Week 8 Exercise

Alissa Trujillo

2022-05-09

Housing Data

a. Gathering the Data Set

```
setwd("/Users/alissa/Documents/Grad/DSC 520/dsc520")
housing_df <- read.csv("data/week-6-housing.csv")
```

b. Analysis

i. Transformations I completed a couple of transformations to make some of the data columns more useful. Some of the data is difficult to use in the format it is presented in. The year_renovated column is particularly difficult to utilize since the homes that were not renovated contain a numeric value, 0. To address this, the first column I added was a boolean value of whether a house has been renovated.

```
housing_df$renovated <- housing_df$year_renovated != 0
```

Next, I added a column that denotes the sale year of each home. Then I converted it to a numeric value in order to be able to use it in future calculations.

```
library(stringi)
housing_df$sale_year <- paste("20", stri_sub(housing_df$sale_date,
    -2, -1), sep = "")
housing_df$sale_year <- as.numeric(housing_df$sale_year)
```

The next step I took was to add a new column that calculates the age of the home upon sale. This is an important piece of information that was not contained within the data that could be a bug predictor of sale price.

```
housing_df$age_at_sale <- housing_df$sale_year - housing_df$year_built
```

```
library(lm.beta)

sale_price_sqft_lm <- lm(sale_price ~ sq_ft_lot, data = housing_df)

sale_price_predictors_lm <- lm(sale_price ~ sq_ft_lot + square_feet_total_living +
    age_at_sale + renovated, data = housing_df)
```

ii. Sale Price Variables For my multiple regression data frame, I have selected both square feet of the lot and square feet of the living space because I think they will be important indicators in the sale price of a home. There should be a positive correlation between square footage and the price of a home The additional factors I included were age at sale, which I calculated in order to determine the age of the home when it was sold. I think there will be a negative correlation between age and sale price, as newer homes with newer appliances, amenities, etc. will hold more value. I also included a variable denoting whether a home has been renovated or not, because I feel like that adds significant value to a home. Out of the information contained in the data table, these factors seem like they will have the most influence on sale price.

```
summary(sale_price_sqft_lm)
```

iii. Summary

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.418e+05  3.800e+03  168.90   <2e-16 ***
## sq_ft_lot   8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

The R2 statistic for this simple regression is 0.01435, meaning that the square feet of the lot can account for 1.4% of variation in sale price. The adjusted R2 is 0.01428 after taking into account the number of independent variables. Since there is only one variable we are testing, the values are very similar, but as we add more variables to our model, R2 will go up and the adjusted R2 will be more conservative because it takes into account that we are adding a number of new variables.

```
summary(sale_price_predictors_lm)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot + square_feet_total_living +
##     age_at_sale + renovated, data = housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2118445  -120235   -42112    47092  3782776
##
## Coefficients:
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.796e+05  1.077e+04  25.949  < 2e-16 ***
## sq_ft_lot                   2.702e-01  5.835e-02   4.630 3.69e-06 ***
## square_feet_total_living    1.668e+02  3.502e+00  47.626  < 2e-16 ***
## age_at_sale                -2.784e+03  1.978e+02 -14.071  < 2e-16 ***
## renovatedTRUE               1.392e+05  2.855e+04   4.877 1.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357400 on 12860 degrees of freedom
## Multiple R-squared:  0.219,  Adjusted R-squared:  0.2188
## F-statistic: 901.6 on 4 and 12860 DF,  p-value: < 2.2e-16
```

The R2 has increased after adding more variables to our regression, now sitting at 0.219. This means that the variables at hand (lot size, size of living space, age at sale, and whether the house is renovated) account for 21.9% of the variation in sale price. The adjusted R squared is slightly lower, at 0.2188, which accounts for the fact that we are testing multiple variables. All of these variables are statistically significant, meaning that we can reject the null hypothesis that they do not have an effect on sale price.

```
lm.beta(sale_price_predictors_lm)
```

iv. Betas

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot + square_feet_total_living +
##     age_at_sale + renovated, data = housing_df)
##
## Standardized Coefficients::
##              (Intercept)                sq_ft_lot square_feet_total_living
##               0.00000000               0.03803923               0.40826482
##              age_at_sale            renovatedTRUE
##              -0.12105550               0.03920664
```

The betas for each parameter indicate the strength of the relationship between each of them and the dependent variable we are measuring. The beta for square_feet_total_living is 0.408, meaning that for every standard deviation increase in sale price, there is also a 0.408 standard deviation increase in living space, all other variables held constant. The beta for sq_ft_lot is 0.038, meaning that for every standard deviation increase in sale price, there is also a 0.038 standard deviation increase in the size of the lot, all other variables held constant. The beta for age_at_sale is -0.121, meaning that for every standard deviation increase in sale price, there is also a 0.121 standard deviation decrease in the age of the home, all other variables held constant. The beta for renovated is 0.039, meaning that for every standard deviation increase in sale price, there is also a 0.039 standard deviation increase in whether a home is renovated, all other variables held constant. This result, however, is a bit more difficult to interpret because it is a boolean variable.

```
confint(sale_price_predictors_lm)
```

v. Confidence Intervals

```
##                                 2.5 %        97.5 %
## (Intercept)               2.584622e+05 300700.647272
## sq_ft_lot                 1.558018e-01      0.384562
## square_feet_total_living  1.599282e+02    173.657627
## age_at_sale              -3.171819e+03  -2396.198318
## renovatedTRUE             8.327419e+04 195207.408420
```

These are the confidence intervals presented at a 95% level. These results indicate that there is a 95% chance of the true value of each of the coefficients falling within these intervals. All of these values are, at a 95% level, significantly different than 0, which supports our ability to reject the null hypothesis that the variables do not have an effect on sale price.

vi. ANOVA

```
sale_price_anova <- aov(sale_price ~ sq_ft_lot, data = housing_df)
sale_price_anova_m <- aov(sale_price ~ sq_ft_lot + square_feet_total_living +
    age_at_sale + renovated, data = housing_df)
summary(sale_price_anova)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## sq_ft_lot       1 3.020e+13 3.020e+13   187.3 <2e-16 ***
## Residuals   12863 2.073e+15 1.612e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(sale_price_anova_m)
```

```
##                             Df    Sum Sq   Mean Sq F value   Pr(>F)
## sq_ft_lot                    1 3.020e+13 3.020e+13  236.38  < 2e-16 ***
## square_feet_total_living     1 4.049e+14 4.049e+14 3169.54  < 2e-16 ***
## age_at_sale                  1 2.258e+13 2.258e+13  176.78  < 2e-16 ***
## renovated                    1 3.038e+12 3.038e+12   23.78 1.09e-06 ***
## Residuals                12860 1.643e+15 1.277e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The original simple regression, after completing an ANOVA assessment, had a very high value of residuals, 2073376756946868. After analysing the variance of the improved multiple regression model, the residuals decreased to 1642849926731011. This is still a huge number but since the numbers we are dealing with, house prices, are also very high, this is to be expected. The value of the residuals dropped by just over 20% which is a significant improvement for our model.

vii. Casewise Diagnostics

```
housing_diagnostics <- housing_df

housing_diagnostics$predicted.probabilities <- fitted(sale_price_predictors_lm)
housing_diagnostics$standardized.residuals <- rstandard(sale_price_predictors_lm)
housing_diagnostics$studentized.residuals <- rstudent(sale_price_predictors_lm)
```

```
housing_diagnostics$dfbeta <- dfbeta(sale_price_predictors_lm)
housing_diagnostics$dffit <- dffits(sale_price_predictors_lm)
housing_diagnostics$leverage <- hatvalues(sale_price_predictors_lm)
housing_diagnostics$covariance.ratios <- covratio(sale_price_predictors_lm)
housing_diagnostics$cooks.distance <- cooks.distance(sale_price_predictors_lm)
head(housing_diagnostics[, c("leverage", "studentized.residuals",
    "dfbeta")])
```

```
##       leverage studentized.residuals dfbeta.(Intercept) dfbeta.sq_ft_lot
## 1 1.374615e-04           -0.12229743      -5.895636e+00     1.094923e-05
## 2 1.630892e-04           -0.31186067      -1.612956e+01     2.824598e-05
## 3 9.431057e-05           -0.33149984       1.200908e+00     5.821082e-05
## 4 2.182803e-04           -0.07439223      -2.957338e+00     7.899769e-06
## 5 1.758316e-04           -0.22247361      -2.189970e+01     3.108804e-06
## 6 3.315064e-04           -2.20526446       1.619102e+02     6.425153e-04
##   dfbeta.square_feet_total_living dfbeta.age_at_sale dfbeta.renovatedTRUE
## 1                   -2.369370e-04       1.598292e-01        -2.205223e+00
## 2                   -7.668939e-04       4.928067e-01        -8.446974e+00
## 3                   -3.722811e-03      -1.345008e-01         1.342399e+01
## 4                    1.146247e-03      -1.257944e-01         5.654299e+00
## 5                    6.555639e-03      -5.950253e-02         5.701714e+00
## 6                   -1.031783e-01       1.334890e+00         3.554542e+01
```

Above I have performed a selection of case wise diagnostics. They are all saved to variables so I am able to use them to analyze my data.

viii. Standardized Residuals

```
large_residuals <- subset(housing_diagnostics$standardized.residuals,
    housing_diagnostics$standardized.residuals > 2 | housing_diagnostics$standardized.residuals <
        -2)

housing_diagnostics$large_residual <- housing_diagnostics$standardized.residuals >
    2 | housing_diagnostics$standardized.residuals < -2
```

This new variable, large_residuals, contains all of the standardized residuals that are higher than 2 and less than -2. The variable added to the table, large_residual, indicates TRUE or FALSE whether the residual is considered large.

ix. Sum of Large Residuals

```
sum(large_residuals)
```

```
## [1] 1189.923
```

Summing together all of our large residuals, we can see that the sum is 1189.923.

x. Variables Containing Large Residuals

```
housing_large_residuals <- subset(housing_diagnostics, housing_diagnostics$large_residual ==
    TRUE)
nrow(housing_large_residuals)
```

```
## [1] 329
```

```
head(housing_large_residuals[c("addr_full", "standardized.residuals",
    "large_residual")])
```

```
##                           addr_full standardized.residuals large_residual
## 6                  8101 229TH DR NE              -2.204933           TRUE
## 25           25149 NE PATTERSON WAY              -2.430486           TRUE
## 115             19656 NE REDMOND RD               3.028049           TRUE
## 160                28527 NE 47TH PL              -2.129685           TRUE
## 178 13414 WOODINVILLE REDMOND RD NE              -2.462153           TRUE
## 239               24103 NE 122ND ST               2.095131           TRUE
```

This table shows us all of the instances in our data of large residuals. According to this, there are 329 individual homes in our table that have residuals greater than an absolute value of 2. This is roughly 2.6% of our data, which is acceptable. Up to 5% of cases may have residuals before we have cause to worry.

xi. Leverage, Cook's Distance, Covariance Ratios

Leverage

```
housing_diagnostics$large_leverage <- housing_diagnostics$leverage >
    (5/12865) * 3
housing_high_leverage <- subset(housing_diagnostics, housing_diagnostics$large_leverage ==
    TRUE)
nrow(housing_high_leverage)
```

```
## [1] 479
```

There are 479 instances of higher than normal leverage, denoted by (k+1)/n. This represents roughly 3.7% of the cases studied, which is less than the 5% that is acceptable. In a normally distributed sample, this is normal to see a small number of data points that are far from the mean. This does not denote a problem in our data.

Cook's Distance

```
housing_diagnostics$cooksd_high <- housing_diagnostics$cooks.distance >
    1
housing_high_cooksd <- subset(housing_diagnostics, housing_diagnostics$cooksd_high ==
    TRUE)
nrow(housing_high_cooksd)
```

```
## [1] 0
```

We will now evaluate the Cook's distance of each of our data points. Anything over a value of 1 is a cause of alarm. According to our data, we do not have any cases where the Cook's distance is larger than 1. This means that none of our individual pieces of data have exaggerated influence on our model.

Covariance Ratio

```
housing_diagnostics$large_cratio <- housing_diagnostics$covariance.ratios >
    (1 + 3 * (5/12865)) | housing_diagnostics$covariance.ratios <
    (1 - 3 * (5/12865))
housing_high_cratio <- subset(housing_diagnostics, housing_diagnostics$large_cratio ==
    TRUE)
nrow(housing_high_cratio)
```

```
## [1] 746
```

The upper limit of the acceptable values for the covariance ratio is 1 + 3 x (5/12865), while the lower limit is 1 + 3 x (5/12865). This means that 5.8% of our data points fall outside our acceptable covariance ratio. This can be problematic, but we can see if these data points are also troublesome in different ways.

```
problematic_data <- subset(housing_diagnostics, housing_diagnostics$large_cratio ==
    TRUE && housing_diagnostics$large_leverage == TRUE)
nrow(problematic_data)
```

```
## [1] 0
```

This table is blank, showing us that there are no data points which both have a higher than normal leverage as well as a covariance ratio outside the acceptable bounds. We know also that there are no pieces of data that have a concerning Cook's distance. Out of the three tests we performed on the data, no data point failed more than one. This helps bolster our confidence that there are no pieces of data that are overbearingly influencing our data.

xii. Independence

```
durbinWatsonTest(sale_price_predictors_lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.722918     0.5541517       0
##  Alternative hypothesis: rho != 0
```

The DWT results in a D-W statistic of 0.554, which is very far off from the preferred value of 2. Since it is less than 1, it suggests that there is positive autocorrelation between the variables. This is a cause for concern, though the nature of the data and housing markets seems to be a likely candidate for issues like this. Since changes in the market tend to trend over time and by area, it makes sense that the data points show correlation within themselves. While this is concerning, it makes sense due to the type of data we are dealing with. Past housing prices are likely to influence future housing prices, and housing market trends change fairly slowly.

xiii. Multicollinearity

```
vif(sale_price_predictors_lm)
```

```
##                sq_ft_lot square_feet_total_living              age_at_sale
##                 1.111411                 1.210025                 1.218680
##                renovated
##                 1.064307
```

The VIF for each variable is fairly close to 1, with the average being also very close to 1. A troublesome VIF is considered to be over 10, so our data looks good here.

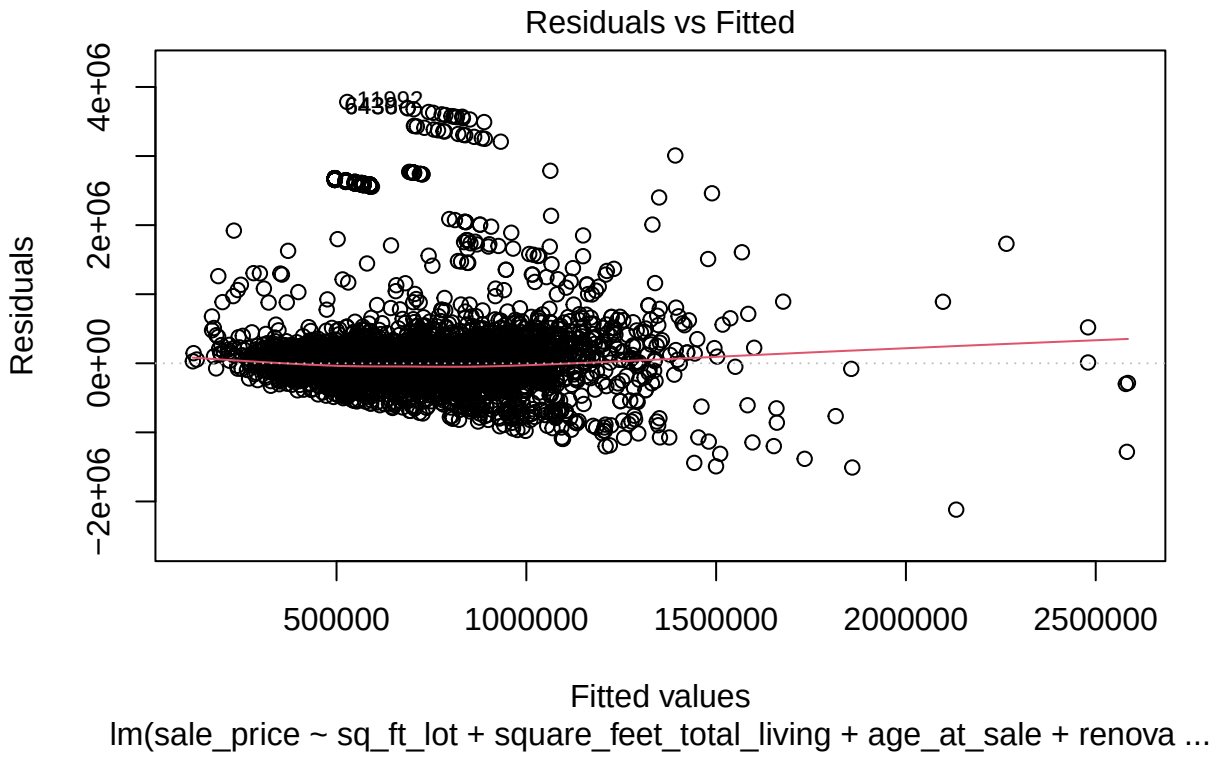
```
1/vif(sale_price_predictors_lm)
```

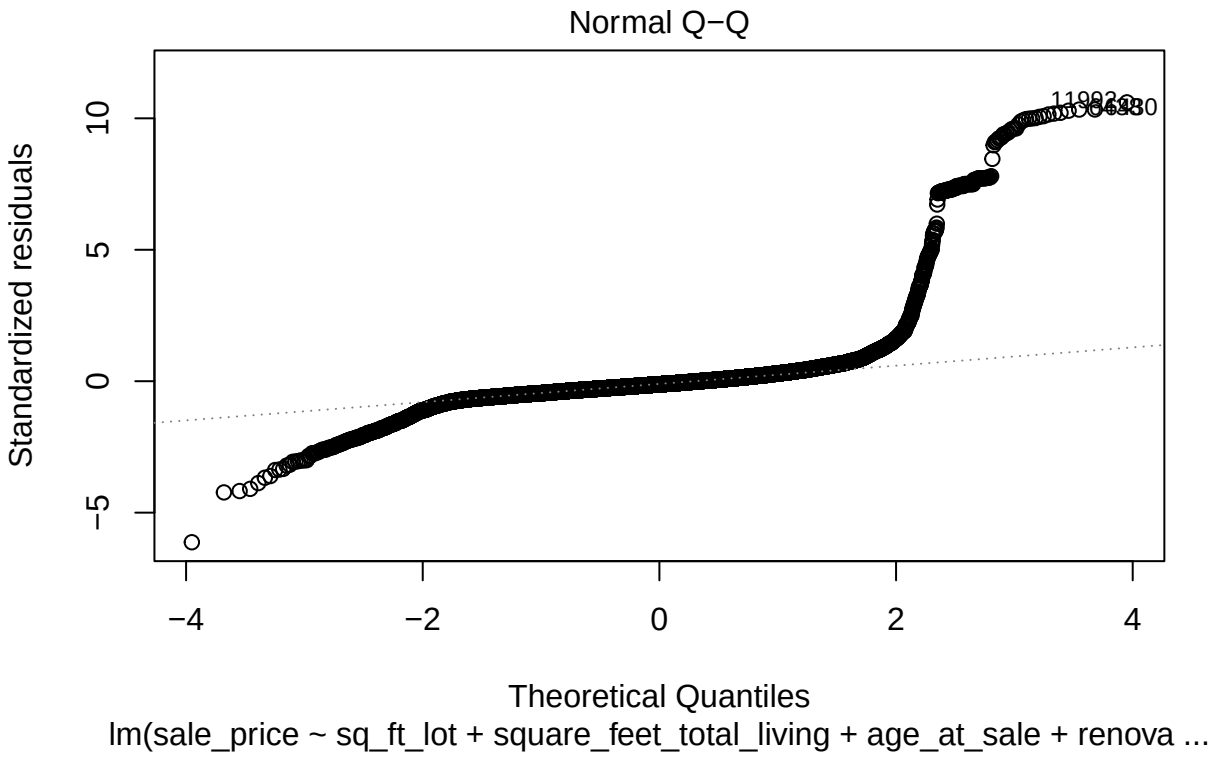
```
##                sq_ft_lot square_feet_total_living              age_at_sale
##                0.8997571                0.8264294                0.8205599
##                renovated
##                0.9395785
```

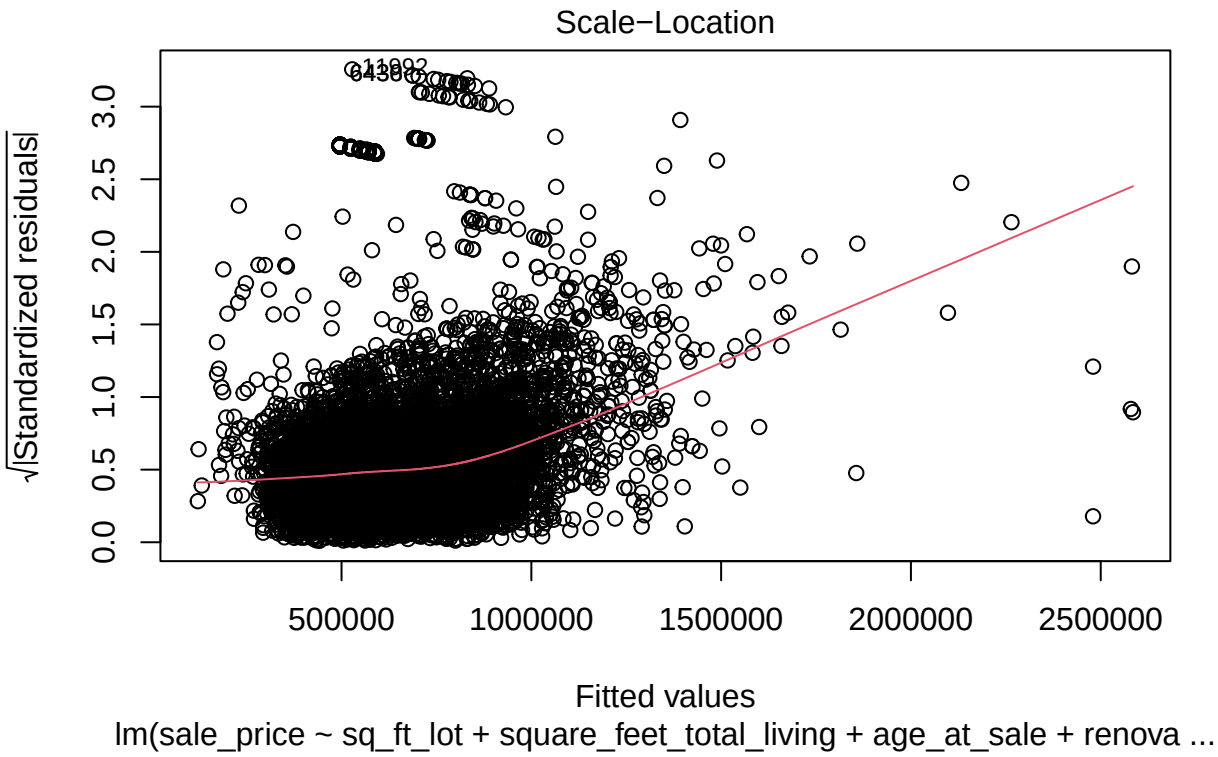
The tolerance for each variable also seems to be close to 1. A value under 0.2 would be a cause for concern, so our data checks out here as well. After analyzing our data, we can be confident that there is no multicollinearity.

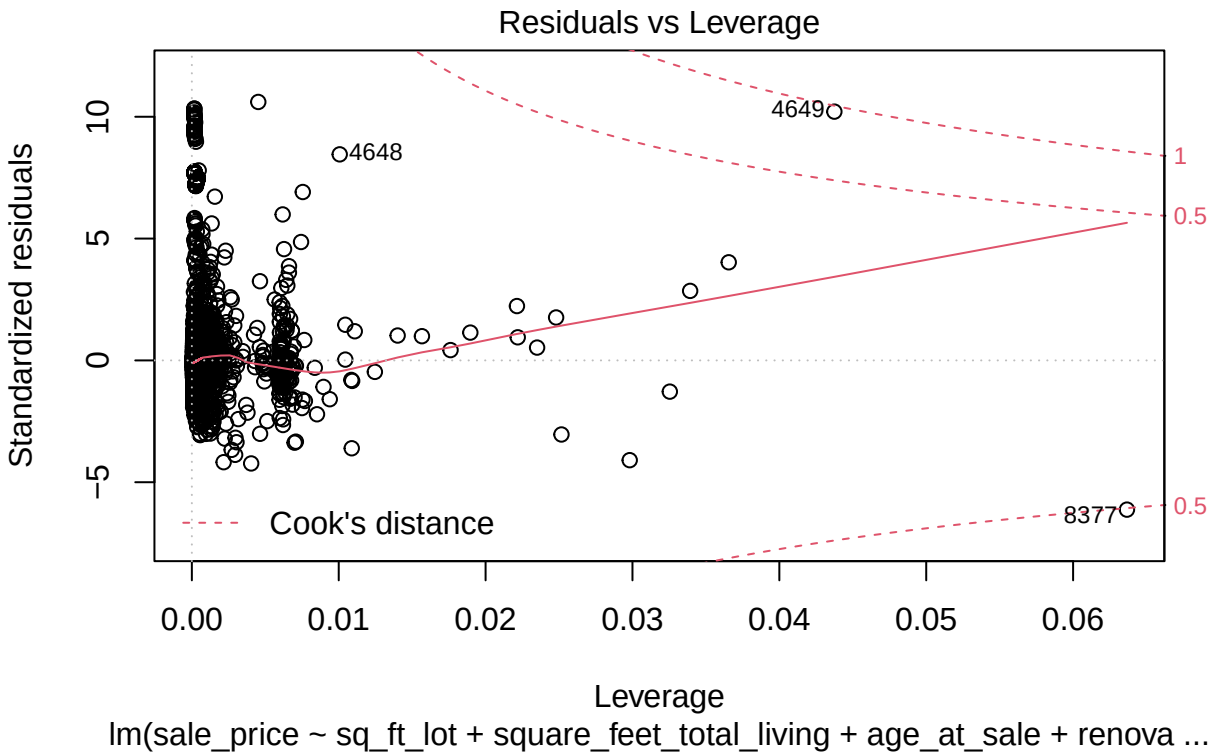
xiv. Residual Graphs

```
plot(sale_price_predictors_lm)
```




Normal Q−Q
Standardized residuals
10
5
0
−5
−4
−2
0
2
4
Theoretical Quantiles
lm(sale_price ~ sq_ft_lot + square_feet_total_living + age_at_sale + renova ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
500000
1000000
1500000
2000000
2500000
Fitted values
lm(sale_price ~ sq_ft_lot + square_feet_total_living + age_at_sale + renova ...



The first plot shows our residuals versus the fitted values. This graph is heavily focused towards the left hand side. There are a number of fitted values that have much higher values than their residual counterpart. This shows that the model predicts much higher values than the actual sale price in many instances. This could be due in part to a cap of sorts to the housing market, where homes above a certain price are too unaffordable and so homeowners do not extend the price of the home as high as it could be in a free economic market.

The Q-Q plot is skewed to the right, meaning that the peak of our data distribution is farther to the right, with more sparse data points to the left-hand side. This indicates that our median sale price is lower than the model would predict. This supports the discussion above, regarding how homes may be priced lower than their amenities may predict. Once again, it could be related to a sort of ceiling to the housing market.

The scale-location plot angles up sharply towards the right, indicating that the data exhibits heteroscedastity. This means that the variability of the data does not seem to match what the model would predict. Our model predicts that there is a lot more variability in sale price of homes than the data actually has.

The Residuals vs. Leverage plot allows us to identify possible outliers. We have a few data points that are towing the line of Cook's distance. We know we do not have anything outside of this barrier, but we do have a few points that near it. There are two specific cases that are near the line, #4649 and #8377.

xv. Conclusion

Our large sample of data allows for us to have a collection of data points where nothing has undue influence over the model. However, the model exhibits heteroscedastity and autocorrelation between the data points These two factors hurt our model's ability to accurately predict future data points. The Q-Q plot skews to the right, meaning that the distribution of our data is farther to the left with a sparing selection of data to the right hand side. These factors all come together to support anomalies in the housing market itself. The

factors assessed in the model predict sale price up to a point, however, trends in the housing market cause our data points to be correlated with each other over time and physical space. This makes it difficult for the model to accurately predict home prices since they are affected by overarching trends we are not able to materialize in our model.