week 8 Assignment

Mithil Patel

2022-05-15

i) The housing dataset was modified to exclude some categorical columns (geographic location, sales warning, etc) that I think will not help to perform quantitative analysis. Additionally, the dataset is shuffled to avoid autocorrelation from occurring.

ii) To perform multiple regression to predict sale price, I have decided to use square feet lot (sq_ft_lot), total living square feet (square_feet_total_living), building grade, and year built as independent variables. The sq_ft_lot shows the area of the house including the front lawn, parking space, backyard, etc. whereas square_feet_total_living shows the total living space available for tenants to use inside the house, which means larger living space should drive the sale price higher. Newly built homes, together with overall building grade, tend to be more valuable than older homes with poor building grades; therefore, looking at the building_grade and year_built should provide better insight into how those variables influence sales price.

R-squared and Adjusted R-squared

Linear regression model R-squared: 0.01435, Adjusted R-squared: 0.01428

Multiple regression model R-squared: 0.2233, Adjusted R-squared: 0.2230

iii) R-squared estimates the percentage of the variance of the dependent variable that can be explained by the independent variable. In a multiple regression model, an additional independent variable will increase the r-squared value regardless of how insignificant the variable is to the dependent variable. Making assumptions about a multiple regression model based on an r-squared value could lead to invalid conclusions about the relationship between variables. Adjusted r-squared is an updated version of r-squared that accounts for only the variables that improve the model. In other words, the adjusted r-squared will increase if the predictors are significant to the model and are likely influencing the dependent variable. Therefore, an adjusted r-squared is ideal for determining a correlation between variables and how much is affected by an additional independent variable. In our case above, additional predictors certainly helped explain large variations found in Sale Price as the adjusted r-squared increased from 0.01428 to 0.2230. Without the new predictors, the adjusted r-squared value is close to zero indicating that there is virtually no correlation. Increased adjusted r-squared shows a higher proportion of variation being explained by new predictors and provides a slight correlation between the variables. However, the adjusted r-squared is not high enough (.60 or above) to make conclusive arguments and further evaluation is required.

Standarized beta for linear model

```
Call:
lm(formula = `Sale Price` ~ sq_ft_lot, data = new_housing_df)

Standardized Coefficients::
(Intercept)   sq_ft_lot
         NA   0.1198122
```

Standarized beta for multiple regression model

```
Call:
lm(formula = `Sale Price` ~ sq_ft_lot + square_feet_total_living +
    building_grade + year_built, data = new_housing_df)

Standardized Coefficients::
             (Intercept)                sq_ft_lot square_feet_total_living
                      NA               0.04170810               0.34523475
          building_grade               year_built
                0.08816384               0.11059981
```

iv) The standardized beta values provide an insight into how relevant a predictor is in a model by evaluating how a change of one standard deviation in the predictor affects the dependent variable by beta standard deviation. By analyzing the multiple regression chart above, the standardized beta value for square_feet_total_living (0.3452) has the highest degree of importance in the model, which means that increasing square_feet_total_living by one standard deviation will cause the sale price to increase by 0. 3452 standard deviations.

```
                   2.5 %       97.5 %
(Intercept) 6.343730e+05 6.492698e+05
sq_ft_lot   7.291208e-01 9.728641e-01
```

```
                               2.5 %        97.5 %
(Intercept)              -5.919468e+06 -4.378856e+06
sq_ft_lot                 1.820603e-01  4.104214e-01
square_feet_total_living  1.314701e+02  1.506150e+02
building_grade            2.398101e+04  4.127804e+04
year_built                2.203431e+03  2.990998e+03
```

v) Confidence interval for each parameter can tell us if the samples from the model represent the true population values in addition to the direction of the relationship. In our case above, sq_ft_lot variable and square_feet_total_living have the tightest confidence interval, the predictors from our model can be representative of the true population values. Compare to tight predictors, building_grade and year_built have a wider interval which indicates a less representative parameter.

```
Analysis of Variance Table

Model 1: `Sale Price` ~ sq_ft_lot
Model 2: `Sale Price` ~ sq_ft_lot + square_feet_total_living + building_grade +
    year_built
  Res.Df              RSS Df       Sum of Sq    F                Pr(>F)
1  12863 2073376756946868
2  12860 1633912364625849  3 439464392321019 1153 < 0.00000000000000022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

vi) Considering a high F value(1153) and a extremely small p value(2.2e+16), we can conclude that the new model significantly improved the fit of the model compare to the original model.

Number of cases where the residual is bigger than 2 or less than -2:

```
[1] 334
```

Cook's distance, leverage, and covariance

Number of extreme cases:

```
[1] 25
```

Note: Extreme cases where the values lie outside of leverage boundary and has a covariance.ratio value below 0.97

CVR Lower: 1 + [3(4+1)/12865] = 1.00117

CVR Upper: 1- [3(4+1)/12865] = 0.99883

leverage Lower: (4+1)/12865 = 0.000387

Leverage Upper: 3(4+1)/12865 = 0.001166

xi) Based on visual analysis, there are several cases where covariance.ratio values fall outside the CVR boundary. Analyzing the cases where the covariance.ratio values are larger than 1.00117, it is apparent, given their Cook's distance and leverage boundary, that they are seemingly less problematic. However, there are approximately 25 extreme cases where the covariance.ratio and leverage values fall well outside our boundary. Therefore, our diagnostic indicates that the model is a slightly unreliable model that has been influenced by a subset of extreme cases. Note that if we assume extreme cases to have covariance value below 0.92 then the model is fairly reliable with no significant influential data points.

Assumption of independence

```
 lag Autocorrelation D-W Statistic p-value
   1     0.003184265      1.993615    0.77
 Alternative hypothesis: rho != 0
```

xii) As shown above, the D-W statistic is close to 2 and p-value is significantly higher than 0.05; therefore, there are conditions for met and we can safely assume the model is independent of errors.

Note: I had to shuffle the original housing dataset because the D-W value was 0.54 and p-value is equal to zero, indicating a positive autocorrelation. Positive autocorrelation occurs when the past data follows a pattern similar to current data. The problem was solved when the dataset was shuffled.

Assumption of no multicollinearity

VIF:

```
     sq_ft_lot square_feet_total_living           building_grade
      1.113594                 2.365744                 2.353068
    year_built
      1.211716
```

Tolerance:

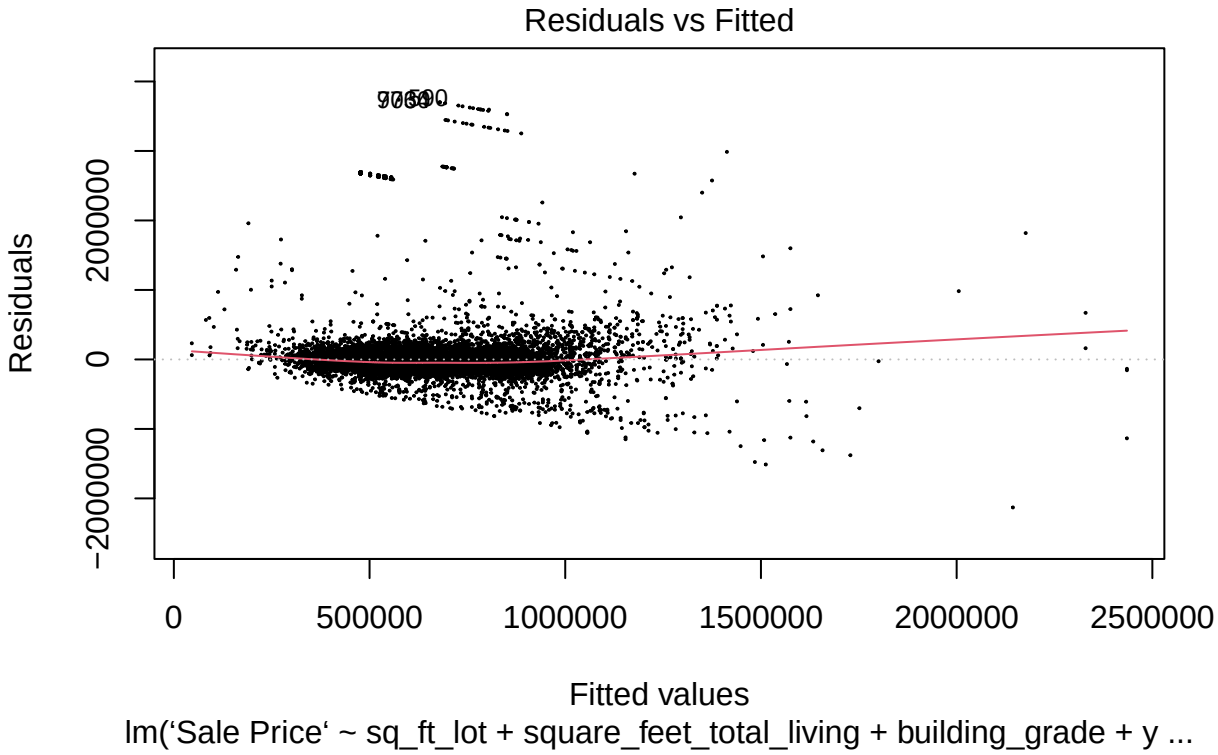
```
     sq_ft_lot square_feet_total_living           building_grade
     0.8979930                0.4226999                0.4249772
    year_built
     0.8252756
```

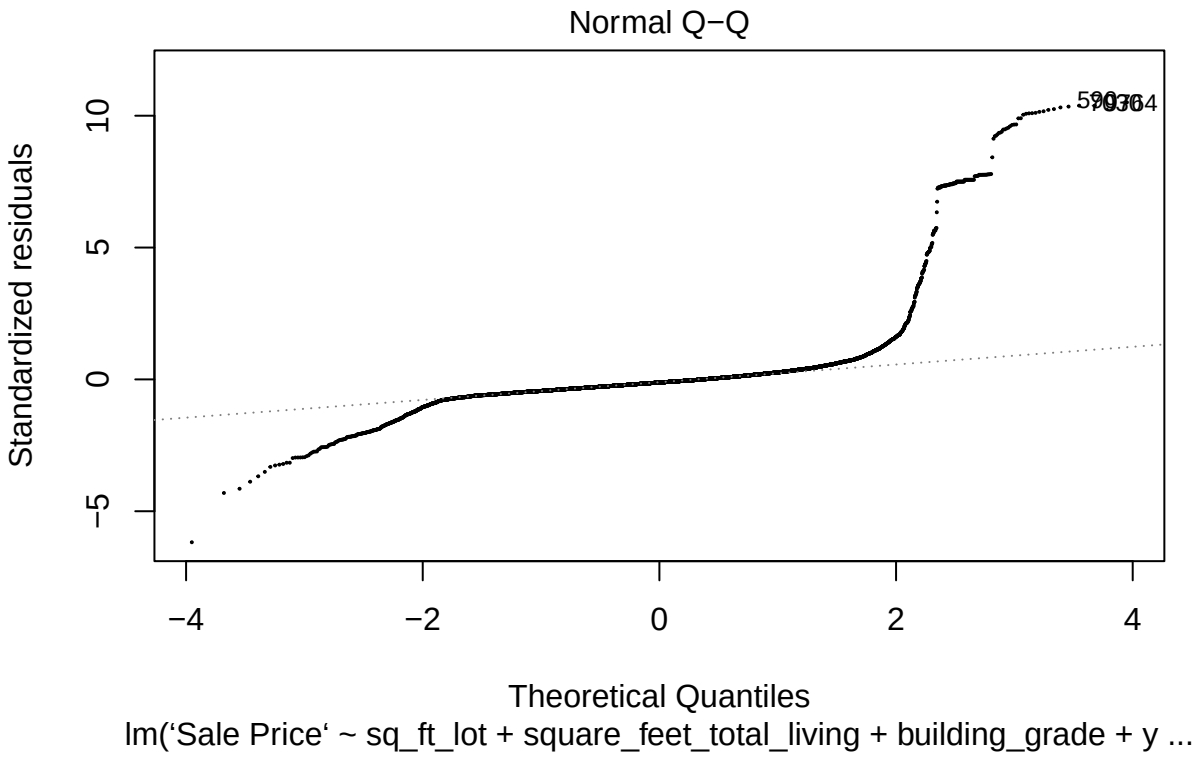
VIF mean:

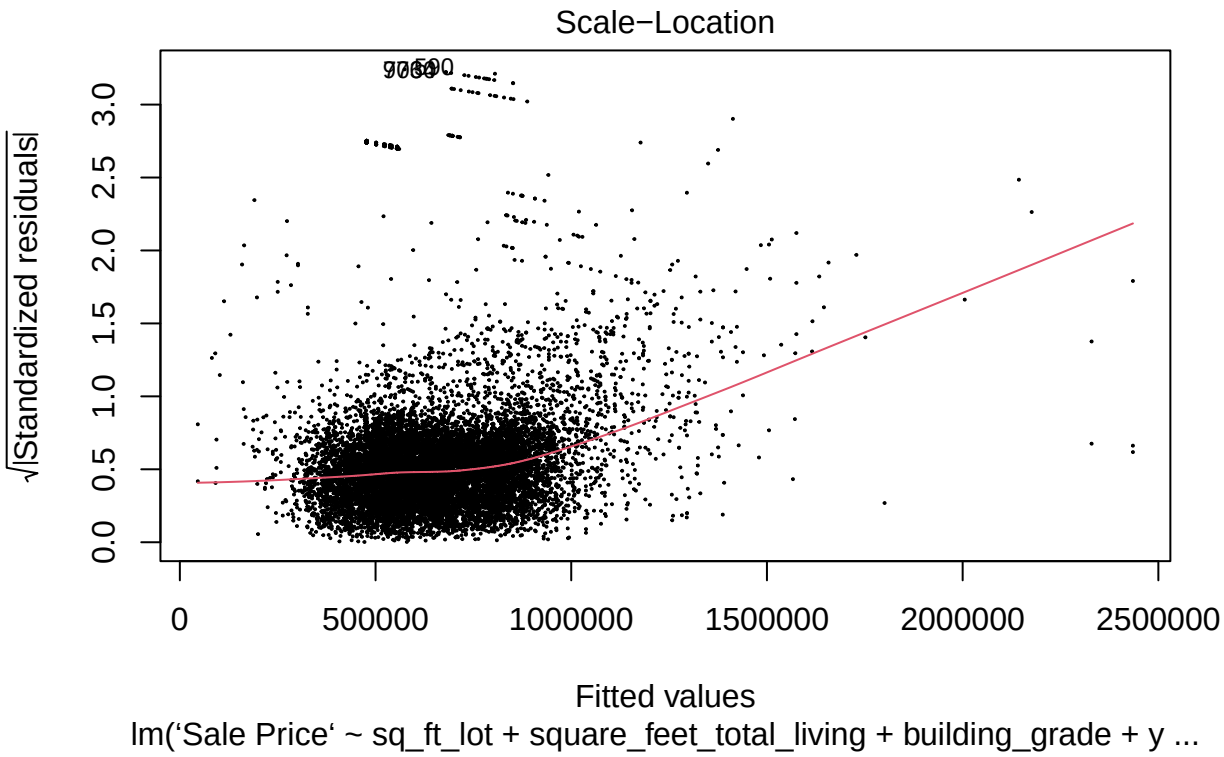
```
[1] 1.761031
```

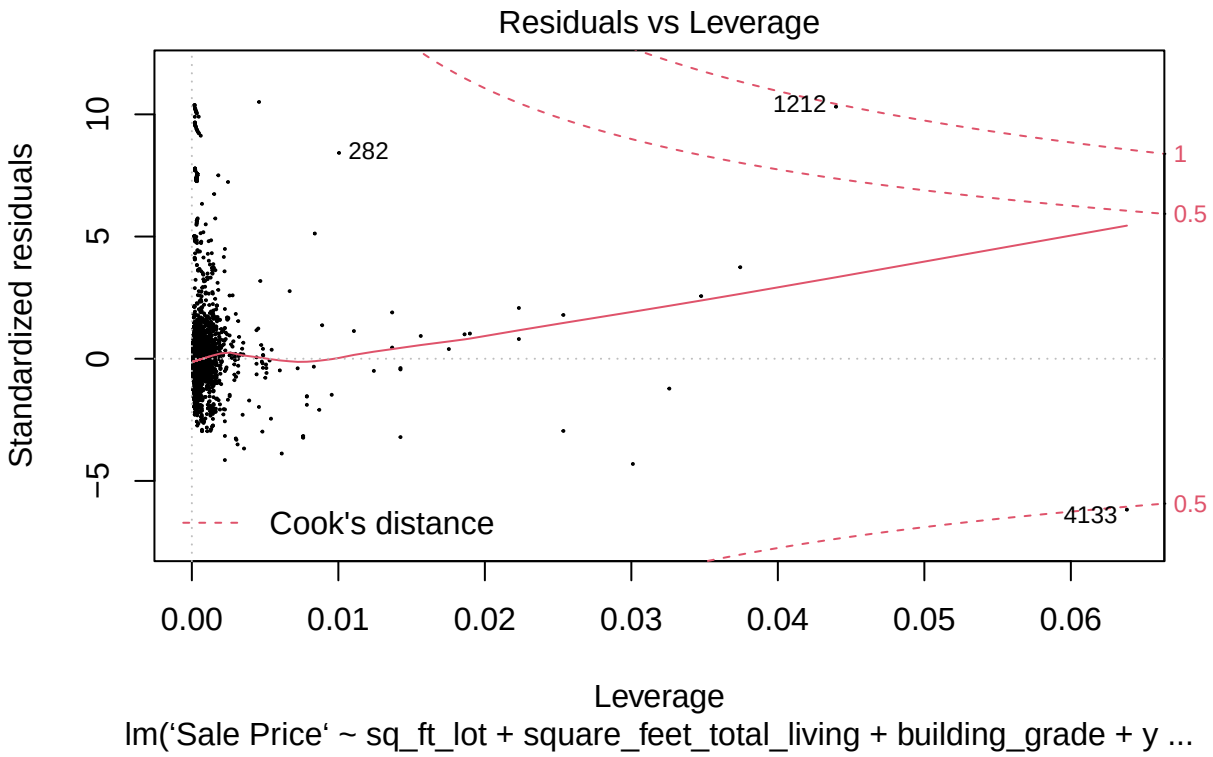
xiii) The vif and the tolerance value for each predictor is above 2.5 and 0.2, respectively. Additionally, the mean vif is relatively close to 1, thus further providing more evidence that the conditions for no multicollinearity are met. In our case, no multicollinearity suggests that the predictors are not highly correlated with one another and they can accurately provide useful information about the dependent variable.

xiv)




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 + building_grade + y ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
500000
1000000
1500000
2000000
2500000
Fitted values
lm(‘Sale Price‘ ~ sq_ft_lot + square_feet_total_living + building_grade + y ...

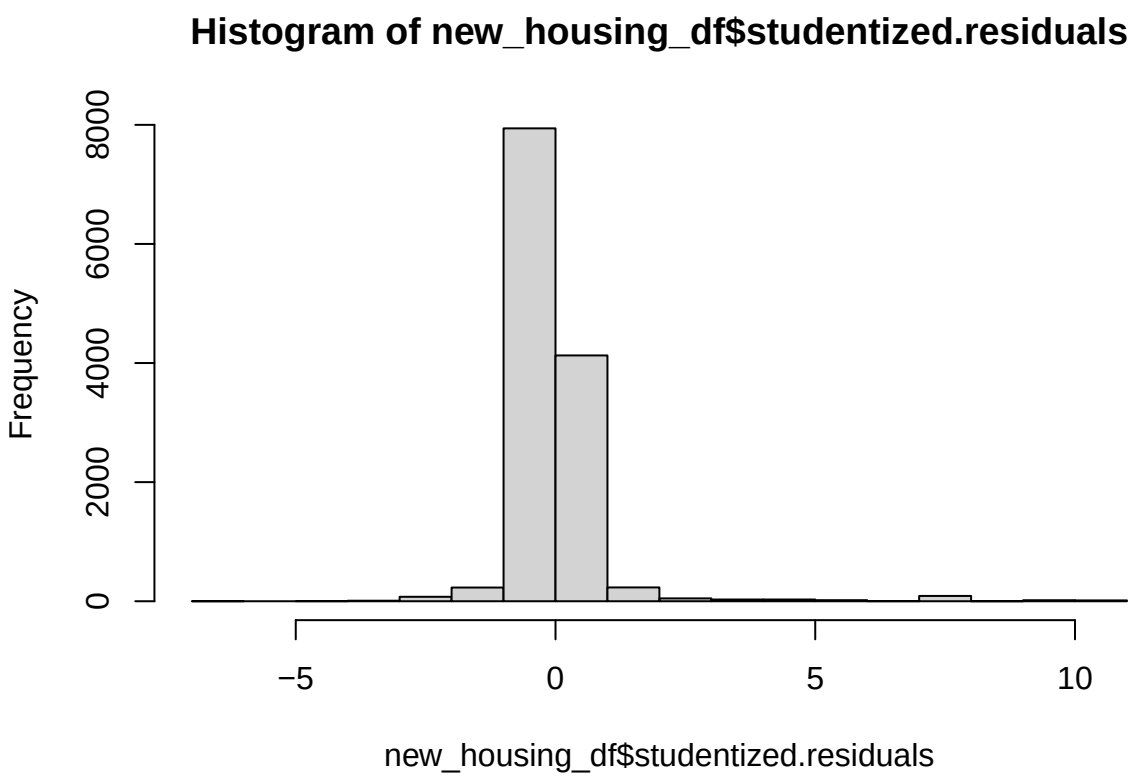


Plot 1: The Residual vs Fitted graph helps us determine whether the model we have used is appropriate for the given dataset. By visually analyzing the Residual vs Fitted plot, we can see the data points are clumped together with near-zero residuals and seem to scatter as the predicted sale price increases. As a result, there exists heteroscedasticity in our data and we must look to add more variables to consistently make an accurate prediction. Visually, there appears to be an equal number of data points above the residual line as well as below the residual linear, indicating linearity in our data. linearity suggests that there is a linear relationship between the sale price and square_feet_total_living, sq_ft_lot, building_grade, etc.

Plot 2: A Q-Q plot can provide insight into whether the residual is normally distributed and the general shape of the distribution. In our case, the two tails pointing in opposite direction indicate we have over-dispersed data. The residual distribution will have a leptokurtic shape with positive kurtosis, which means the distribution has a fatter tail compared to the normal distribution. The tails also suggest a large number of outliers present in our distribution.

Plot 3: The Scale-location plot shows that the data points are clustered together near 0.5 and 2.5 of standardized residuals and the red line is not horizontal; thus, it is a clear indication that our data violates homoscedasticity. The observation provides more evidence to support the analysis in plot 1.

Plot 4: The Residual vs Leverage plot shows any influential data point that could have a significant effect on our linear model. The plot above shows several influential data points (labeled: '5871', '1269','3792', etc.) that lie outside our Cook's distance line. These data points have a high residual and high leverage so removing them will have a considerable impact on the coefficients of our model.



Histogram: The residual histogram above shows a leptokurtic distribution with an increased number of outliers, which matches the prediction made by visually analyzing the Q-Q graph in plot 2.

xv) Since one of the assumptions (homoscedasticity) was not met, the regression model is said to be biased. Therefore, the prediction from our model cannot be generalized to the population model. However, if the regression model was unbiased, we can say that on average the regression model from our sample data can be a good representation of the population model.