Week 8&9 Assignments

Andrew Jordan

2022-05-11

Question 3.A.i

```
library(readxl)
library(ggplot2)
housing_df <- read_excel("data/week-7-housing.xlsx")

housing_df <- within(housing_df,bathrooms <- bath_full_count + (bath_half_count*.5) + (bath_3qtr_count*

housing_df <- housing_df[ ,c("Sale Date","Sale Price","addr_full","zip5","square_feet_total_living","bed

housing_df <- housing_df[housing_df$"Sale Price" < 2000000 & housing_df$"Sale Price" > 99999,]

summary(housing_df)
```

```
##    Sale Date                     Sale Price       addr_full
##  Min.   :2006-01-03 00:00:00.00   Min.   : 100000   Length:12578
##  1st Qu.:2008-07-14 06:00:00.00   1st Qu.: 460000   Class :character
##  Median :2011-11-30 12:00:00.00   Median : 590000   Mode  :character
##  Mean   :2011-08-06 01:16:35.45   Mean   : 624213
##  3rd Qu.:2014-06-17 00:00:00.00   3rd Qu.: 740000
##  Max.   :2016-12-16 00:00:00.00   Max.   :1990000
##       zip5       square_feet_total_living    bedrooms        bathrooms
##  Min.   :98052   Min.   :  240            Min.   : 0.000   Min.   : 0.000
##  1st Qu.:98052   1st Qu.: 1822            1st Qu.: 3.000   1st Qu.: 2.250
##  Median :98052   Median : 2420            Median : 4.000   Median : 2.500
##  Mean   :98053   Mean   : 2526            Mean   : 3.476   Mean   : 2.469
##  3rd Qu.:98053   3rd Qu.: 3100            3rd Qu.: 4.000   3rd Qu.: 2.750
##  Max.   :98074   Max.   :13540            Max.   :11.000   Max.   :23.500
##    year_built   year_renovated      sq_ft_lot       building_grade
##  Min.   :1900   Min.   :   0.00   Min.   :    785   Min.   : 2.000
##  1st Qu.:1979   1st Qu.:   0.00   1st Qu.:   5400   1st Qu.: 8.000
##  Median :1998   Median :   0.00   Median :   7978   Median : 8.000
##  Mean   :1993   Mean   :  24.94   Mean   :  20945   Mean   : 8.237
##  3rd Qu.:2007   3rd Qu.:   0.00   3rd Qu.:  12510   3rd Qu.: 9.000
##  Max.   :2016   Max.   :2016.00   Max.   :1008414   Max.   :13.000
##   present_use
##  Min.   :  0.000
##  1st Qu.:  2.000
##  Median :  2.000
##  Mean   :  6.464
```

```
##  3rd Qu.:  2.000
##  Max.   :300.000
```

##Question 3.B.i As the focus for this dataset's usage is Sale Price and other variables that are possible predictors there are a few modifications I made to the dataset to exclude irrelevant variables. Bathrooms were consolidated from three separate categories of full bath, half bath, and three-quarter bath, into one category of bathrooms. Several redundant variables were excluded, including the city name and postal city name as those are covered by the zipcode category, and longitude and latitude as those are more reasonably covered by the full street address. The most relevant remaining variables included are the sale price, sale date, number of bedrooms and bathrooms, square footage of the property and the lot, building grade, each property's present use, the year the property was built, and-when applicable-the year it was most recently renovated. Outliers in Sale Price were also removed, with properties sold for over two million and under one hundred thousand being excluded.

##Question 3.B.ii

```
priceByLot_lm <- lm(housing_df$"Sale Price"~housing_df$sq_ft_lot,data=housing_df)

priceByVars_lm <- lm(housing_df$"Sale Price"~housing_df$zip5 + housing_df$bedrooms + housing_df$bathroo
```

As the location, square footage, number of rooms, and age of a house are all considered important factors when buying a home, these are the predictors I have chosen to include.

##Question 3.B.iii

```
summary(priceByLot_lm)
```

```
##
## Call:
## lm(formula = housing_df$"Sale Price" ~ housing_df$sq_ft_lot,
##     data = housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1144425  -158777   -30701   119692  1376759
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.051e+05  2.318e+03  261.03   <2e-16 ***
## housing_df$sq_ft_lot 9.116e-01  4.518e-02   20.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 237300 on 12576 degrees of freedom
## Multiple R-squared:  0.03135,    Adjusted R-squared:  0.03127
## F-statistic:   407 on 1 and 12576 DF,  p-value: < 2.2e-16
```

```
summary(priceByVars_lm)
```

```
##
## Call:
## lm(formula = housing_df$"Sale Price" ~ housing_df$zip5 + housing_df$bedrooms +
##     housing_df$bathrooms + housing_df$year_built + housing_df$square_feet_total_living,
```

```
##     data = housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1516642   -85424   -10168    71983  1392518
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -1.999e+08  8.955e+07  -2.232   0.0256 *
## housing_df$zip5                    2.029e+03  9.135e+02   2.222   0.0263 *
## housing_df$bedrooms               -2.085e+04  2.267e+03  -9.197  < 2e-16 ***
## housing_df$bathrooms               1.485e+04  3.516e+03   4.223 2.43e-05 ***
## housing_df$year_built              5.650e+02  1.013e+02   5.577 2.50e-08 ***
## housing_df$square_feet_total_living  1.783e+02  2.566e+00  69.509  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 170900 on 12572 degrees of freedom
## Multiple R-squared:  0.4977, Adjusted R-squared:  0.4975
## F-statistic:  2491 on 5 and 12572 DF,  p-value: < 2.2e-16
```

As the R2 of the first model is .03135, this indicates that the square footage of a lot accounts for only 3.14% of variation in the sale price. The second model's R2 value of .4977 shows a 49.77% relation between the predictors and variation in sale price. The inclusion of the additional predictors in the second model provides an explanation for nearly half of the variation found in sale price. The adjusted R2 values for the first and second models have a difference of .0008 and .0002 respectively. This shows that there is an expected .08% and .02% difference for the respective models were the data derived from the entire population, as opposed to a sample.

##Question 3.B.iV

```
library(lm.beta)
lm.beta(priceByVars_lm)
```

```
##
## Call:
## lm(formula = housing_df$"Sale Price" ~ housing_df$zip5 + housing_df$bedrooms +
##     housing_df$bathrooms + housing_df$year_built + housing_df$square_feet_total_living,
##     data = housing_df)
##
## Standardized Coefficients::
##                           (Intercept)                        housing_df$zip5
##                                    NA                             0.01423231
##                   housing_df$bedrooms                    housing_df$bathrooms
##                            -0.07539724                             0.04166350
##                housing_df$year_built housing_df$square_feet_total_living
##                             0.04015785                             0.69895662
```

The standardized beta results indicate a high degree of importance for square_feet_total_living as it shows for every one standard deviation in Sale Price, there should be a .70 rise of square_feet_total_living as well. zip5, bathrooms, and year_built all have lower degrees of importance, with results less than .1 per one standard deviation. ##Question 3.B.v

```
confint(priceByVars_lm)
```

```
##                                           2.5 %        97.5 %
## (Intercept)                        -3.754389e+08 -2.435690e+07
## housing_df$zip5                     2.388310e+02  3.819840e+03
## housing_df$bedrooms                -2.528885e+04 -1.640289e+04
## housing_df$bathrooms                7.954032e+03  2.173682e+04
## housing_df$year_built               3.664351e+02  7.636138e+02
## housing_df$square_feet_total_living  1.733158e+02  1.833744e+02
```

##Question 3.B.vi

```
anova(priceByLot_lm, priceByVars_lm)
```

```
## Analysis of Variance Table
##
## Model 1: housing_df$"Sale Price" ~ housing_df$sq_ft_lot
## Model 2: housing_df$"Sale Price" ~ housing_df$zip5 + housing_df$bedrooms +
##     housing_df$bathrooms + housing_df$year_built + housing_df$square_feet_total_living
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  12576 7.0840e+14
## 2  12572 3.6737e+14  4 3.4103e+14 2917.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##Question 3.B.vii

```
housing_df$residuals <- resid(priceByVars_lm)
housing_df$standardizedResiduals <- rstandard(priceByVars_lm)
summary(housing_df)
```

```
##    Sale Date                        Sale Price       addr_full
##  Min.   :2006-01-03 00:00:00.00   Min.   : 100000   Length:12578
##  1st Qu.:2008-07-14 06:00:00.00   1st Qu.: 460000   Class :character
##  Median :2011-11-30 12:00:00.00   Median : 590000   Mode  :character
##  Mean   :2011-08-06 01:16:35.45   Mean   : 624213
##  3rd Qu.:2014-06-17 00:00:00.00   3rd Qu.: 740000
##  Max.   :2016-12-16 00:00:00.00   Max.   :1990000
##       zip5       square_feet_total_living    bedrooms        bathrooms
##  Min.   :98052   Min.   :  240            Min.   : 0.000   Min.   : 0.000
##  1st Qu.:98052   1st Qu.: 1822            1st Qu.: 3.000   1st Qu.: 2.250
##  Median :98052   Median : 2420            Median : 4.000   Median : 2.500
##  Mean   :98053   Mean   : 2526            Mean   : 3.476   Mean   : 2.469
##  3rd Qu.:98053   3rd Qu.: 3100            3rd Qu.: 4.000   3rd Qu.: 2.750
##  Max.   :98074   Max.   :13540            Max.   :11.000   Max.   :23.500
##    year_built   year_renovated      sq_ft_lot       building_grade
##  Min.   :1900   Min.   :   0.00   Min.   :    785   Min.   : 2.000
##  1st Qu.:1979   1st Qu.:   0.00   1st Qu.:   5400   1st Qu.: 8.000
##  Median :1998   Median :   0.00   Median :   7978   Median : 8.000
##  Mean   :1993   Mean   :  24.94   Mean   :  20945   Mean   : 8.237
##  3rd Qu.:2007   3rd Qu.:   0.00   3rd Qu.:  12510   3rd Qu.: 9.000
```

```
##  Max.   :2016   Max.   :2016.00   Max.   :1008414   Max.   :13.000
##   present_use        residuals        standardizedResiduals
##  Min.   :  0.000   Min.   :-1516642   Min.   :-8.903382
##  1st Qu.:  2.000   1st Qu.:  -85424   1st Qu.:-0.499844
##  Median :  2.000   Median :  -10168   Median :-0.059495
##  Mean   :  6.464   Mean   :       0   Mean   :-0.000048
##  3rd Qu.:  2.000   3rd Qu.:   71983   3rd Qu.: 0.421176
##  Max.   :300.000   Max.   : 1392518   Max.   : 8.153907
```

##Question 3.B.viii

```
housing_df$greaterResiduals <- housing_df$standardizedResiduals > 2
housing_df$lesserResiduals <- housing_df$standardizedResiduals < -2
summary(housing_df)
```

```
##    Sale Date                        Sale Price       addr_full
##  Min.   :2006-01-03 00:00:00.00   Min.   : 100000   Length:12578
##  1st Qu.:2008-07-14 06:00:00.00   1st Qu.: 460000   Class :character
##  Median :2011-11-30 12:00:00.00   Median : 590000   Mode  :character
##  Mean   :2011-08-06 01:16:35.45   Mean   : 624213
##  3rd Qu.:2014-06-17 00:00:00.00   3rd Qu.: 740000
##  Max.   :2016-12-16 00:00:00.00   Max.   :1990000
##       zip5       square_feet_total_living    bedrooms        bathrooms
##  Min.   :98052   Min.   :  240            Min.   : 0.000   Min.   : 0.000
##  1st Qu.:98052   1st Qu.: 1822            1st Qu.: 3.000   1st Qu.: 2.250
##  Median :98052   Median : 2420            Median : 4.000   Median : 2.500
##  Mean   :98053   Mean   : 2526            Mean   : 3.476   Mean   : 2.469
##  3rd Qu.:98053   3rd Qu.: 3100            3rd Qu.: 4.000   3rd Qu.: 2.750
##  Max.   :98074   Max.   :13540            Max.   :11.000   Max.   :23.500
##    year_built   year_renovated      sq_ft_lot       building_grade
##  Min.   :1900   Min.   :   0.00   Min.   :    785   Min.   : 2.000
##  1st Qu.:1979   1st Qu.:   0.00   1st Qu.:   5400   1st Qu.: 8.000
##  Median :1998   Median :   0.00   Median :   7978   Median : 8.000
##  Mean   :1993   Mean   :  24.94   Mean   :  20945   Mean   : 8.237
##  3rd Qu.:2007   3rd Qu.:   0.00   3rd Qu.:  12510   3rd Qu.: 9.000
##  Max.   :2016   Max.   :2016.00   Max.   :1008414   Max.   :13.000
##   present_use        residuals        standardizedResiduals greaterResiduals
##  Min.   :  0.000   Min.   :-1516642   Min.   :-8.903382     Mode :logical
##  1st Qu.:  2.000   1st Qu.:  -85424   1st Qu.:-0.499844     FALSE:12216
##  Median :  2.000   Median :  -10168   Median :-0.059495     TRUE :362
##  Mean   :  6.464   Mean   :       0   Mean   :-0.000048
##  3rd Qu.:  2.000   3rd Qu.:   71983   3rd Qu.: 0.421176
##  Max.   :300.000   Max.   : 1392518   Max.   : 8.153907
##  lesserResiduals
##  Mode :logical
##  FALSE:12332
##  TRUE :246
##
##
##
```

##Question 3.B.ix

```
sum(housing_df$greaterResiduals)
```

```
## [1] 362
```

##Question 3.B.x

```
housing_df[housing_df$greaterResiduals, c("Sale Price","zip5","square_feet_total_living","bedrooms","ba
```

```
## # A tibble: 362 x 5
##    `Sale Price`  zip5 square_feet_total_living bedrooms bathrooms
##           <dbl> <dbl>                   <dbl>    <dbl>     <dbl>
##  1      1392000 98052                    3740        4      4.75
##  2      1053649 98053                    2680        2      2.5
##  3      1900000 98053                    6610        4      4.25
##  4      1080135 98053                    2700        3      2.75
##  5      1520000 98052                    4640        5      4.25
##  6      1390000 98053                     660        0      1
##  7      1390000 98053                    3280        3      2.75
##  8      1300000 98052                    4240        4      3.5
##  9      1588359 98053                    3360        2      2.5
## 10      1450000 98052                    3480        3      2.5
## # ... with 352 more rows
```

##Question 3.B.xi

```
housing_df$cooksDistance <- cooks.distance(priceByVars_lm)
housing_df$leverage <- hatvalues(priceByVars_lm)
housing_df$covarianceRatios <- covratio(priceByVars_lm)

housing_df[housing_df$greaterResiduals, c("cooksDistance", "leverage", "covarianceRatios")]
```

```
## # A tibble: 362 x 3
##    cooksDistance leverage covarianceRatios
##            <dbl>    <dbl>            <dbl>
##  1      0.00224  0.00141             0.997
##  2      0.000352 0.000468            0.999
##  3      0.00388  0.00239             0.998
##  4      0.000160 0.000173            0.998
##  5      0.00314  0.00182             0.997
##  6      0.0122   0.00187             0.984
##  7      0.000539 0.000252            0.995
##  8      0.000413 0.000528            0.999
##  9      0.00254  0.000737            0.991
## 10      0.00160  0.000648            0.994
## # ... with 352 more rows
```

##Question 3.B.xii

```
library(car)
```

```
## Loading required package: carData
```

```
durbinWatsonTest(priceByVars_lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3139655      1.372035       0
##  Alternative hypothesis: rho != 0
```

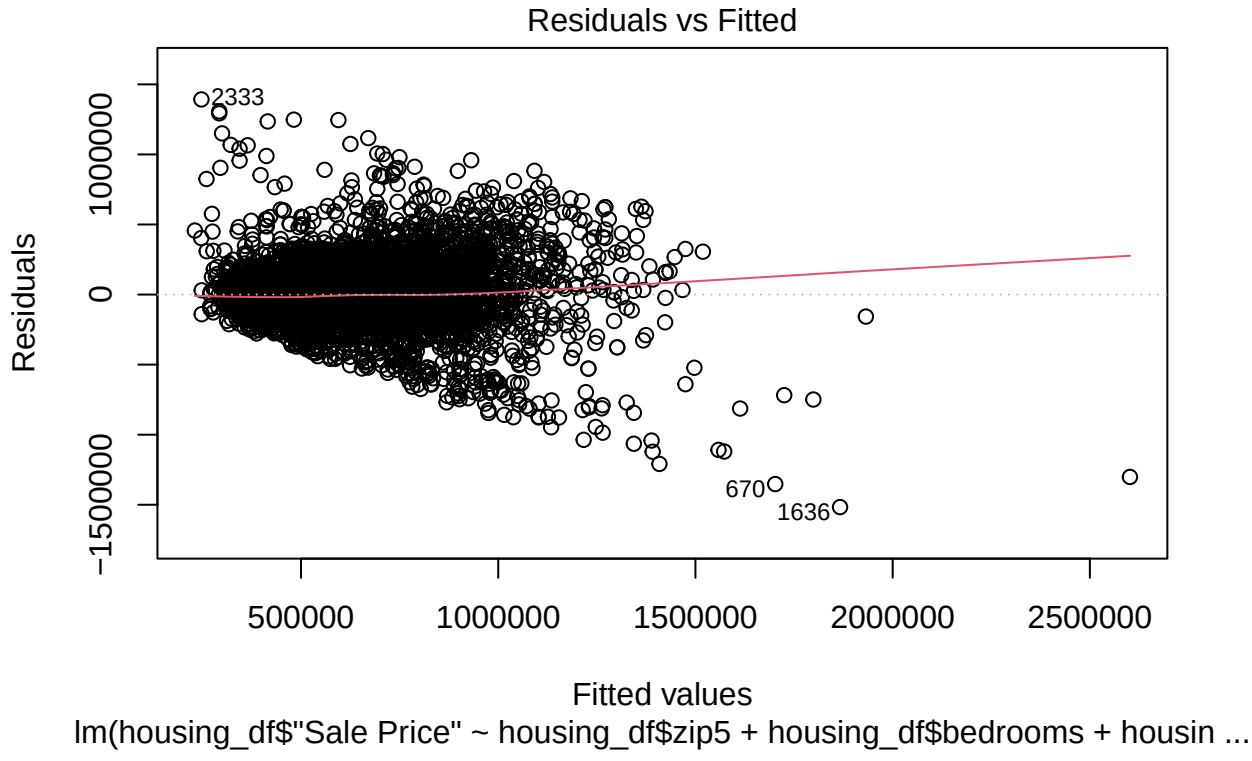
##Question 3.B.xiii

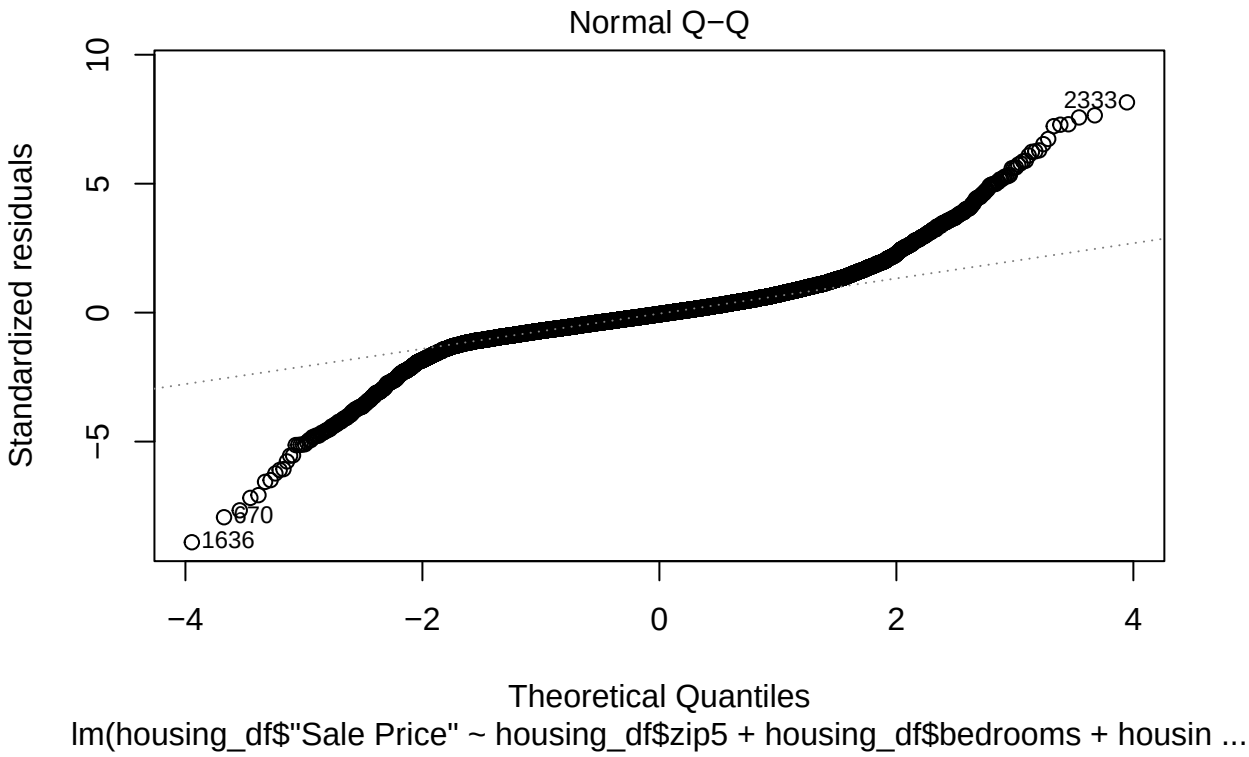
```
vif(priceByVars_lm)
```

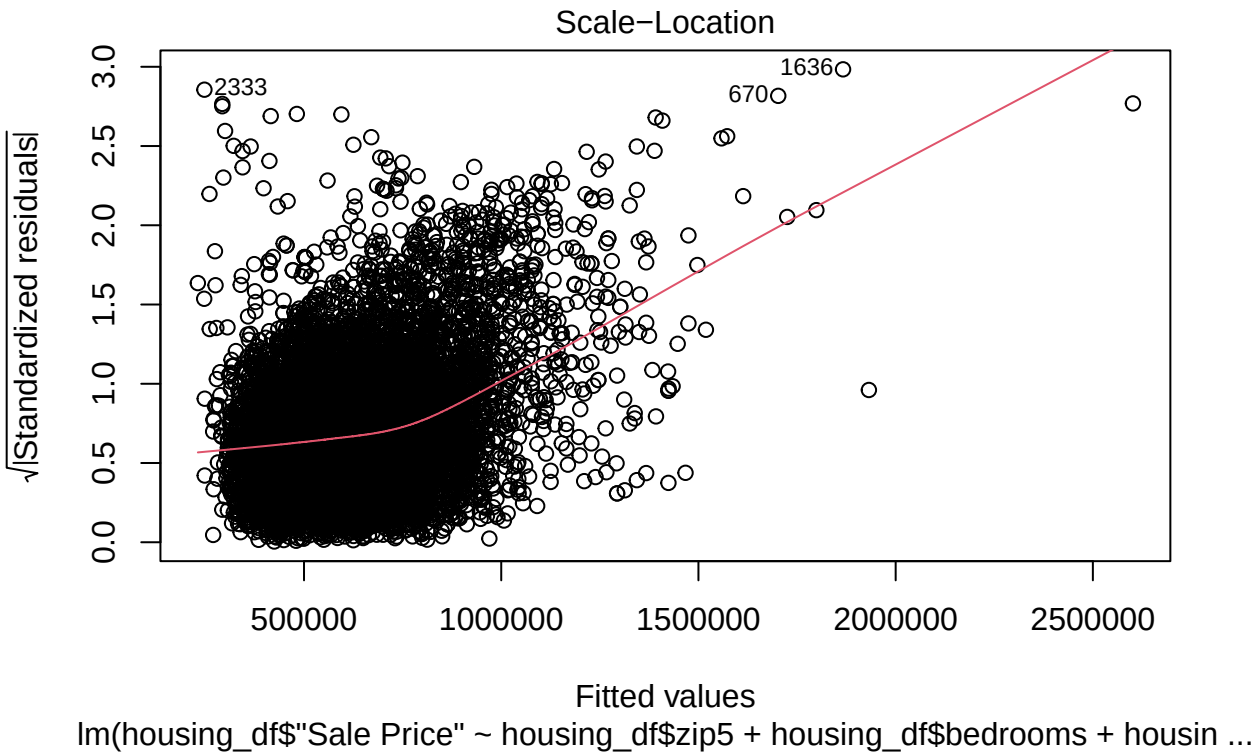
```
##                       housing_df$zip5                 housing_df$bedrooms
##                              1.027139                              1.682112
##                 housing_df$bathrooms                housing_df$year_built
##                              2.436551                              1.297637
## housing_df$square_feet_total_living
##                              2.530631
```

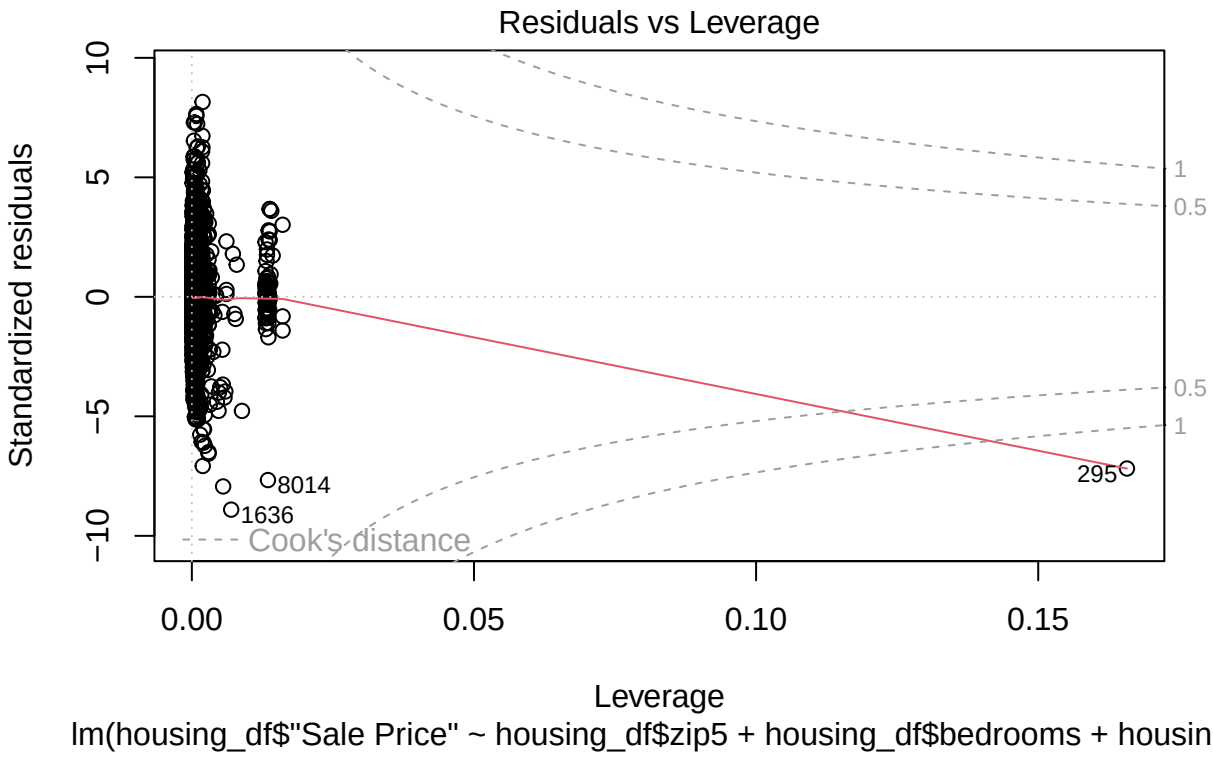
##Question 3.B.xiV

```
plot(priceByVars_lm)
```

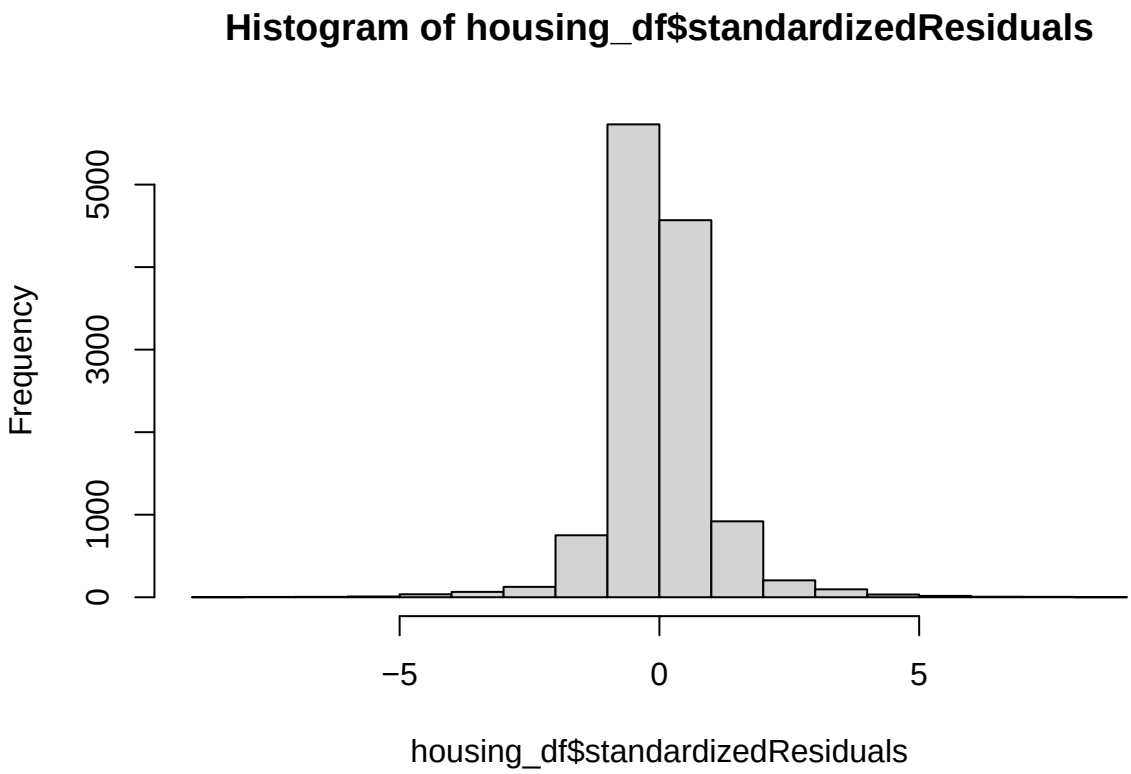



Normal Q−Q
Standardized residuals
10
5
0
−5
2333
970
1636
−4
−2
0
2
4
Theoretical Quantiles
lm(housing_df$"Sale Price" ~ housing_df$zip5 + housing_df$bedrooms + housin ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2333
670
1636
500000
1000000
1500000
2000000
2500000
Fitted values
lm(housing_df$"Sale Price" ~ housing_df$zip5 + housing_df$bedrooms + housin ...




```
hist(housing_df$standardizedResiduals)
```



##Question 3.B.xv