wavy and has a broad confidence interval.

#### Histogram of Lot Size:

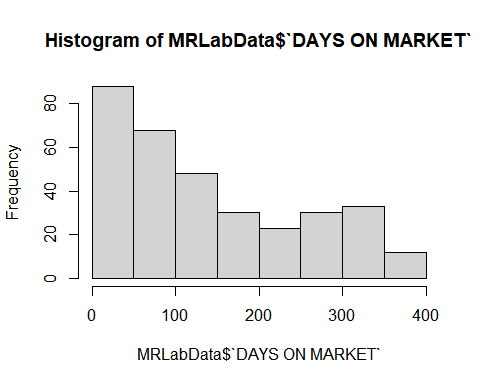


The histogram for lot size is not normal. We see a very high bar with most houses between 0 to 10000 and a long tail towards the right. It would be helpful to see how the histogram changes by removing the outliers from the right-hand side with high lot size.

#### Days on market vs Price Scatterplots:

|  |  |
| --- | --- |
|  |  |

#### Histogram of Days on Market:

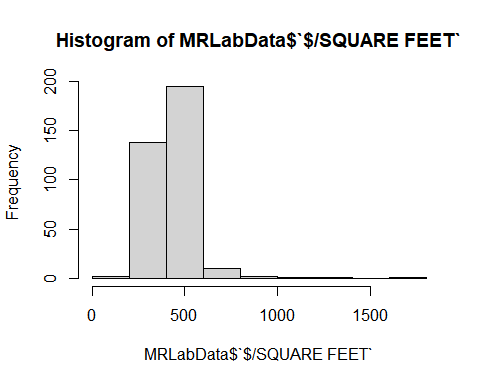


Days on market data points do not show any correlation with price. Also, the histogram shows that the distribution is not normal.

#### $/Square Feet vs Price:

|  |  |
| --- | --- |
|  |  |

#### Histogram of $/Square feet:



Dollar per square feet scatterplot shows data points concentrated in one area with no linear correlation with price. It can be helpful to see after removing outliers. The histogram is slightly normal with right skewness.

# Delete missing data:

After selecting a subset of dataset in MRLAB2 which includes price, beds, baths, square feet, lot size and year built columns, we delete the missing data from these variables using complete cases.

summary(MRLAB2)

## PRICE BEDS BATHS SQUARE FEET   
## Min. : 471000 Min. :0.000 Min. :0.750 Min. : 790   
## 1st Qu.: 705000 1st Qu.:3.000 1st Qu.:1.750 1st Qu.:1600   
## Median : 795500 Median :4.000 Median :2.500 Median :2075   
## Mean : 921413 Mean :3.737 Mean :2.401 Mean :2233   
## 3rd Qu.:1053750 3rd Qu.:4.000 3rd Qu.:2.750 3rd Qu.:2745   
## Max. :4925000 Max. :7.000 Max. :4.500 Max. :6070   
## NA's :3   
## LOT SIZE YEAR BUILT   
## Min. : 2347 Min. :1920   
## 1st Qu.: 7200 1st Qu.:1969   
## Median : 8030 Median :1977   
## Mean : 9754 Mean :1983   
## 3rd Qu.: 9875 3rd Qu.:1996   
## Max. :90604 Max. :2020   
##

MRLAB3 <- MRLAB2[complete.cases(MRLAB2), ]  
summary(MRLAB3)

## PRICE BEDS BATHS SQUARE FEET   
## Min. : 471000 Min. :1.000 Min. :0.750 Min. : 790   
## 1st Qu.: 706000 1st Qu.:3.000 1st Qu.:1.750 1st Qu.:1600   
## Median : 796000 Median :4.000 Median :2.500 Median :2080   
## Mean : 923869 Mean :3.769 Mean :2.401 Mean :2237   
## 3rd Qu.:1055000 3rd Qu.:4.000 3rd Qu.:2.750 3rd Qu.:2745   
## Max. :4925000 Max. :7.000 Max. :4.500 Max. :6070   
## LOT SIZE YEAR BUILT   
## Min. : 2347 Min. :1920   
## 1st Qu.: 7200 1st Qu.:1969   
## Median : 8045 Median :1977   
## Mean : 9780 Mean :1983   
## 3rd Qu.: 9900 3rd Qu.:1997   
## Max. :90604 Max. :2020

From the MRLAB2 summary output below we can see that there were 3 NA’s in baths columns. After applying complete cases to select the rows without NA’s we can see the output of MRLAB3 dataset has no NA’s in bath column.

# Regression

Below is the output of linear regression using all five variables.

HouseFit1<-lm(PRICE~BATHS+BEDS+`SQUARE FEET`+`LOT SIZE`+ `YEAR BUILT`, data=MRLAB3)  
summary(HouseFit1)  
## Call:  
## lm(formula = PRICE ~ BATHS + BEDS + `SQUARE FEET` + `LOT SIZE` +   
## `YEAR BUILT`, data = MRLAB3)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -1284375 -83322 -11306 68093 2524611   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7340240.366 1914836.223 -3.833 0.000151 \*\*\*  
## BATHS -8583.197 40450.777 -0.212 0.832087   
## BEDS -120682.856 24337.497 -4.959 0.00000112 \*\*\*  
## `SQUARE FEET` 372.956 33.980 10.976 < 0.0000000000000002 \*\*\*  
## `LOT SIZE` 2.226 2.252 0.988 0.323666   
## `YEAR BUILT` 3974.981 978.300 4.063 0.00006013 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 271600 on 341 degrees of freedom  
## Multiple R-squared: 0.5642, Adjusted R-squared: 0.5578   
## F-statistic: 88.29 on 5 and 341 DF, p-value: < 0.00000000000000022

**Model:** The R – squared value for this model is 0.5578 and the P- value is very low, almost zero. Hence this model is significant, and it will explain 55.78% variation in the house price.

We see that three out of five independent variables are statistically significant except for Baths and Lot size which have P-value higher than 0.05.

**Bath:** The bath coefficient has negative around 8k, so one bath increase in bath will cause 8583$ decrease in price. The standard error is very high 40k as compared to the variation estimate. And the P value is greater 0.05 so this bath variable is not significant.

**Beds:** The P-value is less than 0.05 and almost equal to 0. We see it marked with 3 \*’s showing it is significant variable. The estimate value is very high negative value which means that one bath increase in beds will cause 120683$ decrease in price of the house.

**Square feet:** We observe that the P-value of square feet is very less almost 0. Hence, square feet is the most significant among all the selected variables. The estimate coefficient is 373, which will be the increase in price for every square feet increase.

**Lot size:** The estimate lot size coefficient is 2 which would be the $ price increase per increase in lot size holding the other variables constant. The P-value is higher than 0.05 and hence lot size is not significant.

**Year Built:** The P-value for year built is very low, almost zero and due to which this variable is significant. For per year increase in year built the price increase would be 3975$.

Next, we perform the regression again by choosing the significant variables which are beds, square feet and year built.

HouseFit2<-lm(PRICE~BEDS+`SQUARE FEET`+`YEAR BUILT`, data=MRLAB3)  
summary(HouseFit2)  
## Call:  
## lm(formula = PRICE ~ BEDS + `SQUARE FEET` + `YEAR BUILT`, data = MRLAB3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1307370 -87885 -13082 69375 2502577   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6449736.38 1618235.42 -3.986 0.000082241 \*\*\*  
## BEDS -127259.32 23265.71 -5.470 0.000000087 \*\*\*  
## `SQUARE FEET` 381.39 24.52 15.552 < 0.0000000000000002 \*\*\*  
## `YEAR BUILT` 3529.56 826.42 4.271 0.000025239 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 271200 on 343 degrees of freedom  
## Multiple R-squared: 0.5629, Adjusted R-squared: 0.5591   
## F-statistic: 147.2 on 3 and 343 DF, p-value: < 0.00000000000000022

We see that there was minor improvement in R-square from 0.5578 to 0.5591. And the p-value remains the same so this model significant and little better considering that we haven’t explored the non-linear correlations and data transformations.

We check another regression model by considering just square feet variable as it is the most significant one.

HouseFit3<-lm(PRICE~`SQUARE FEET`,data=MRLAB3)  
summary(HouseFit3)  
## Call:  
## lm(formula = PRICE ~ `SQUARE FEET`, data = MRLAB3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1127269 -106356 4235 85629 2694319   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 161353.14 43684.52 3.694 0.000257 \*\*\*  
## `SQUARE FEET` 340.91 18.26 18.668 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 288500 on 345 degrees of freedom  
## Multiple R-squared: 0.5025, Adjusted R-squared: 0.5011   
## F-statistic: 348.5 on 1 and 345 DF, p-value: < 0.00000000000000022

The R-square value actually decreased to 0.5011, so we conclude that the second model Housefit2 is the best model in this case.

# Residuals:

|  |  |
| --- | --- |
|  |  |
|  |  |

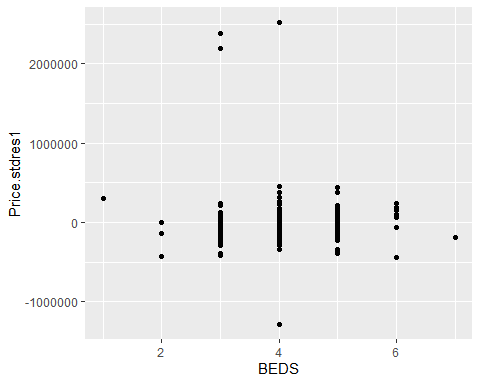
The residual points are close to the center line with few outliers. The redline for the plotted residuals is pretty straight line except for slightly distanced towards the start and end.

The QQ plot shows a decent straight light which signifies that the residuals have normal distribution except for few outliers shown in right top corner.

We see that the red fitted line for standardized residuals which is slightly wavy but still close to flat line.

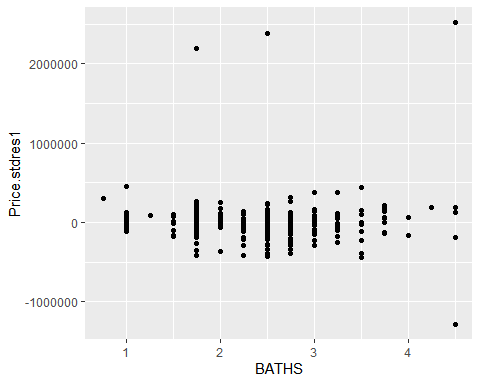
Residuals vs leverage shows two residual points 272 and 265 which are very close the cook’s distance. These observations might be causing leverage and worth further exploration.

#### Beds vs Residuals:



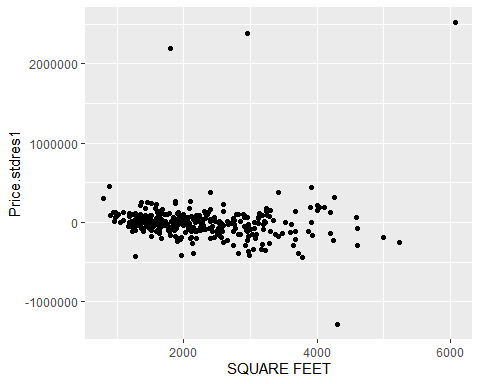
We see a few outliers also observed in the plots before. It might be worth exploring the if the model improves by removing these outliers. We see some concentration towards the center at 3,4 & 5 beds which reduces the heteroscedasticity.

#### Baths vs Residuals:



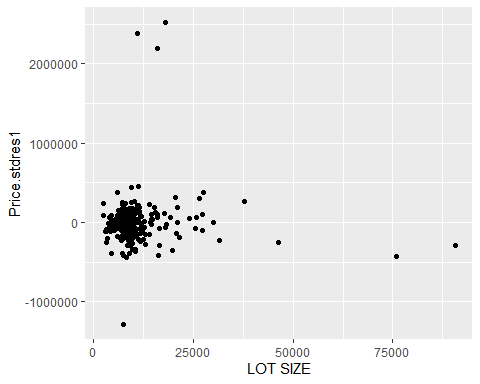
Baths vs residuals has residual points more dispersed than beds showing better heteroscedasticity. We see some outliers on the top with very high prices.

#### Square feet vs Residuals:



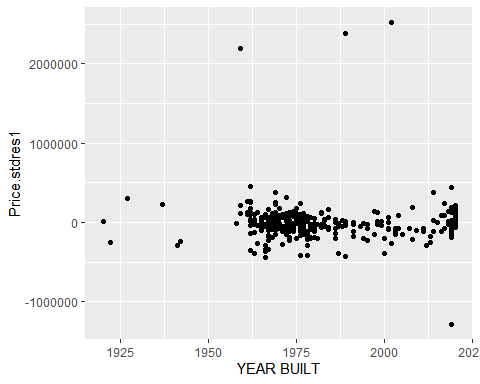
We see that the residual points are more cloud shaped and spread out. There are some outliers which is consistent observation for all the residual plots.

#### Lot size vs Residuals:



The residual point for lot size is more like a blog shaped concentrated in one area which show a pattern.

#### Year built vs Residuals:



For year built vs residuals we see the points concentrated in one or two areas causing less heteroscedasticity.

# Log Transformation:

As we noticed nonlinearity in the scatter plots, log data transformation is worth exploring.

MRLAB3$LOGPRICE <- log(MRLAB3$PRICE)  
MRLAB3$LOGBed <- log(MRLAB3$BEDS)  
MRLAB3$LOGBath <- log(MRLAB3$BATHS)  
MRLAB3$LOGSqFoot <- log(MRLAB3$`SQUARE FEET`)  
MRLAB3$LOGLotSize <- log(MRLAB3$`LOT SIZE`)  
MRLAB3$LOGYearBuilt <- log(MRLAB3$`YEAR BUILT`)

HouseFitlog<-lm(LOGPRICE~LOGBed+LOGBath+LOGSqFoot+LOGLotSize+LOGYearBuilt, data=MRLAB3)  
summary(HouseFitlog)

##   
## Call:  
## lm(formula = LOGPRICE ~ LOGBed + LOGBath + LOGSqFoot + LOGLotSize +   
## LOGYearBuilt, data = MRLAB3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.15115 -0.08018 0.00044 0.07983 1.33862   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -84.255602 10.331674 -8.155 0.00000000000000675 \*\*\*  
## LOGBed -0.235351 0.063704 -3.694 0.000257 \*\*\*  
## LOGBath 0.003562 0.059861 0.060 0.952582   
## LOGSqFoot 0.576572 0.055583 10.373 < 0.0000000000000002 \*\*\*  
## LOGLotSize 0.132471 0.026630 4.975 0.00000103886385922 \*\*\*  
## LOGYearBuilt 12.199632 1.359067 8.976 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.187 on 341 degrees of freedom  
## Multiple R-squared: 0.6768, Adjusted R-squared: 0.672   
## F-statistic: 142.8 on 5 and 341 DF, p-value: < 0.00000000000000022

The R-squared value of the above model which uses log transformation of variables is significantly high than the previous model. The adjusted R-squared increased to 0.672 from 0.5591 and the model is significant as the P- value is less than 0.05 and very close to zero.

Only LOGBath has P- value higher than 0.05 but by removing the LOGBath the model did not improve much. The adjusted R -squared value without LOGBath was 0.673.