Statistical Data Mining | Hunters Green Home Sales

Pablo X Zumba

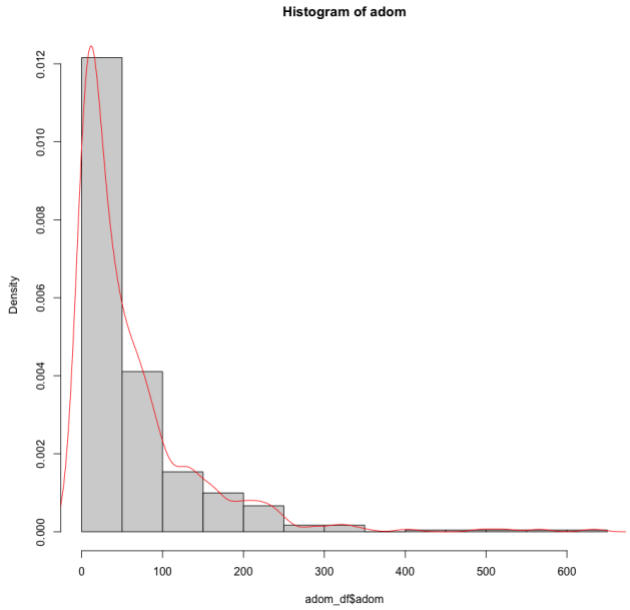
Create statistical models to predict: (1) agent days on market (adom), which is essentially the number of days it took to close the sale from the date of listing, and (2) sale price of the home (price sold) based on relevant attributes in this data set.

1. Create a table of relevant predictors, hypothesized direction of effect (+/-), and rationale for each hypothesized effect. (2 points)

Predictor	Effect	Rationale
DV: adom		
sqft	+/-	The price of a house increases with its size, making it harder to sell, so the square footage and "adom" are directly proportional.
pricesold	-/+	It's very likely that it will take less time to sell a house if the price sold is lower than the list price (the "adom" is decreased).
listprice	+/-	The list price of a house could be deemed more appealing to potential buyers at a particular moment, resulting in a faster or slower sale (directly proportional to adom).
yrhouse	+	The original variable "yrblt" was replaced by "yrhouse" (how many years a house is) by subtracting 2022 from the current year (see data transformation in appendix). The adom could be reduced if the house is newer, tending to zero years, and longer if the house is older so its relationship is directly proportional to predict adom and inversely proportional to predict price sold.
garages	-	Until the number of garages reaches three or four, the garage-adom relationship is directly proportional, since too many garages don't make a house more desirable.
Excluded: **slnoskm**(identifier hence not useful), **status**(all values are sold, it will introduce bias), **bathstotal, bathsfull, bathshalf**(repetitive and very correlated to sqft), **address**(same area code all observations), **subdivn**(same as address variable), **pendingdate**(it can be usefull for time series regression but no for this problem), **datesold**(it can be usefull for time series regression but no for this problem), **splsale**(Not useful based on the table, since it has 453 None observations so it would introduce bias to the prediction), **beds** (very correlated to sqft), **roof**(it has 9 categorical values but only 2 are relevant), **lotsqft(**very correlated to sqft hence no necessary)**, pool** (predominance of private, not too relevant), **spa(**not relevant for adom)**, cdom**(Cdom is supposed to be equal to or greater than adom, but not less, and as the correlation plot shows(see appendix), it is highly correlated, but it makes no sense to take it into account for predicting adom), **lppersqft**(very correlated to sqft), **sppersqft**(very correlated to lppersqft).		
DV: pricesold		
sqft	+/-	Directly proportional to the price sold.
bathstotal	+	Directly proportional to the price sold and at the same time very important for potential buyers when deciding.
yrhouse	+/-	The newer the house, the more expensive is in comparison with older houses. Directly proportional to the pricesold.
garages	+/-	The more garages, the expensive the house is. No parking garage could decrease its price.
lotsqft	+	Typically, houses with lotsqft that are larger than square footage will have extra space such as gardens, gazebos, and green areas, so the price will be higher.

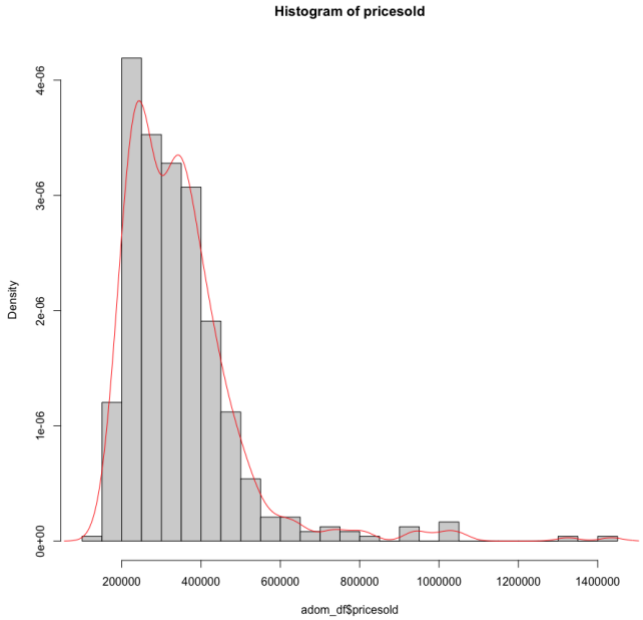
Other variables appear to be unrelated with the pricesold. See “Determining Importance of variables using ‘caret’ package” at appendix section.

Descriptive analysis of adom.



Analysis: Adom values near zero are more commonly observed in the histogram, which shows a non-normally distributed dependent variable and right-skewed distribution.

Descriptive analysis of price sold.

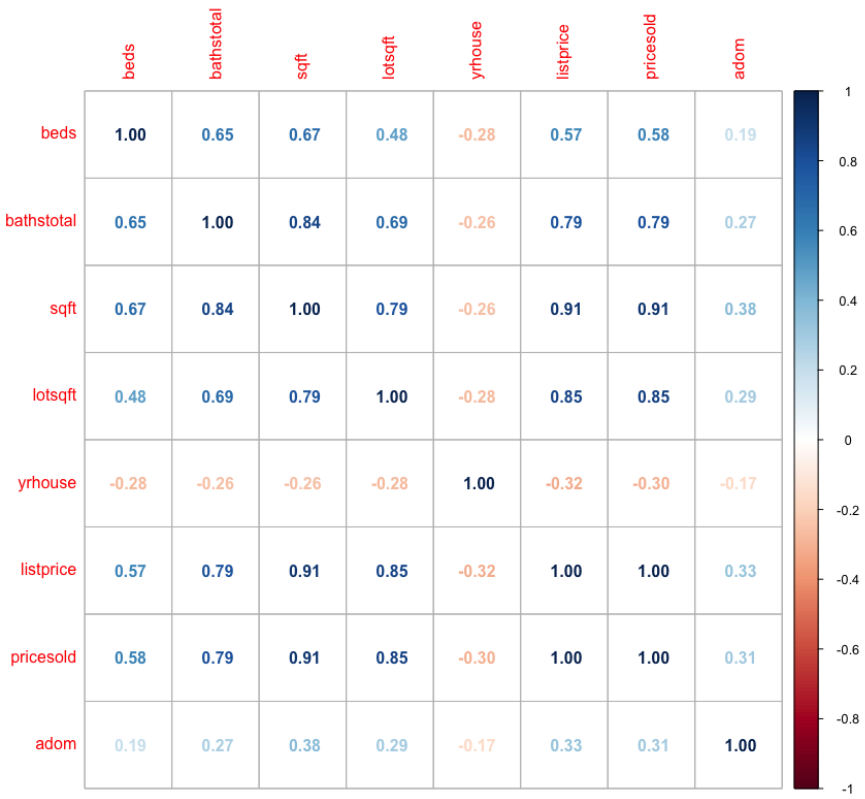


Analysis: The histogram tends to be bimodal, normally distributed, and right-skewed, however, skewness can be reduced by eliminating outliers.

Analysis of correlation

This graph shows a correlation matrix between the variables selected for "adom" and “pricesold” prediction, excluding “garages” since this is a variable with no continuous data.

In this case, the list price cannot be used as a predictor of the sold price since there is a correlation of 1 between them. We can also observe that there is a negative correlation between the created variable “yrhouse” which tells us that the newer the house (the fewer years have), the greater the price sold or the list price.
A negative correlation can also be observed between the created variable "yrhouse" which tells us that the newer the home (the fewer years it has), the higher the price sold or the list price.



	beds	bathstotal	sqft	lotsqft	yrhouse	listprice	pricesold	adom
beds	1.00	0.65	0.67	0.48	-0.28	0.57	0.58	0.19
bathstotal	0.65	1.00	0.84	0.69	-0.26	0.79	0.79	0.27
sqft	0.67	0.84	1.00	0.79	-0.26	0.91	0.91	0.38
lotsqft	0.48	0.69	0.79	1.00	-0.28	0.85	0.85	0.29
yrhouse	-0.28	-0.26	-0.26	-0.28	1.00	-0.32	-0.30	-0.17
listprice	0.57	0.79	0.91	0.85	-0.32	1.00	1.00	0.33
pricesold	0.58	0.79	0.91	0.85	-0.30	1.00	1.00	0.31
adom	0.19	0.27	0.38	0.29	-0.17	0.33	0.31	1.00

Run a set of three reasonable models for each DV. Copy and paste the R code for the three models and the combined output using stargazer. (3 points)

Models for agent days on market:

```
ols1 <- lm(adom ~ sqft+pricesold+listprice+yrhouse+garages,data=adom_df)
ols2 <- lm((adom) ~ sqft*pricesold*listprice,data=adom_df)
ols3 <- lm((adom^6) ~ sqft*pricesold*listprice,data=adom_df)
```

OLS Analysis of adom

	Dependent variable:		
	adom	(adom)	(adom3)
	(1)	(2)	(3)
sqft	0.052***	0.078***	17,666.940***
	(0.010)	(0.019)	(2,942.308)
pricesold	-0.001***	-0.004***	-861.947***
	(0.0002)	(0.001)	(106.247)
listprice	0.001***	0.003***	1,010.094***
	(0.0002)	(0.001)	(121.034)
yrhouse	-2.074		
	(1.369)		
garages	-13.912*		
	(7.923)		
sqft:pricesold		0.00000***	0.281***
		(0.00000)	(0.029)
sqft:listprice		-0.00000***	-0.344***
		(0.00000)	(0.031)
pricesold:listprice		-0.000	-0.0002***
		(0.000)	(0.00004)
sqft:pricesold:listprice		0.000***	0.00000***
		(0.000)	(0.000)
Constant	60.113	-33.536	-36,434,656.000***
	(44.750)	(38.471)	(5,877,151.000)
Observations	478	482	482
R^2	0.219	0.307	0.646
Adjusted R^2	0.211	0.297	0.641
Residual Std. Error	71.418 (df = 472)	67.418 (df = 474)	10,299,297.000 (df = 474)
F Statistic	26.446*** (df = 5; 472)	29.993*** (df = 7; 474)	123.758*** (df = 7; 474)
Note:			*p<0.1; **p<0.05; ***p<0.01

Models for sales price:

ols1 <- lm(pricesold ~ bathstotal+sqft+yrhouse+garages+lotsqft,data=adom_df)

ols2 <- lm((pricesold) ~ bathstotal*sqft*yrhouse*lotsqft,data=adom_df)

ols3 <- lm(log(pricesold) ~ sqft*lotsqft+bathstotal,data=adom_df)

OLS Analysis of saleprice

	Dependent variable:		
	pricesold	(pricesold)	log(pricesold)
	(1)	(2)	(3)
bathstotal	11,156.480**	63,395.310	0.046***
	(5,490.126)	(179,548.700)	(0.014)
sqft	103.484***	72.198	0.0003***
	(6.635)	(177.046)	(0.00002)
yrhouse	-2,477.634**	26,446.260	
	(1,006.355)	(17,834.630)	
garages	10,473.630*		
	(5,924.630)		
lotsqft	7.954***	88.585**	0.00004***
	(0.643)	(34.494)	(0.00000)
bathstotal:sqft		41.631	
		(49.680)	
bathstotal:yrhouse		-4,563.395	
		(6,463.656)	
sqft:yrhouse		0.931	
		(6.449)	
bathstotal:lotsqft		-19.751**	
		(8.529)	
sqft:lotsqft		-0.002	-0.000***
		(0.007)	(0.000)
yrhouse:lotsqft		-3.376***	
		(1.286)	
bathstotal:sqft:yrhouse		-1.251	
		(1.823)	
bathstotal:sqft:lotsqft		0.001	
		(0.002)	
bathstotal:yrhouse:lotsqft		0.948***	
		(0.334)	
sqft:yrhouse:lotsqft		0.00003	
		(0.0003)	
bathstotal:sqft:yrhouse:lotsqft		-0.00003	
		(0.0001)	
Constant	-4,033.719	-608,591.200	11.384***
	(33,902.270)	(488,615.300)	(0.040)
Observations	478	482	482
R^2	0.878	0.898	0.845
Adjusted R^2	0.876	0.895	0.843
Residual Std. Error	53,547.050 (df = 472)	49,175.800 (df = 466)	0.141 (df = 477)
F Statistic	678.056*** (df = 5; 472)	274.970*** (df = 15; 466)	648.173*** (df = 4; 477)
Note:			*p<0.1; **p<0.05; ***p<0.01

2. Select the best model from each set and examine whether it meets the assumptions of the regression model. Which of the five regression assumptions are met for the final models? (2 points)

Assumption	Model: adom ols3	Model: saleprice ols2
Linearity	Yes	Yes

Normality	No. Based on Shapiro-Wilk's and Kolmogorov-Smirnov test, p-value<0.05 hence data is not normally distributed.	Despite Shapiro-Wilk's giving pval<0.05, the residuals plot seems to be normally distributed (see appendix for reference).
Homoskedasticity	No. p-value<0.05 thus, at least two population variances differ.	No. p-value<0.05 thus, at least two population variances differ.
Multicollinearity	Based on Durbin-Watson test, DW=2.3 clo se to 2; hence no autocorrelation	DW = 1.6885 shows no autocorrelation.
Independence	Yes	No

3. Using your best models, select the top three predictors of adom and pricesold, and explain their marginal effects on the dependent variables. Remember that significance is not important. (2 points)

The top 3 predictors for adom are sqft, price sold, and list price.

- From model ols3, sqft has a beta coefficient of 17,666 which means for 1000 increase in sqft, adom increases by 17,666.
- The price sold has a beta coefficient of -806 which means for 1000 decrease in price sold, adom decreases by -806.
- The list price has a beta coefficient of 1010, which means for 1000 increase in listprice, adom increases by 1010.

The top 3 predictors for saleprice are sqft, bathsfull and lotsqft.

- From model ols2, sqft has a beta coefficient of 72.1 which means for 1000 increase in sqft, price sold increase by 72.198
- From model ols2,bathsfull has a beta coefficient of 63,395 which means for 1000 increase in bathsfull, sale price increases 63,395.
- From model ols2,lotsqft has a beta coefficient of 88.5 which means for 1000 increase in lotsqft, sale price increases 88.5.

Note: By reviewing the beta coefficient in the present models, we lose the sense of interpretation due to the nonlinear interactions between variables.

Appendix – Markdown Format

Statistical Data Mining | Hunters Green Home Sales

Pablo X Zumba

Preprocessing and exploring data

```
rm(list=ls())
df=rio::import("HuntersGreenHomeSales.xlsx", sheet="Data")
colnames(df)=tolower(make.names(colnames(df)))
```

Cleaning, transforming and filtering data based on predicting “adom” variable.

```
#yrblt will be transformed into how many years the house is at then current year (2022)
df['yrhouse'] = 2022 - df$yrblt
#Replacing NA values on spa column.
df["spa"][is.na(df["spa"])] = FALSE
#Changing column names on adom and cdom
names(df)[names(df) == 'adom_agentdaysonmarket'] = 'adom'
names(df)[names(df) == 'cdom_cumuldaysmls'] = 'cdom'
#Checking Correlation between adom and cdom
cor(df$adom,df$cdom)#0.8684742 adom and cdom are very correlated.
```

```
## [1] 0.8684742
```

```
#Based on the correlation that exists between adom & cdom (0.87), I decided to disregard
cdom since it will introduce bias. From a business perspective and the concept and adom a
nd cdom. The value of adom is always going to be less than cdom and since we're trying to
predict adom, it does not make sense to me to take into account a variable that is always
going to be greater and highly correlated.
#Changing name on lppersqft & sppersqft
names(df)[names(df) == 'lppersqft'] = 'listprice_psqft'
names(df)[names(df) == 'sppersqft'] = 'pricesold_psqft'
#Checking if the house was sold in a price greater than listed.
list_sold_price=ifelse(df$pricesold>df$listprice,"Yes","No")
table(list_sold_price)#There are 42 houses that were sold in a price greater than listed.
```

```
## list_sold_price
##  No Yes
## 440  42
```

```
#Checking if the house was sold in a price greater than listed per square foot.
list_sold_price_psqft=ifelse(df$pricesold_psqft>df$listprice_psqft,"Yes","No")
table(list_sold_price_psqft)#Same results as the previous. 42 houses.
```

```
## list_sold_price_psqft
##  No Yes
## 440  42
```

```
adom_df = df[, c('beds','bathstotal','sqft','garages','roof','lotsqft','yrhouse','pool','
spa','listprice','pricesold','adom')]
```

```
#We ended up having 11 variables/columns and need to convert categorical into factor vari
able.
adom_df$roof=as.factor(adom_df$roof)
adom_df$pool=as.factor(adom_df$pool)
adom_df$spa=as.factor(adom_df$spa)
#Only have 3 categorical variables: roof, pool and spa.
attach(adom_df)
str(adom_df)

#View(adom_df)
```

Descriptive analysis on adom

```
#hist(adom)
den <- density(adom_df$adom)                          # Density function
#plot(den, main="Kernel Density of adom", col="red")
hist(adom_df$adom, breaks=20, prob=T, main="Histogram of adom") #different knots to make
it moreprecise
lines(den, col="red")
```

Descriptive analysis on pricesold

```
den <- density(adom_df$pricesold)                          # Density function
#plot(den, main="Kernel Density of adom", col="red")
hist(adom_df$pricesold, breaks=20, prob=T, main="Histogram of pricesold") #different knot
s to make it moreprecise
lines(den, col="red")
```

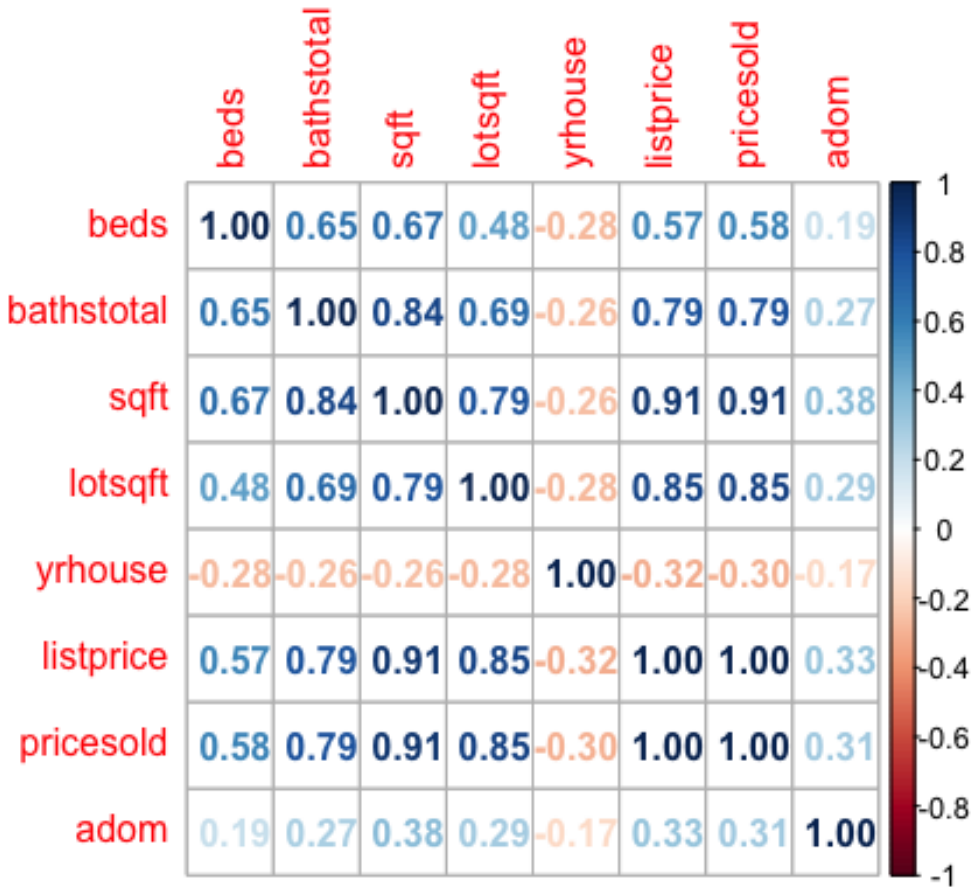
Descriptive analysis on interactions between independent variables and adom

```
#plot(adom ~ beds, data=adom_df)# 4 and 5 beds have the highest adom.
#plot(adom ~ bathstotal, data=adom_df)#2 and 3 baths total are the most tend to have the
lowes adom
#plot(adom ~ sqft, data=adom_df)#The sqft between 1000-400 has a low adom. Based on resea
rch, this variable seems to ve very relecant for adom.
#plot(adom ~ garages, data=adom_df)#2 and 3 garages has more adom observations.
#plot(adom ~ roof, data=adom_df)#Shingle and Tile are more related to adom.
#plot(adom ~ lotsqft, data=adom_df)#The lotsqft between 200-1200 has a low adom
#plot(adom ~ yrhouse, data=adom_df)#Hard to see a relationship.
#plot(adom ~ pool, data=adom_df)#Based on the median of the box plot, all types of pool a
ppear to have a similar adom, so we can get rid of it.
#plot(adom ~ spa, data=adom_df)#There is no significant difference between adom when the
house has spa or does not.
#plot(adom ~ sale_ratio, data=adom_df)#The majority of the houses have a sales ratio betw
een 0.95 and 1, which means they sold faster when the sold price was slightly lower than
the list price.
```

Analyzing correlation between variables

```
#plot(adom_df[,c(1,2,3,4,5,6,7,8,9,10,11)],pch=19,main="Continuous Variables only")
#Garages can be consider as a categorical variable so it's not at the corrplot.
judge_cor = round(cor(adom_df[,c(1,2,3,6,7,10,11,12)]),12)
library(corrplot)

## corrplot 0.92 loaded
```

```
corrplot(judge_cor,method="number")
```



- There is multicollinearity “correlation between predictor/independent variables” on variables that are related to space such as: sqft, lotsqft, and number of baths and beds, which make sense. - There is some positive correlation between the dependent variable “adom” with bathstotal, sqft, and lotsqft and negative correlation with sale_ratio. - The negative correlation with sale_ratio could be explained because it’s more likely to sold a house when the listing price is reduced. - Based on the correlation graph, we still need to get rid of some columns, speccially the ones that are related to space. - yearhouse does not seem to be relevant any more. - We can combine beds, bathstotal and garages to get a total number useful spaces in the house. The variable is going to be called: bed_bad_gar - We can also combine the sqft & lotsqft variables to get a sqft_ratio.

```
#Trying a different approach to eliminate more variables:
#adom_df['bed_bad_gar'] = beds+bathstotal+garages
#adom_df['sqft_ratio'] = lotsqft / sqft
#adom_df = adom_df[, c('bed_bad_gar','roof','sqft_ratio','sale_ratio','adom')]
#attach(adom_df)
#str(adom_df)
```

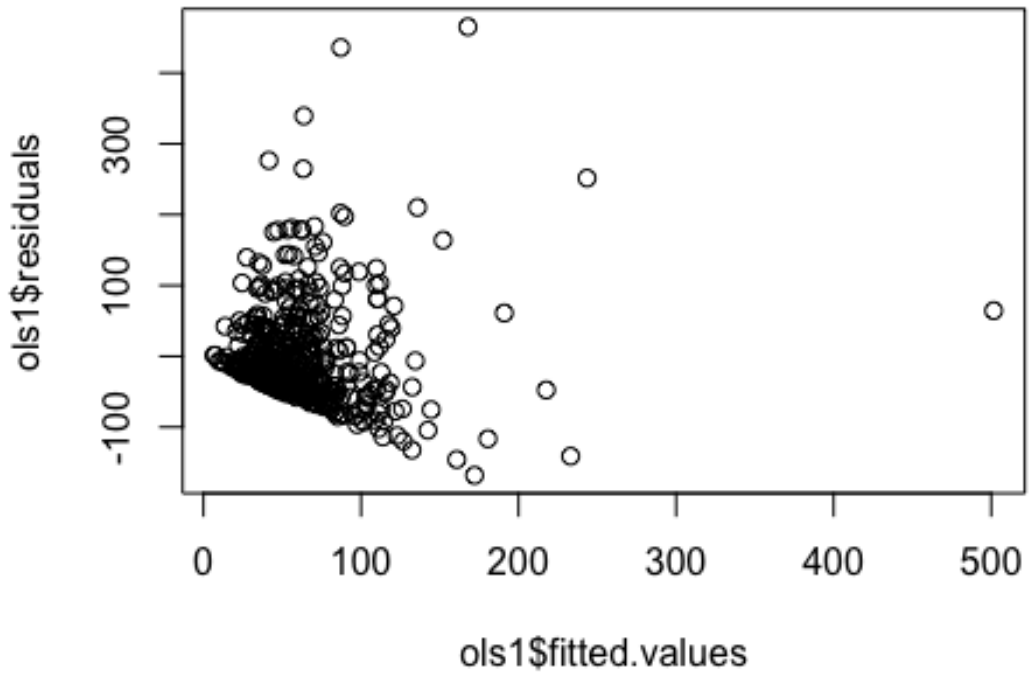
#‘adom’ Models

##OLS Estimation on “adom” **First Model ols1**

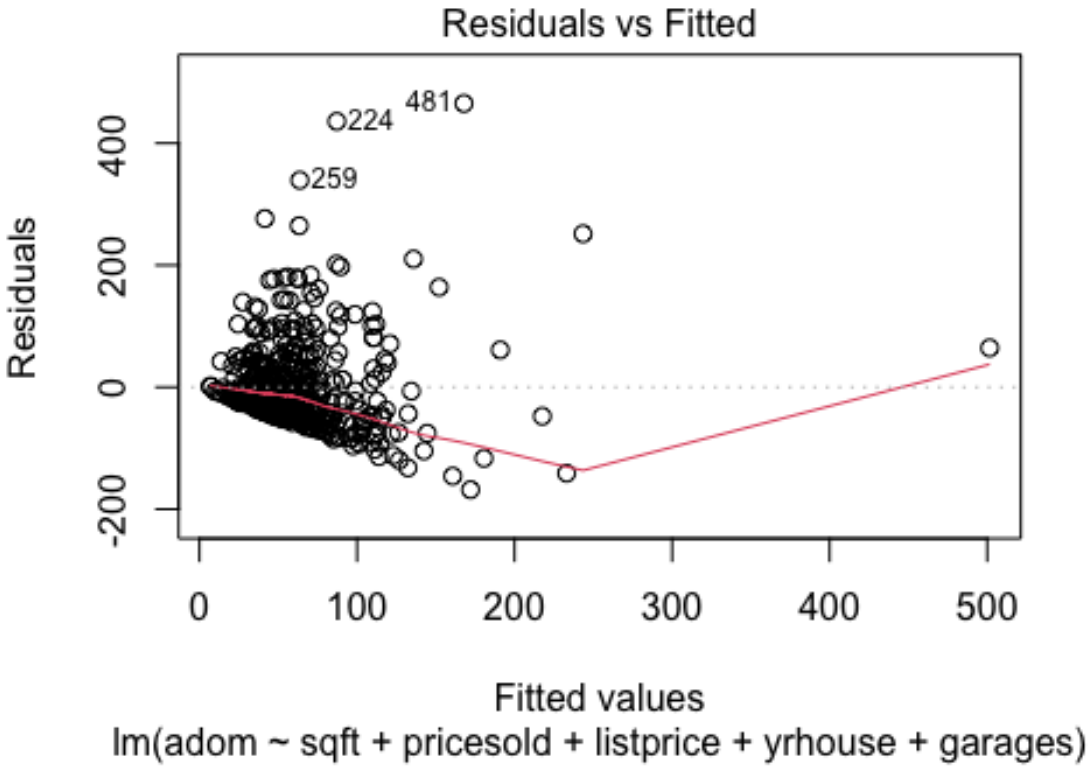
```
ols1 <- lm(adom ~ sqft+pricesold+listprice+yrhouse+garages,data=adom_df)
summary(ols1)
```

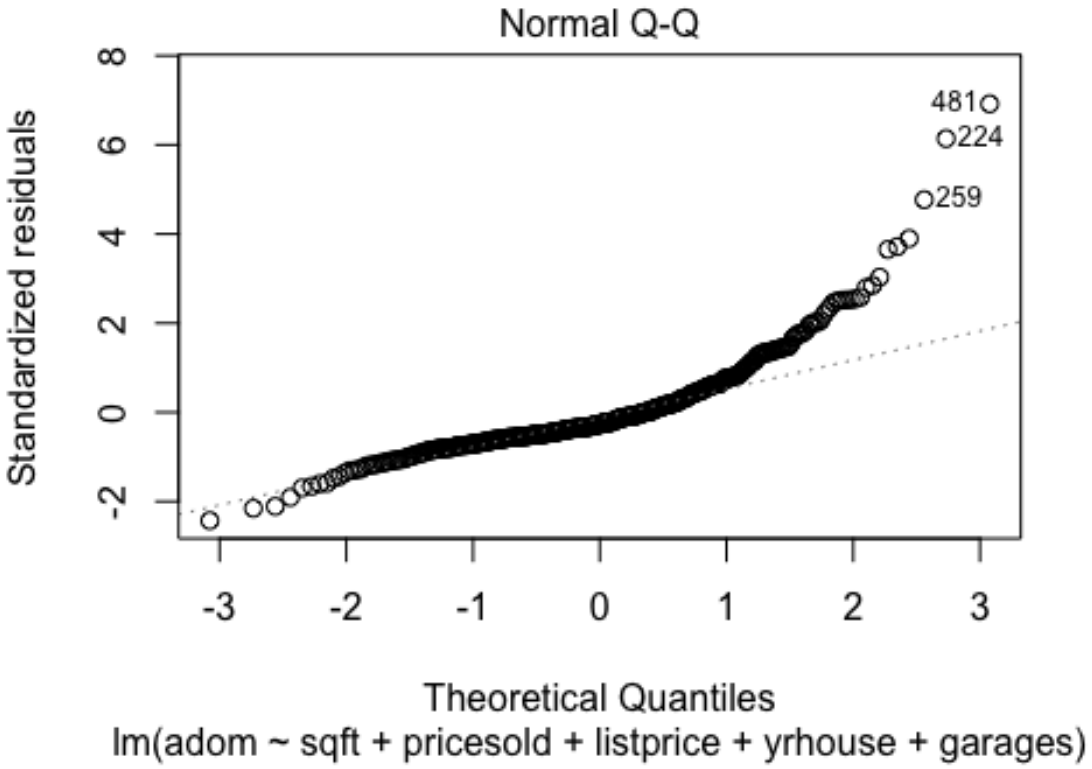
```
##
## Call:
## lm(formula = adom ~ sqft + pricesold + listprice + yrhouse +
##     garages, data = adom_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -168.07  -40.28  -19.20   22.18  465.14
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.011e+01  4.475e+01   1.343   0.1798
## sqft         5.182e-02  9.761e-03   5.308 1.71e-07 ***
## pricesold   -1.412e-03  2.437e-04  -5.796 1.25e-08 ***
## listprice    1.252e-03  2.270e-04   5.514 5.79e-08 ***
## yrhouse     -2.074e+00  1.369e+00  -1.515   0.1304
## garages     -1.391e+01  7.923e+00  -1.756   0.0797 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.42 on 472 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2106
## F-statistic: 26.45 on 5 and 472 DF,  p-value: < 2.2e-16
```

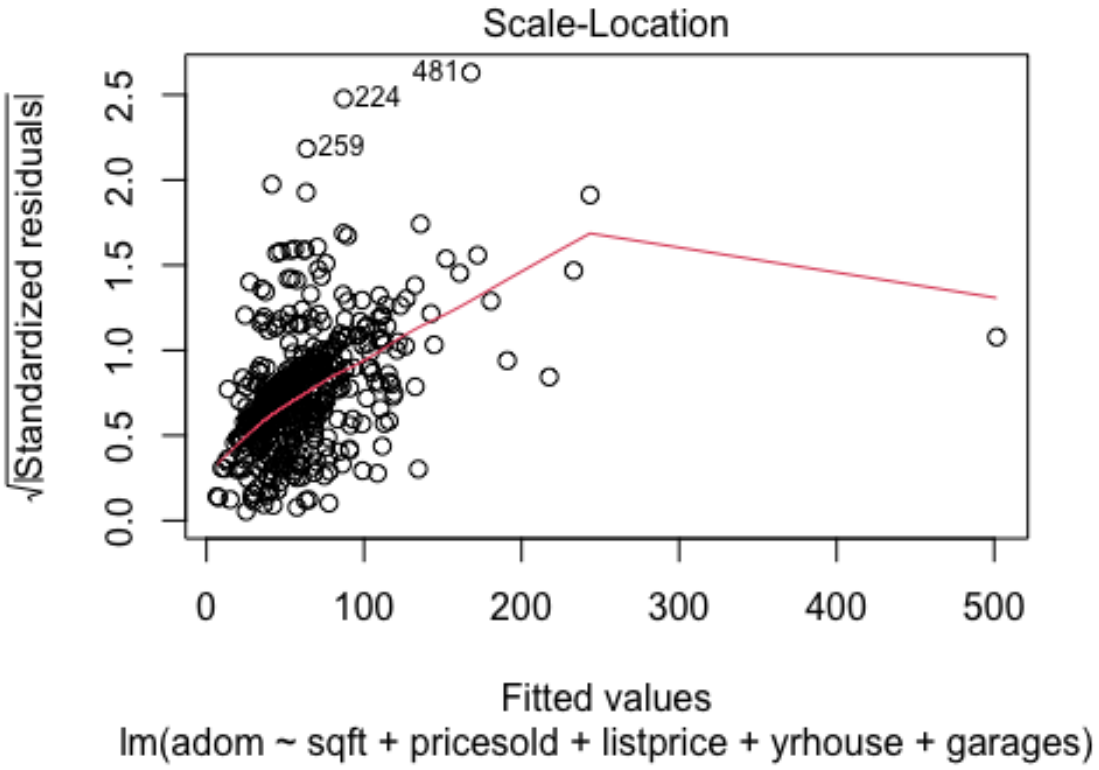
```
plot(ols1$residuals~ols1$fitted.values)
```

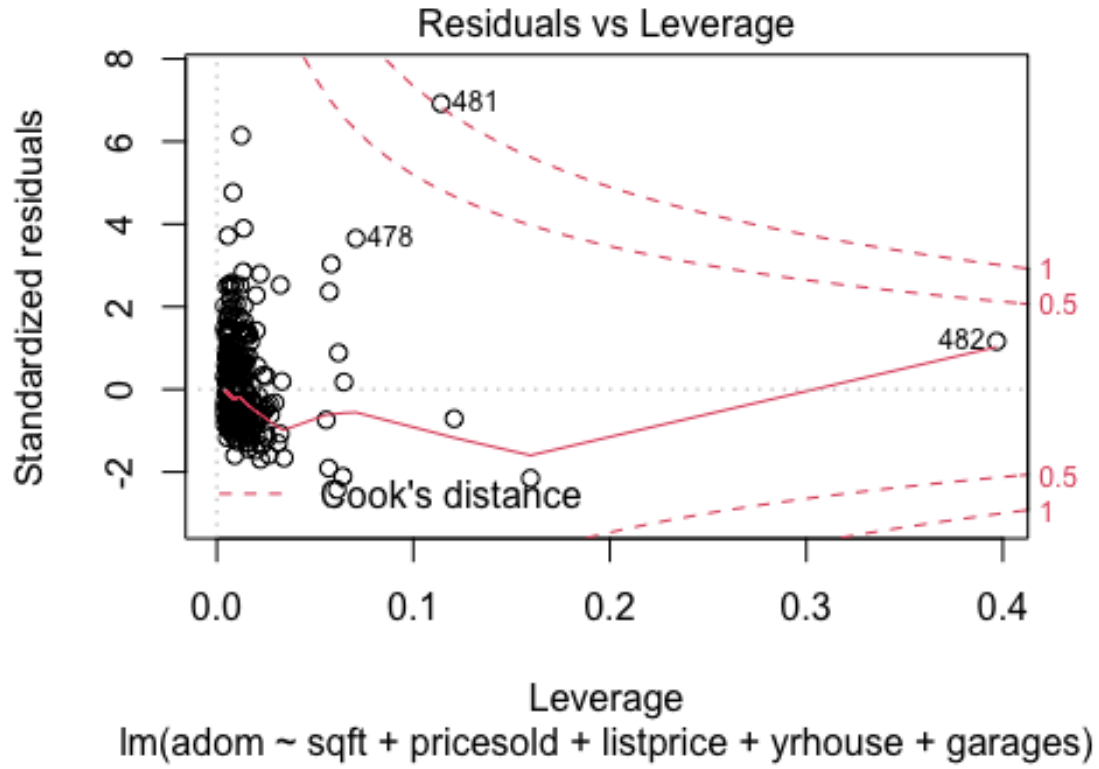


```
plot(ols1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(adom ~ sqft + pricesold + listprice + yrhouse + garages)
481
224
259
-200
0
200
400
0
100
200
300
400
500


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(adom ~ sqft + pricesold + listprice + yrhouse + garages)
481
224
259
-2
0
2
4
6
8
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
481
224
259
0.0
0.5
1.0
1.5
2.0
2.5
0
100
200
300
400
500
Fitted values
lm(adom ~ sqft + pricesold + listprice + yrhouse + garages)


Residuals vs Leverage
Standardized residuals
481
478
482
Cook's distance
1
0.5
0.5
1
-2
0
2
4
6
8
0.0
0.1
0.2
0.3
0.4
Leverage
lm(adom ~ sqft + pricesold + listprice + yrhouse + garages)

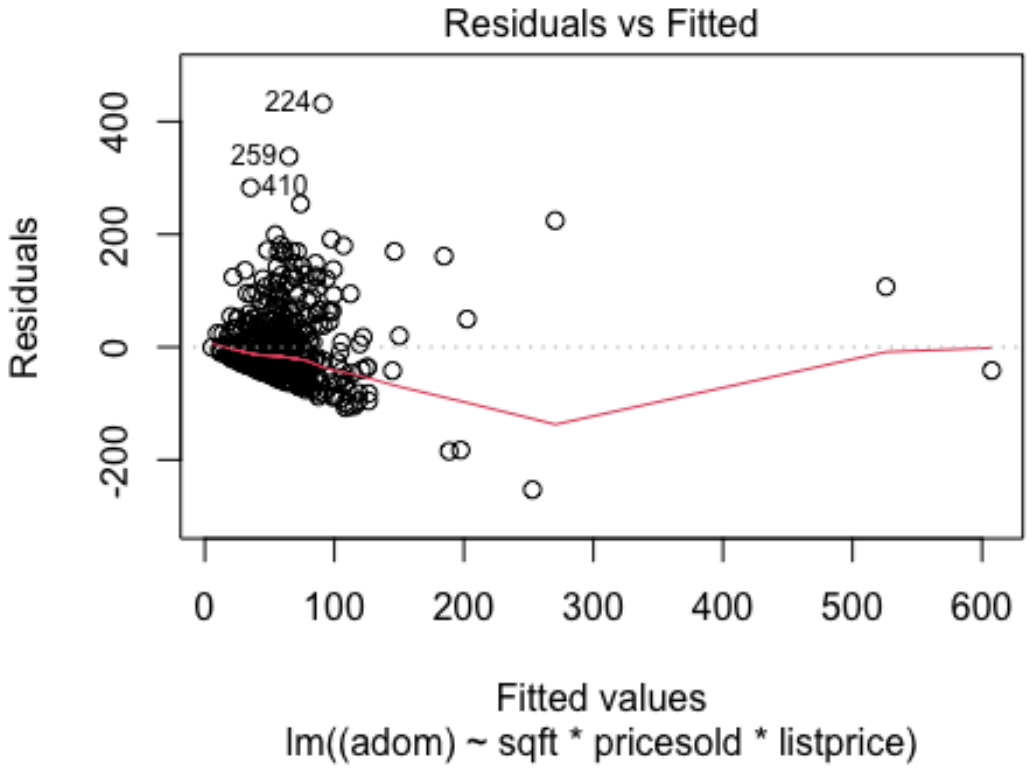
#Determining Importance of variables using “caret” package.

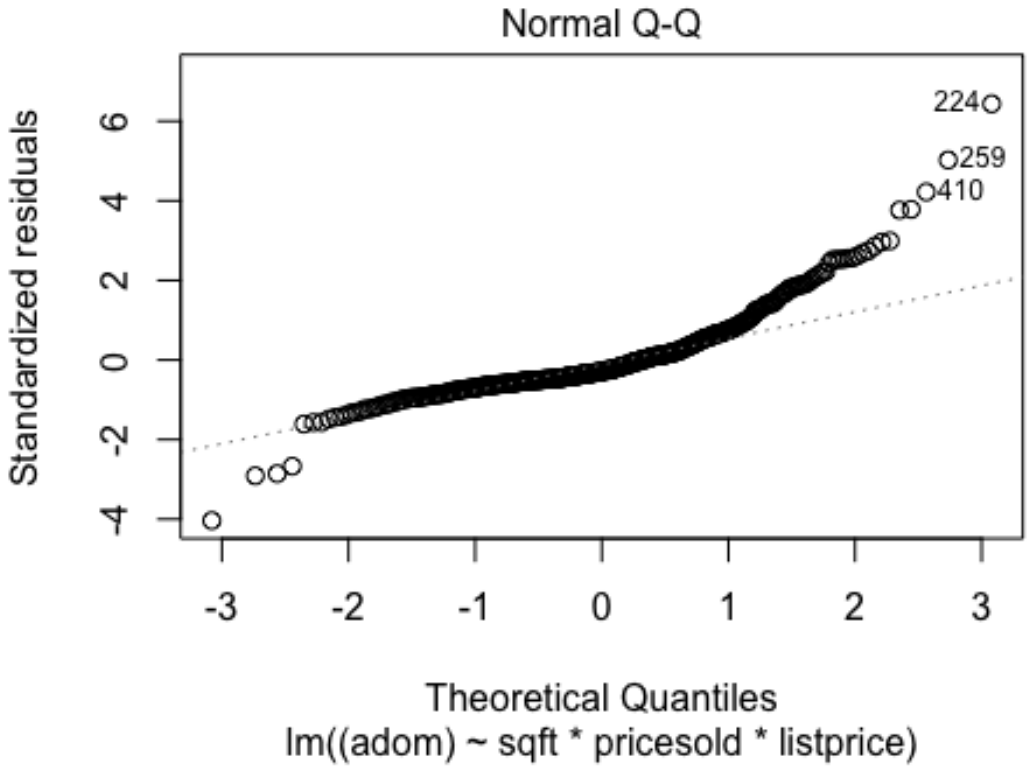
```
library(caret)

## Warning: package 'caret' was built under R version 4.1.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.1.2

## Loading required package: lattice

ols1Imp = varImp(ols1, scale=FALSE)
ols1Imp
```

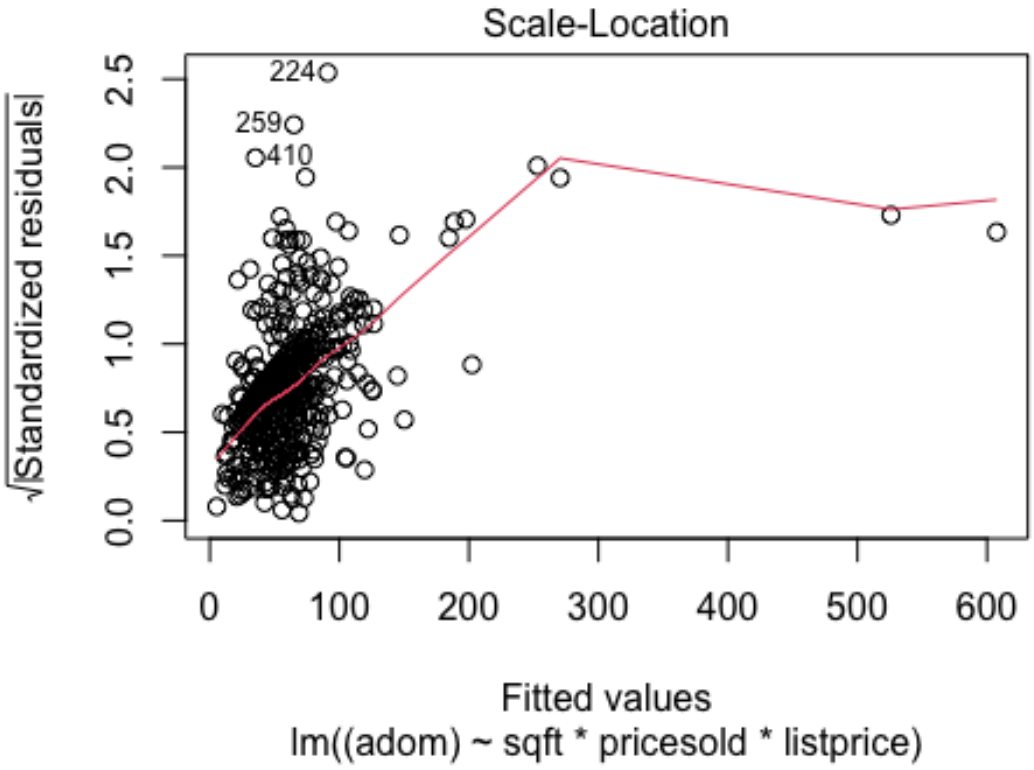
Based on the table above, the top 5 important variables are: sqft, pricesold, listprice, yrhouse, and garages so we will use those in the new model.

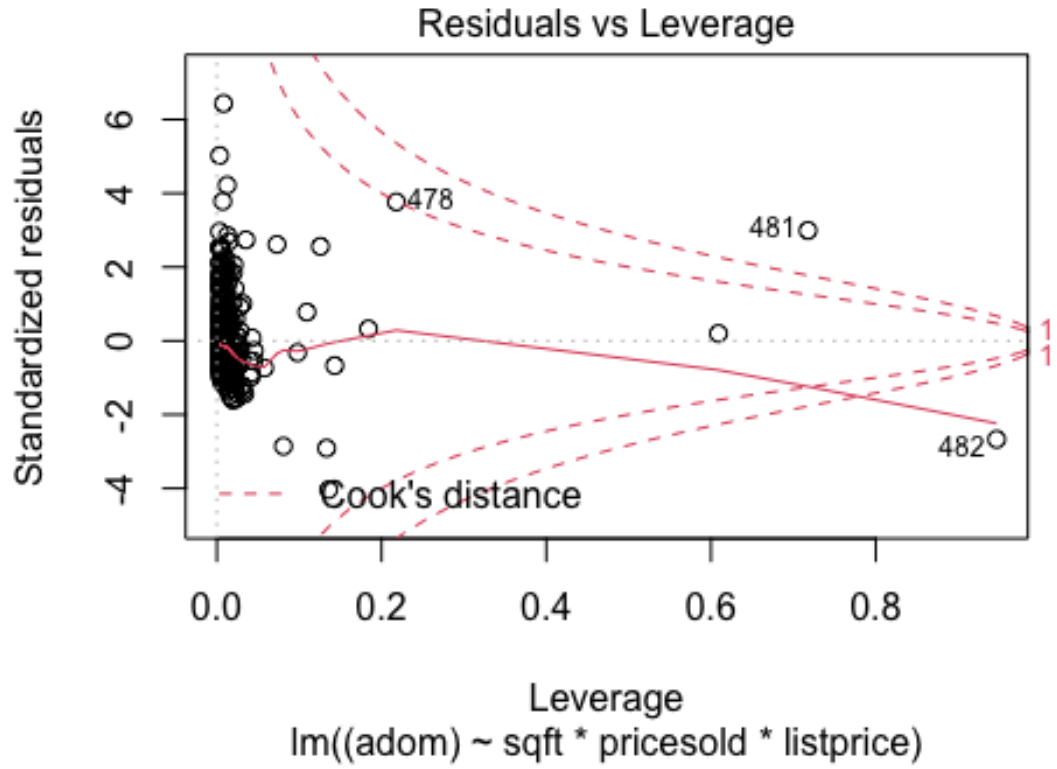
Second Model using most important variables ols2

```
ols2 <- lm((adom) ~ sqft*pricesold*listprice,data=adom_df)
summary(ols2)

##
## Call:
## lm(formula = (adom) ~ sqft * pricesold * listprice, data = adom_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -252.90  -37.94  -19.80   21.94  432.03
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.354e+01  3.847e+01  -0.872 0.383795
## sqft                      7.777e-02  1.926e-02   4.038 6.28e-05 ***
## pricesold                -3.559e-03  6.955e-04  -5.117 4.52e-07 ***
## listprice                 3.437e-03  7.923e-04   4.338 1.76e-05 ***
## sqft:pricesold            8.210e-07  1.901e-07   4.319 1.91e-05 ***
## sqft:listprice           -9.138e-07  2.050e-07  -4.458 1.03e-05 ***
## pricesold:listprice      -4.207e-10  2.753e-10  -1.528 0.127139
## sqft:pricesold:listprice  1.667e-13  4.370e-14   3.816 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.42 on 474 degrees of freedom
## Multiple R-squared:  0.307,  Adjusted R-squared:  0.2967
## F-statistic: 29.99 on 7 and 474 DF,  p-value: < 2.2e-16

plot(ols2)
```


Residuals vs Fitted
Residuals
Fitted values
lm((adom) ~ sqft * pricesold * listprice)
224
259
410


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm((adom) ~ sqft * pricesold * listprice)
224
259
410



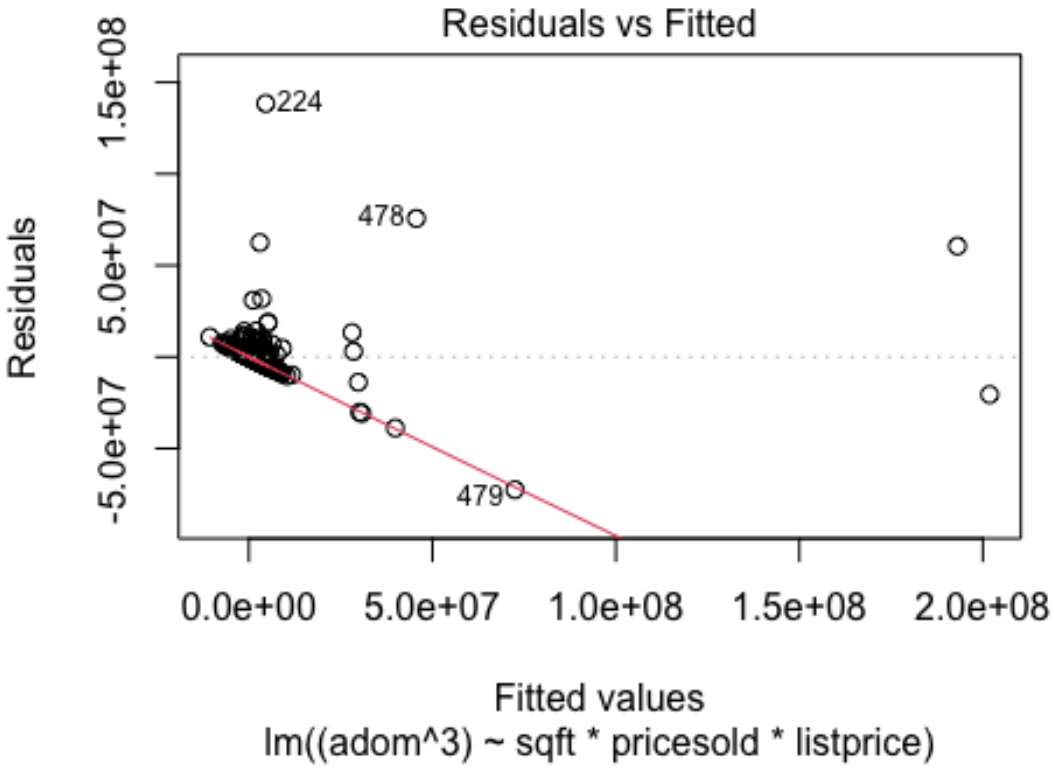


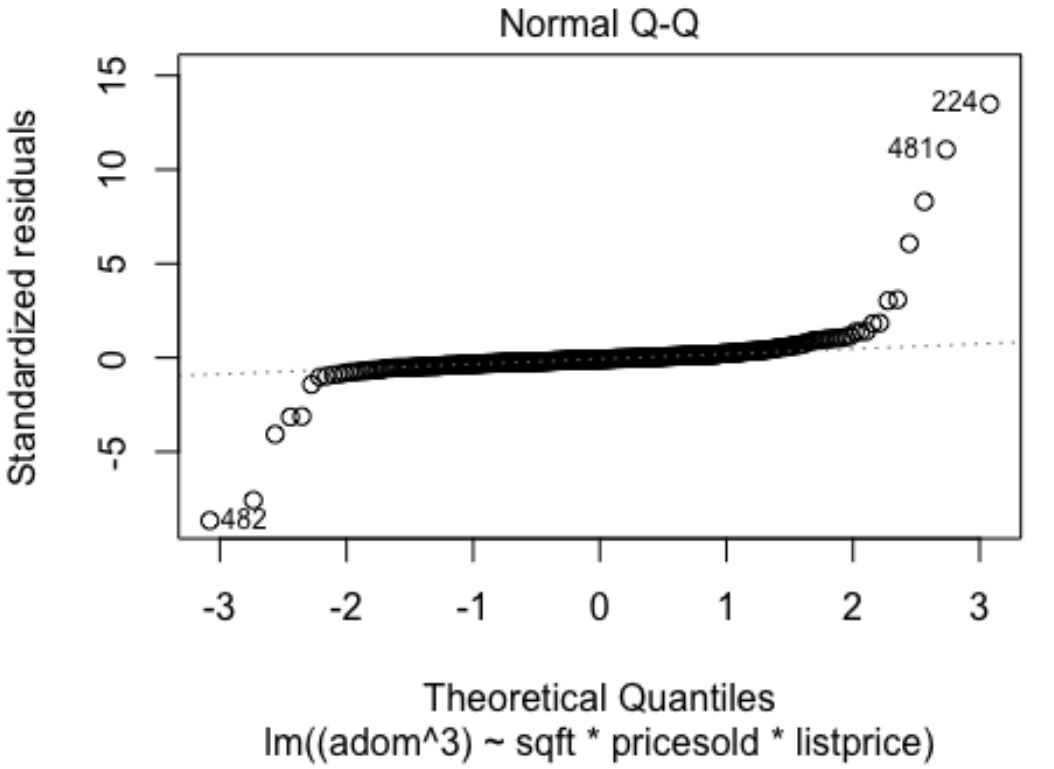
Third Model using most important variables ols3

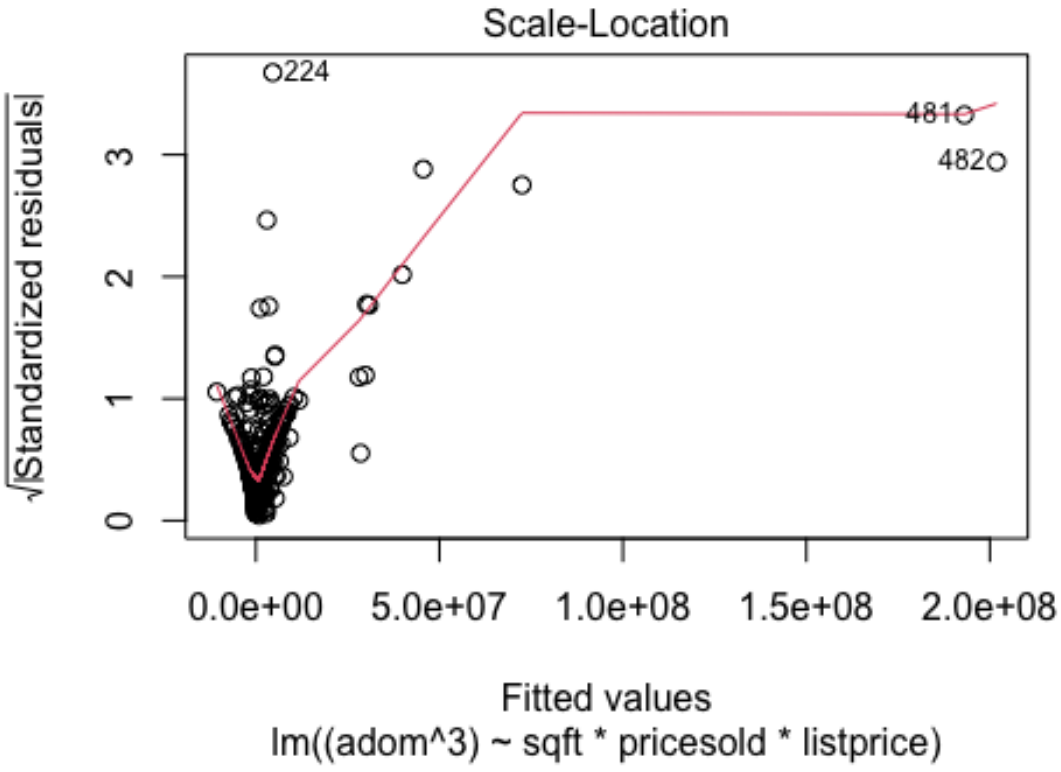
```r
ols3 <- lm((adom^3) ~ sqft*pricesold*listprice,data=adom_df)
summary(ols3)
```

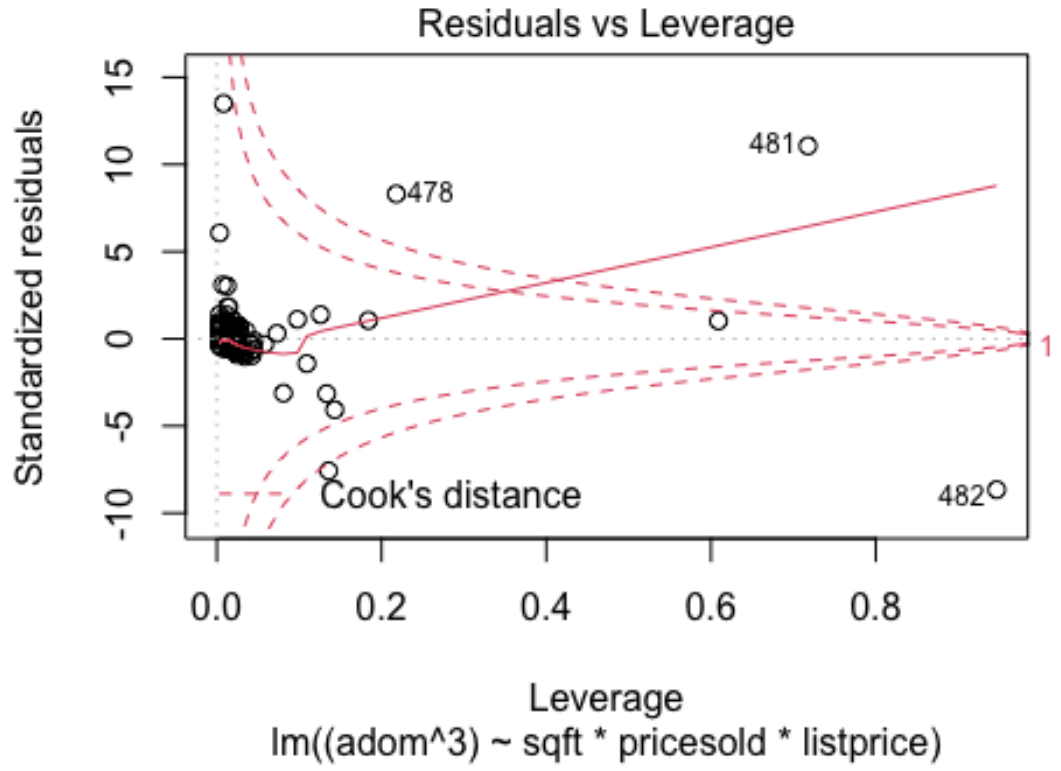
```
##
## Call:
## lm(formula = (adom^3) ~ sqft * pricesold * listprice, data = adom_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -72505614  -2558285   -803695   1164269 138445136
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -3.643e+07  5.877e+06  -6.199 1.23e-09 ***
## sqft                       1.767e+04  2.942e+03   6.004 3.82e-09 ***
## pricesold                 -8.619e+02  1.062e+02  -8.113 4.28e-15 ***
## listprice                  1.010e+03  1.210e+02   8.346 7.82e-16 ***
## sqft:pricesold             2.812e-01  2.904e-02   9.684  < 2e-16 ***
## sqft:listprice            -3.438e-01  3.131e-02 -10.979  < 2e-16 ***
## pricesold:listprice       -2.451e-04  4.206e-05  -5.827 1.04e-08 ***
## sqft:pricesold:listprice   8.398e-08  6.676e-09  12.579  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10300000 on 474 degrees of freedom
## Multiple R-squared:  0.6463, Adjusted R-squared:  0.6411
## F-statistic: 123.8 on 7 and 474 DF,  p-value: < 2.2e-16
```

```r
plot(ols3)
```

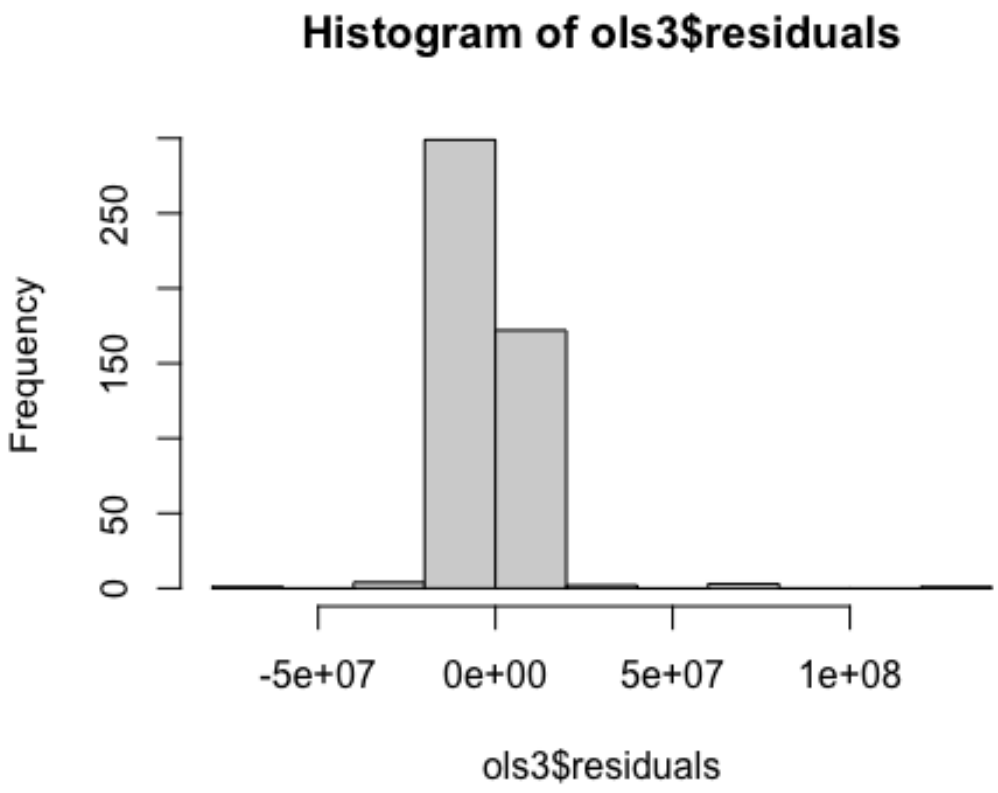

Residuals vs Fitted
Residuals
Fitted values
lm((adom^3) ~ sqft * pricesold * listprice)
224
478
479
-5.0e+07
5.0e+07
1.5e+08
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm((adom^3) ~ sqft * pricesold * listprice)
224
481
482
-5
0
5
10
15
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
224
481
482
0
1
2
3
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Fitted values
lm((adom^3) ~ sqft * pricesold * listprice)


Residuals vs Leverage
Standardized residuals
481
478
482
Cook's distance
1
-10
-5
0
5
10
15
0.0
0.2
0.4
0.6
0.8
Leverage
lm((adom^3) ~ sqft * pricesold * listprice)

```
hist(ols3$residuals)
```



#Testing Assumptions **Formal test for Normality**

```
# Shapiro-Wilk's test of multivariate normality for small samples (n<2000)
shapiro.test(ols3$res) #p-value<0.05 thus data is not normally distributed

##
##  Shapiro-Wilk normality test
##
## data:  ols3$res
## W = 0.42207, p-value < 2.2e-16

# Kolmogorov-Smirnov test:
norm <- rnorm(200) #bencharm sample of two hund.
ks.test(norm, ols3$res)#p-value<0.05 thus data is not normally distributed

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  norm and ols3$res
## D = 0.63071, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

- The data is not normally distributed.

Bartlett's test of heterokedasticity

```
library("car")

## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.2

bartlett.test(list(ols3$res, ols3$fit))#p-value<0.05 thus, at least two population varian
ces differ

##
##  Bartlett test of homogeneity of variances
##
## data:  list(ols3$res, ols3$fit)
## Bartlett's K-squared = 43.036, df = 1, p-value = 5.374e-11
```

Levene's test of homoskedasticity

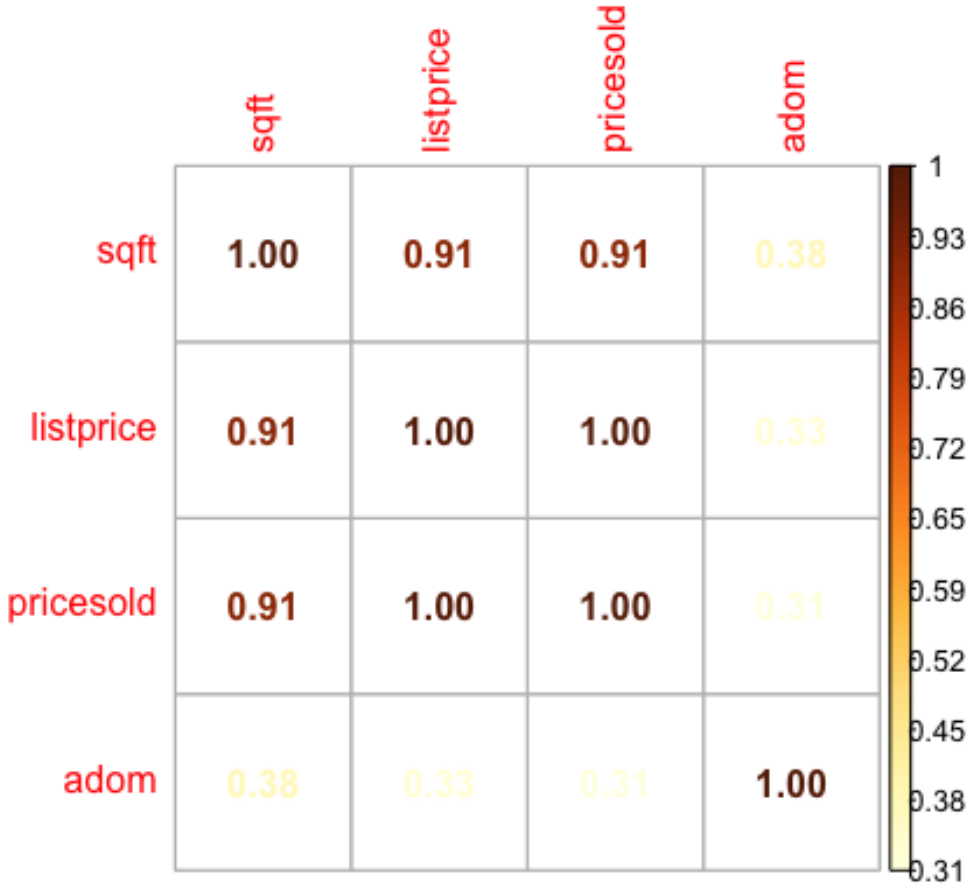
```
#levene.df = data.frame(ols3$residuals, ols3$fitted.values)
#leveneTest(ols3.residuals~ols3.fitted.values,data=levene.df,center=median)
#Levene's test is not appropriate with quantitative explanatory variables.
```

Durbin-Watson test of autocorrelation

```
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.1.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

#d ~ [0, 4]; values around 2 (i.e., 1.5 to 2.5) suggests no autocorrelation
dwtest(ols3)#DW close to 2; hence no autocorrelation

##
##  Durbin-Watson test
##
## data:  ols3
## DW = 2.3759, p-value = 1
## alternative hypothesis: true autocorrelation is greater than 0
```

#Checkind independence in model 3. sqft,pricesold,listprice

```
#plot(adom_df[,c(1,2,3,4,5,6,7,8,9,10,11)],pch=19,main="Continuous Variables only")
#Garages can be consider as a categorical variable so it's not at the corrplot.
independence = round(cor(adom_df[,c(3,10,11,12)]),12)
library(corrplot)
corrplot(independence,method="number",is.corr = F)
```



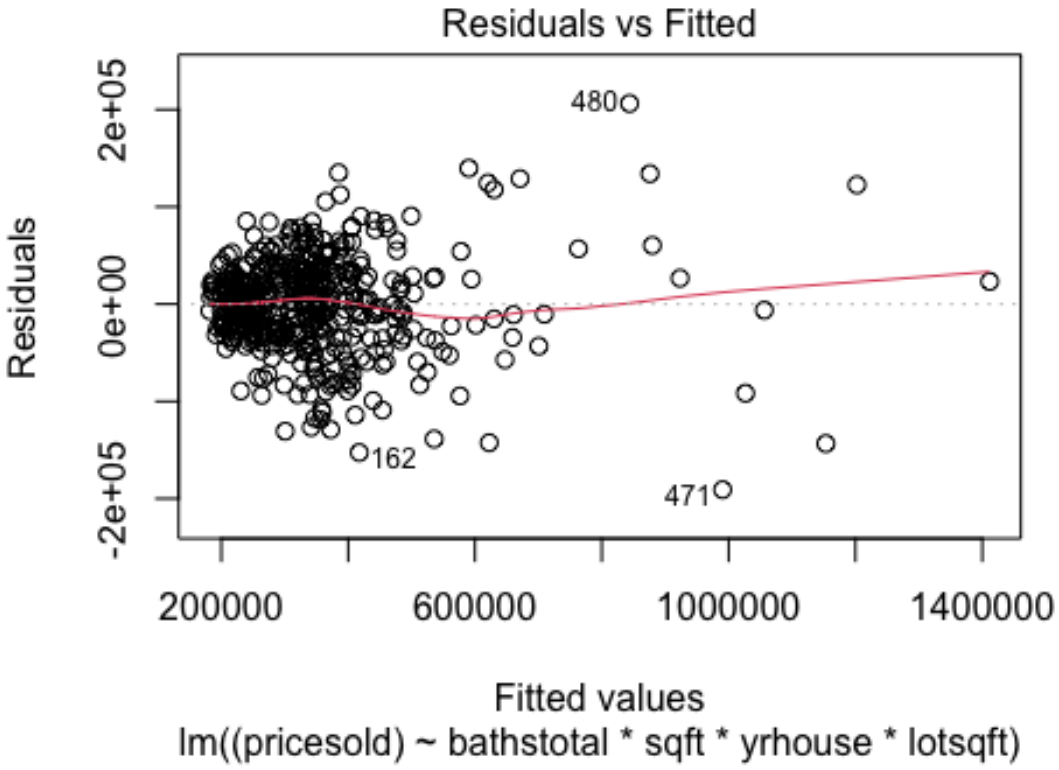
Model 2 for ols2 on predicting price sold.

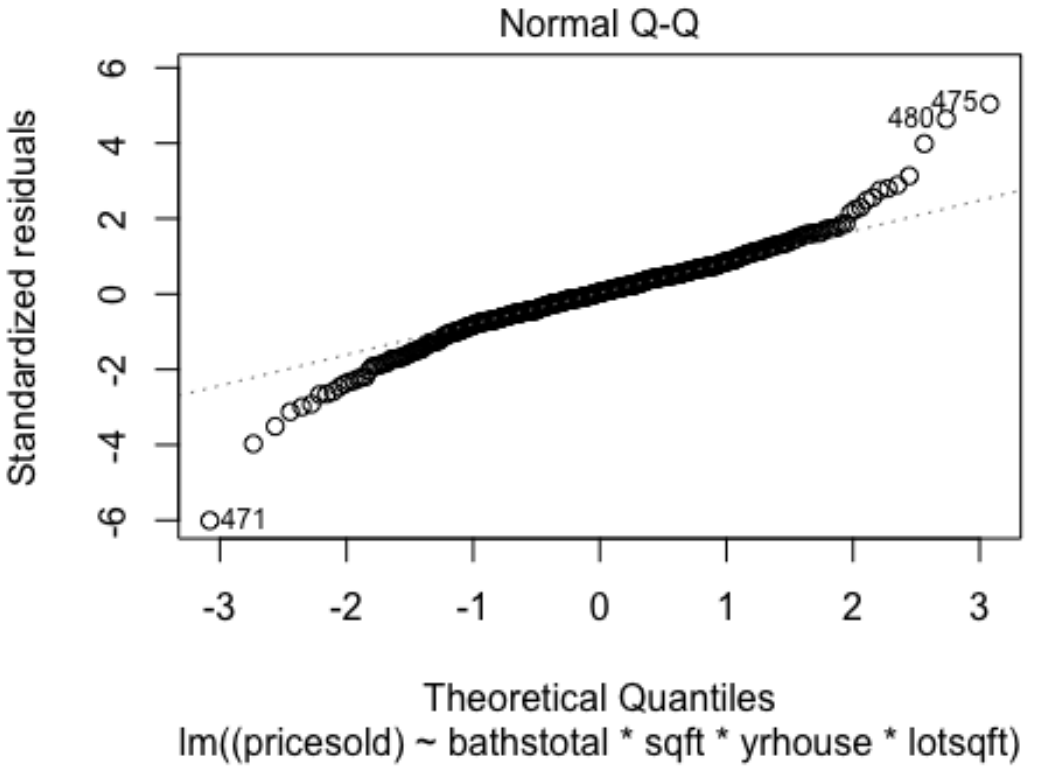
Second Model using most important variables ols2

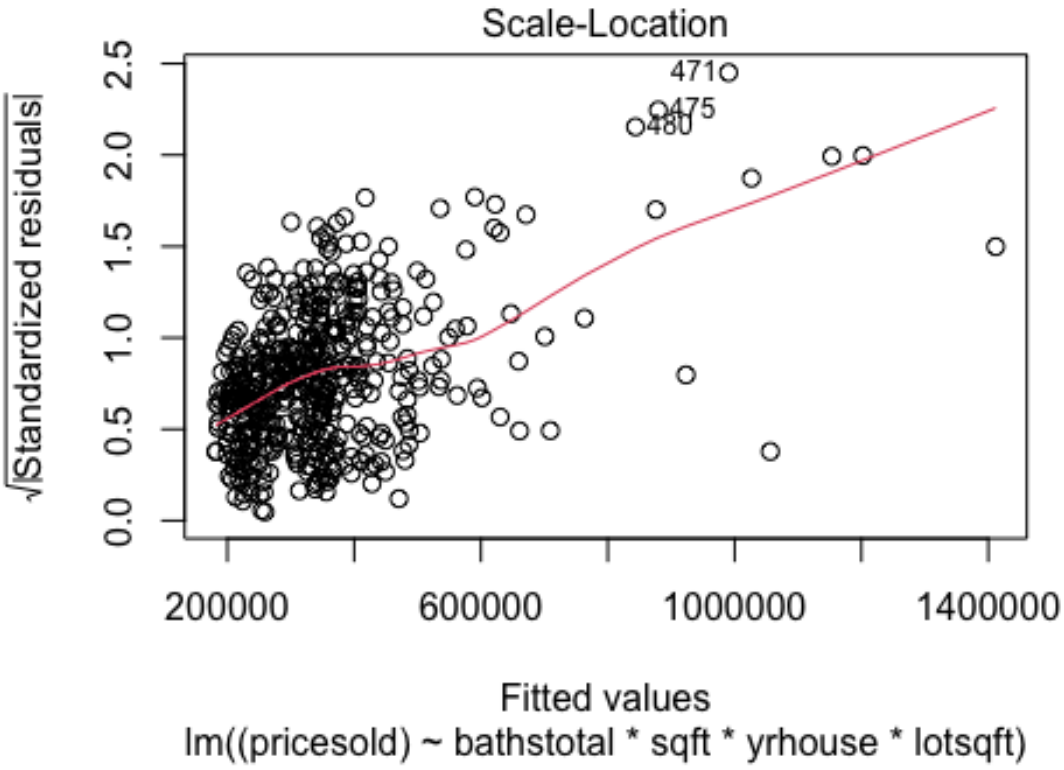
```
ols2 <- lm((pricesold) ~ bathstotal*sqft*yrhouse*lotsqft,data=adom_df)
summary(ols2)
```

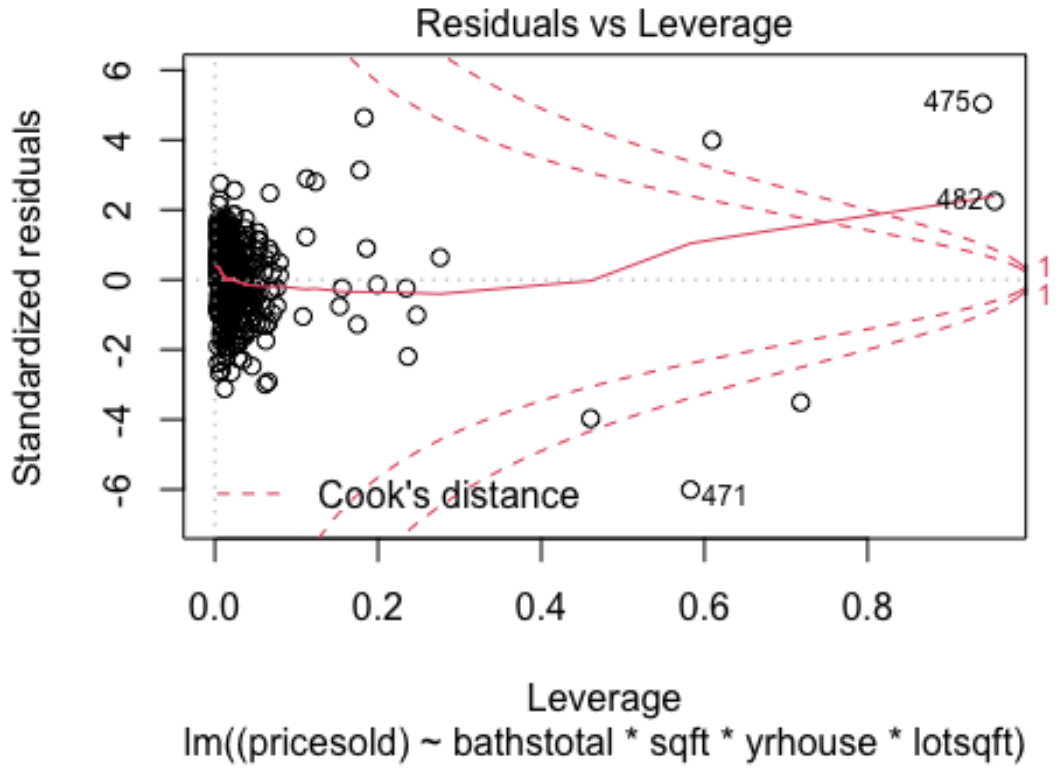
```
##
## Call:
## lm(formula = (pricesold) ~ bathstotal * sqft * yrhouse * lotsqft,
##     data = adom_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190824  -25431    1937   27812  206233
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -6.086e+05  4.886e+05  -1.246  0.21356
## bathstotal                     6.340e+04  1.795e+05   0.353  0.72419
## sqft                           7.220e+01  1.770e+02   0.408  0.68361
## yrhouse                        2.645e+04  1.783e+04   1.483  0.13879
## lotsqft                        8.858e+01  3.449e+01   2.568  0.01054 *
## bathstotal:sqft                4.163e+01  4.968e+01   0.838  0.40247
## bathstotal:yrhouse            -4.563e+03  6.464e+03  -0.706  0.48054
## sqft:yrhouse                   9.313e-01  6.449e+00   0.144  0.88523
## bathstotal:lotsqft            -1.975e+01  8.529e+00  -2.316  0.02101 *
```

```
## sqft:lotsqft                    -2.286e-03  7.022e-03  -0.325  0.74496
## yrhouse:lotsqft                 -3.376e+00  1.286e+00  -2.626  0.00893 **
## bathstotal:sqft:yrhouse         -1.251e+00  1.823e+00  -0.686  0.49291
## bathstotal:sqft:lotsqft          5.646e-04  1.709e-03   0.330  0.74124
## bathstotal:yrhouse:lotsqft       9.478e-01  3.343e-01   2.836  0.00477 **
## sqft:yrhouse:lotsqft             3.217e-05  2.625e-04   0.123  0.90252
## bathstotal:sqft:yrhouse:lotsqft -3.217e-05  6.456e-05  -0.498  0.61852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49180 on 466 degrees of freedom
## Multiple R-squared:  0.8985, Adjusted R-squared:  0.8952
## F-statistic:   275 on 15 and 466 DF,  p-value: < 2.2e-16

plot(ols2)
```

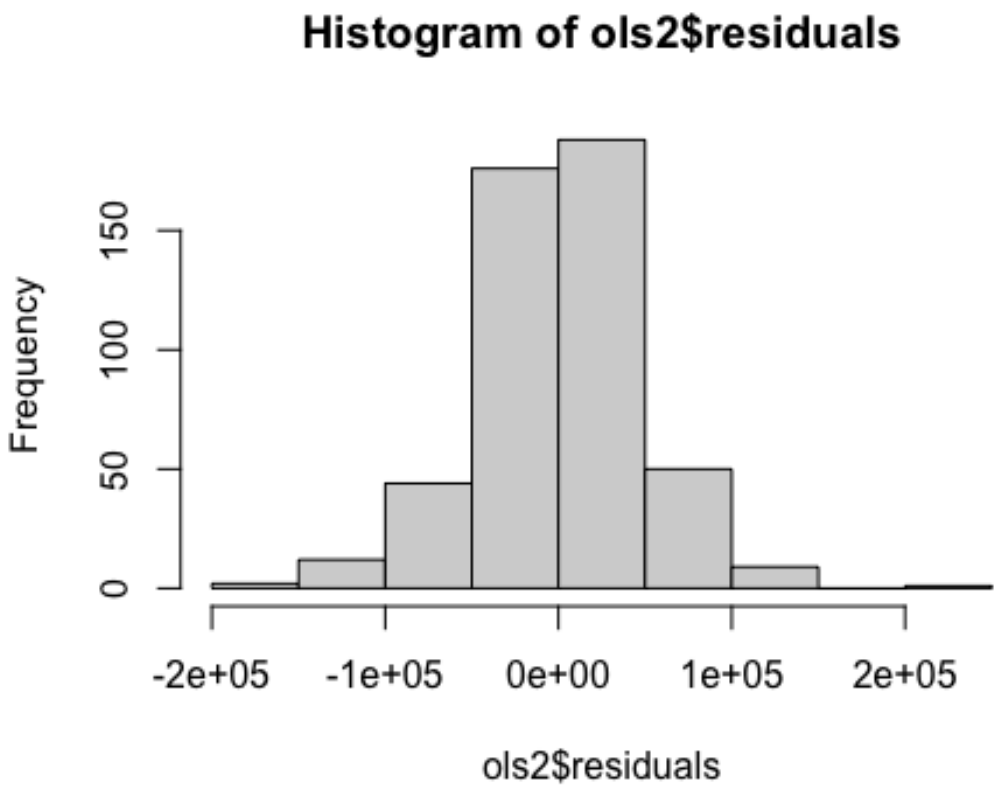

Residuals vs Fitted
Residuals
2e+05
0e+00
-2e+05
480
162
471
200000
600000
1000000
1400000
Fitted values
lm((pricesold) ~ bathstotal * sqft * yrhouse * lotsqft)


Normal Q-Q
Standardized residuals
6
4
2
0
-2
-4
-6
475
480
471
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm((pricesold) ~ bathstotal * sqft * yrhouse * lotsqft)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
471
475
480
200000
600000
1000000
1400000
Fitted values
lm((pricesold) ~ bathstotal * sqft * yrhouse * lotsqft)


Residuals vs Leverage
Standardized residuals
-6
-4
-2
0
2
4
6
475
482
1
1
Cook's distance
471
0.0
0.2
0.4
0.6
0.8
Leverage
lm((pricesold) ~ bathstotal * sqft * yrhouse * lotsqft)

```
hist(ols2$residuals)
```



#Checkind independence in model 2. bathstotal,sqft,yrhouse,lotsqft

```
#plot(adom_df[,c(1,2,3,4,5,6,7,8,9,10,11)],pch=19,main="Continuous Variables only")
#Garages can be consider as a categorical variable so it's not at the corrplot.
independence = round(cor(adom_df[,c(2,3,6,7,11)]),11)
library(corrplot)
corrplot(independence,method="number",is.corr = F)
```


bathstotal
sqft
lotsqft
yrhouse
pricesold
bathstotal 1.00 0.84 0.69 -0.26 0.79
sqft 0.84 1.00 0.79 -0.26 0.91
lotsqft 0.69 0.79 1.00 -0.28 0.85
yrhouse -0.26 -0.26 -0.28 1.00 -0.30
pricesold 0.79 0.91 0.85 -0.30 1.00
1
0.87
0.74
0.61
0.48
0.35
0.22
0.09
-0.04
-0.17
-0.3