Regression Analysis Part 4

Cody

2024-06-03

Refitting a new base model

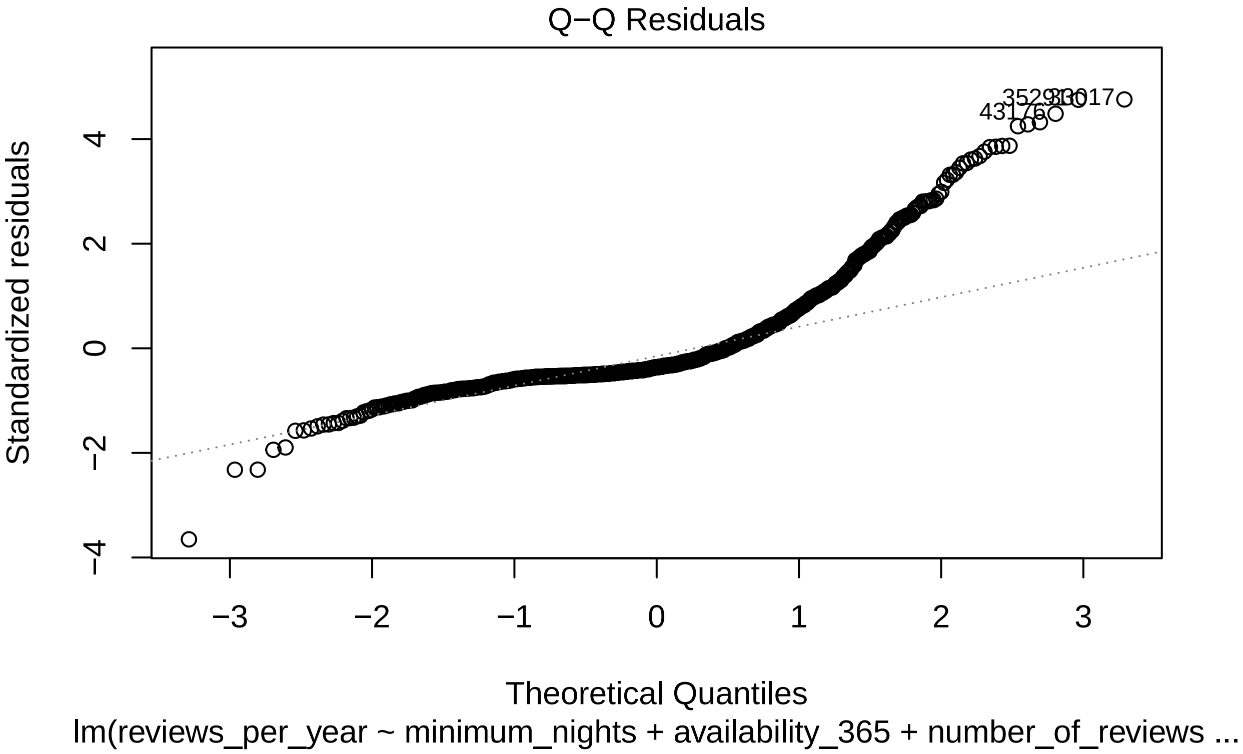
```
Call:
lm(formula = reviews_per_year ~ minimum_nights + availability_365 + 
    number_of_reviews + neighbourhood_group, data = NYC)

Coefficients:
                 (Intercept)                minimum_nights  
                    15.54790                      -0.11136  
            availability_365             number_of_reviews  
                     0.01052                       0.22083  
 neighbourhood_groupBrooklyn  neighbourhood_groupManhattan  
                    -6.99795                      -7.24796  
   neighbourhood_groupQueens  
                    -1.88167  
```

- The base linear model parameters are shown above

Assess base model assumption violations.

Normality

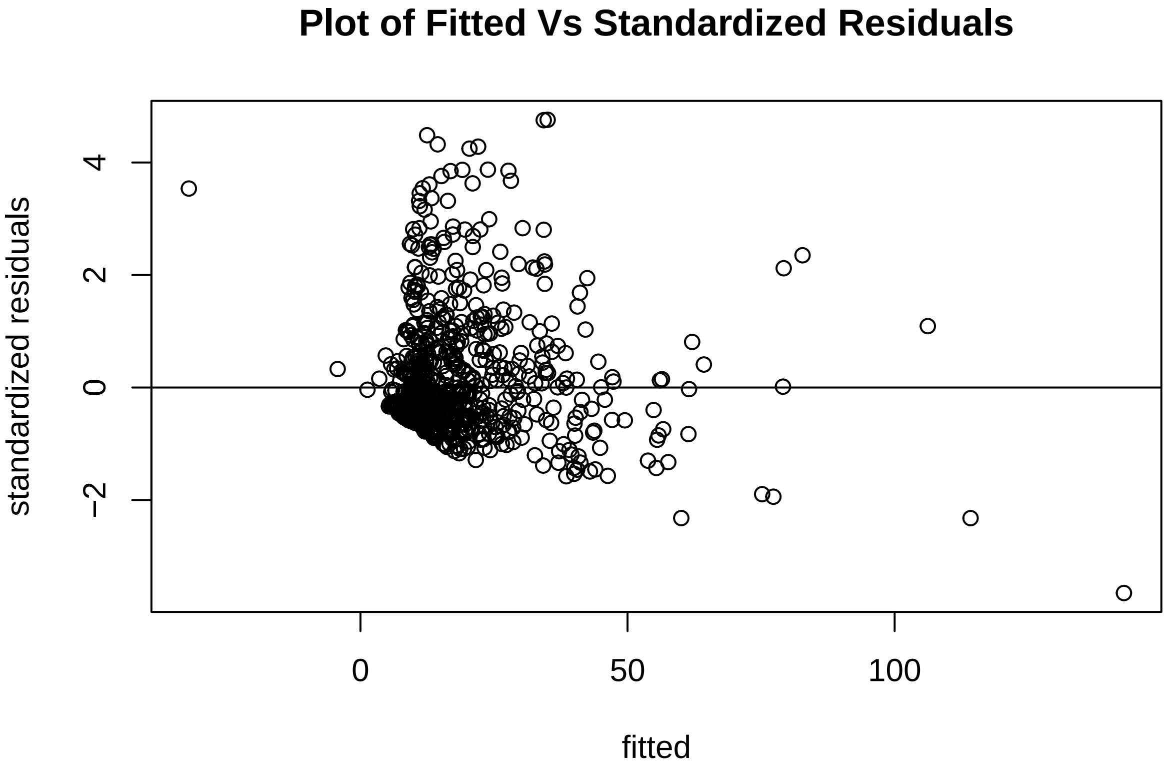


```
    Shapiro-Wilk normality test

data:  rstandard(NYC.lm)
W = 0.79358, p-value < 2.2e-16
```

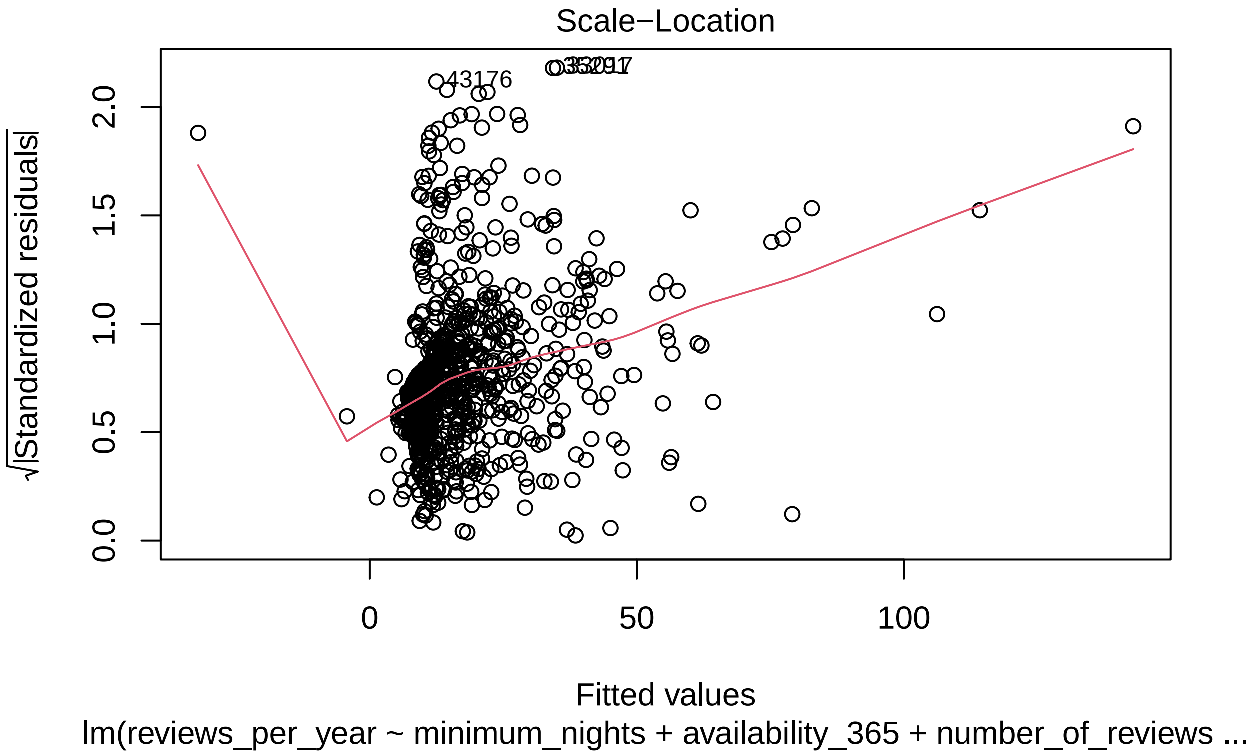
- Residuals are clearly non-normally distributed as shown by the qqplot. There is a right skew in the data. The standardized residuals tend to fall higher than predicted by the current regression model.
- The Shapiro-Wilkes test also verifies what we observed in the QQ plot, the data does not appear to be normally distributed.

Model Classification



- The fitted vs. residuals plot shows that there is not an even scattering of standardized residuals across the horizontal line drawn at zero. The residuals look like they skew in the positive direction.
- The residuals also appear to fan out as the fitted values get larger.
- This indicates an adjustment needs to be made in the base model for a better fit

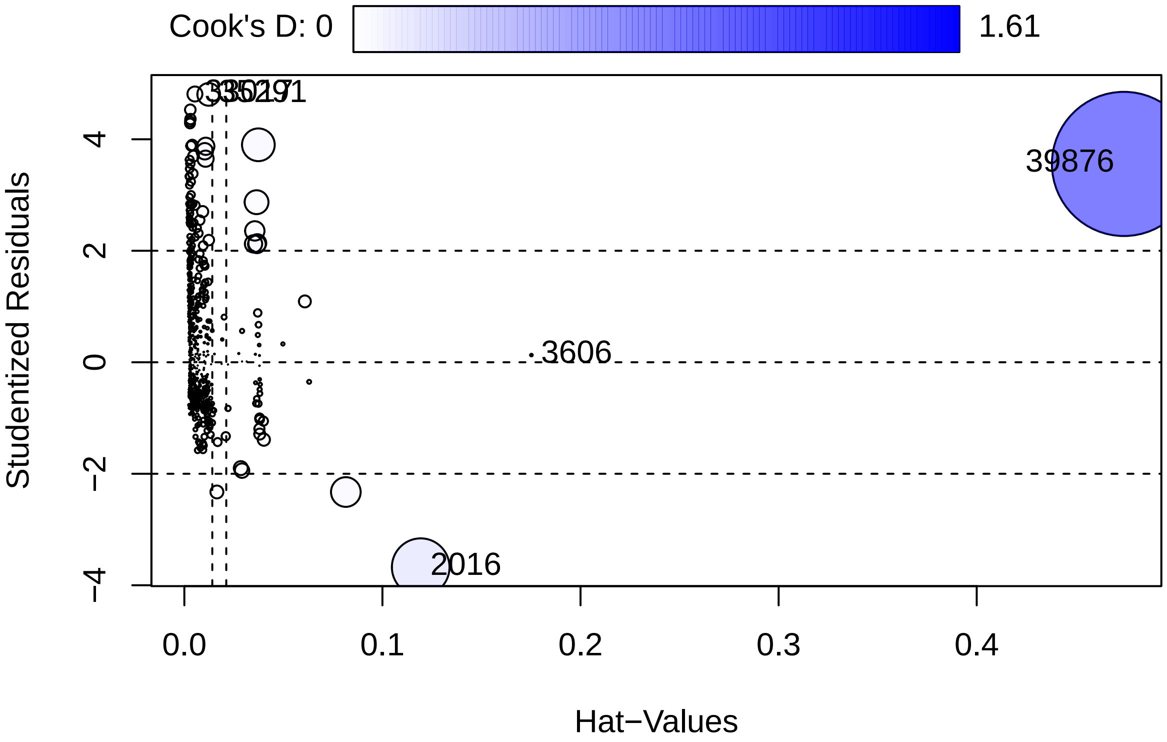
Homoscedacity assumption



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 111.1251, Df = 1, p = < 2.22e-16
```

- The Scale-Location plot shows additional verification for heteroscedacity that was hinted at in the last plot. There is a steady increasing linear trend in the absolute value of the standardized residuals. This means the residuals fan out as the fitted values become larger.
- The NVC test having a very low p value is one additional indicator that equivalent variance is not a safe assumption here

Influential points



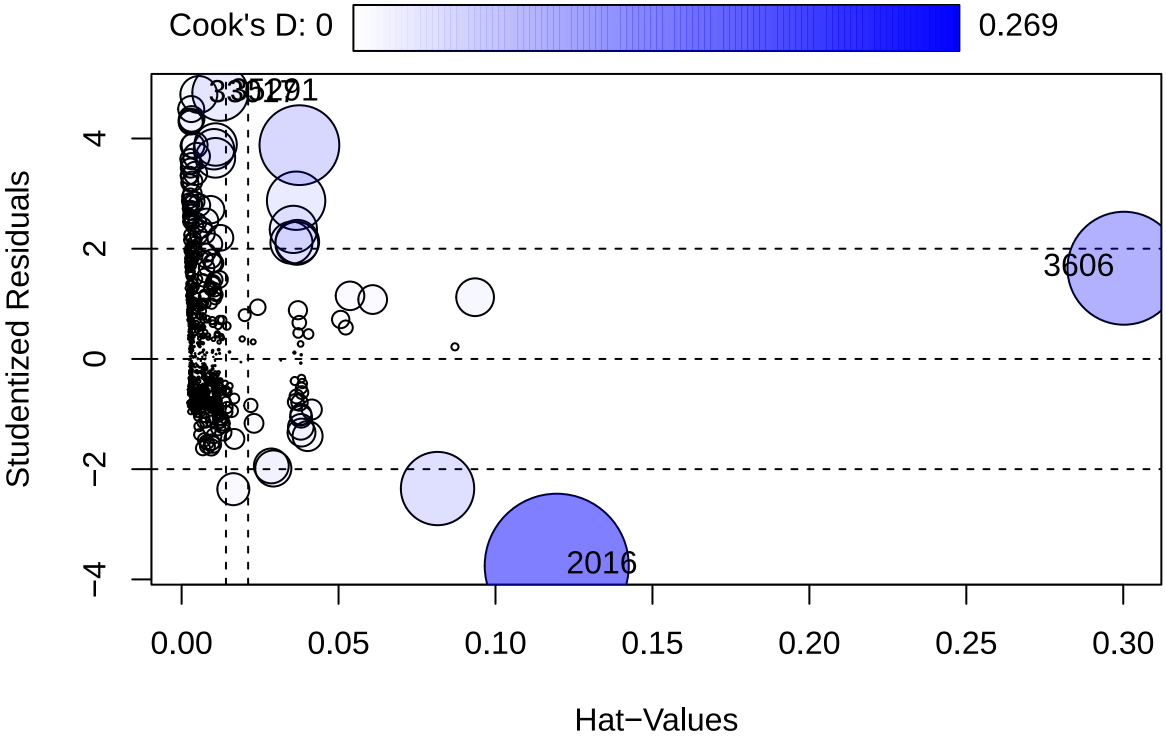
```
        StudRes         Hat        CookD
35291  4.8054255 0.012307744 0.0402058732
2016  -3.6760895 0.119319311 0.2582758358
3606   0.1288235 0.175218226 0.0005041576
33017  4.8119216 0.005338263 0.0173621590
39876  3.5577209 0.474323049 1.6124689207
```

- Clearly point 39876 shown above is an influential point, on the plot it is shown that its cook's distance is far greater than any other points in the dataset.

Addressing problems diagnosed with base model

Outlier

Removing the extremely influential point

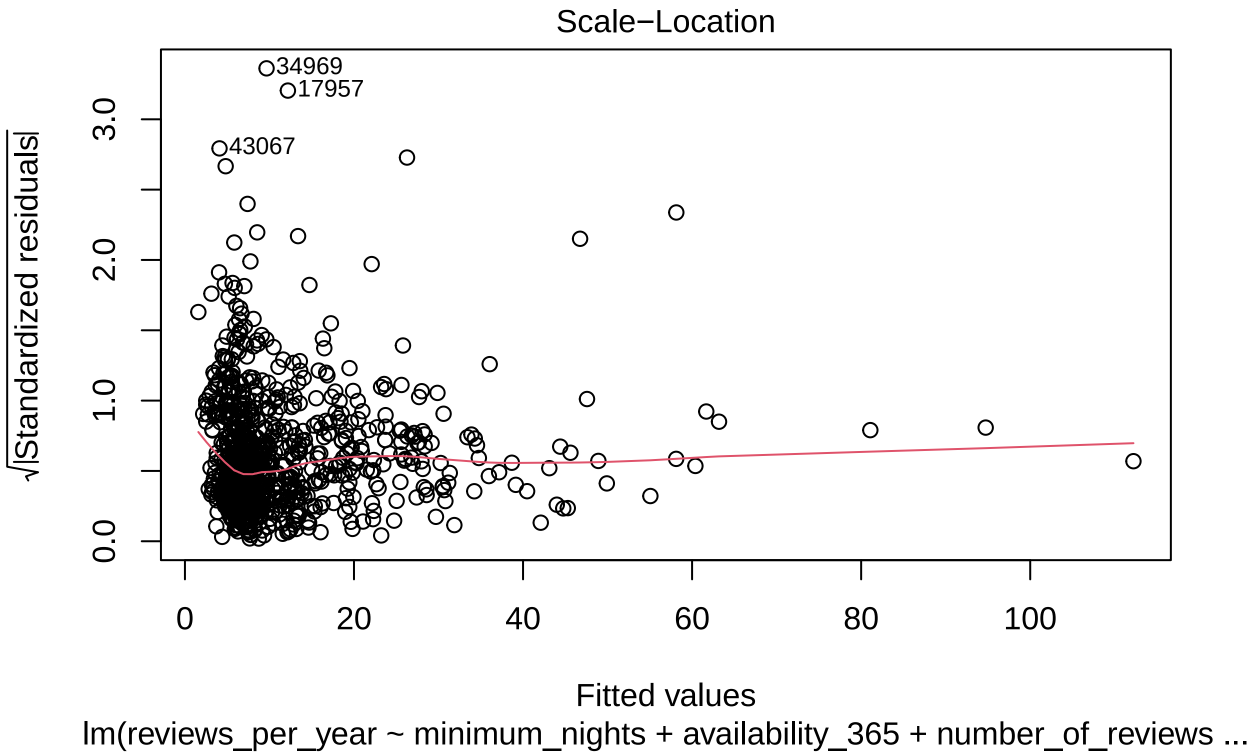


```
        StudRes         Hat      CookD
35291  4.828080 0.012310756 0.04058621
2016  -3.752319 0.119512591 0.26943694
3606   1.645848 0.300221370 0.16573290
33017  4.798015 0.005476203 0.01771241
```

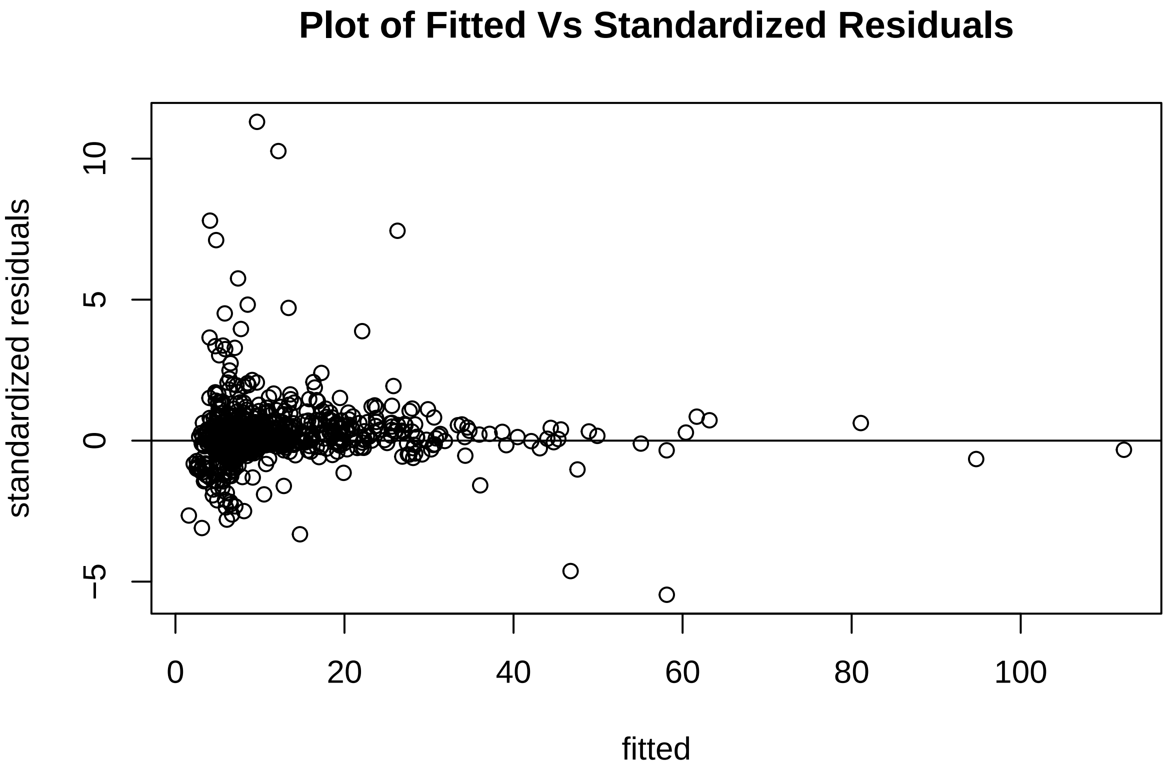
- Removing point 39876 as its cook's distance is incredibly large at 1.61 compared to other data. The largest cooks distance is now only 0.2694369.

Variance and model classification issue

Weighted least Squares approach to residuals



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.04022649, Df = 1, p = 0.84104
```



- Weighting based on the required minimum nights to stay drastically improved the model's homoscedacity as shown by the horizontal trendline in the scale-location plot. The NCV test results back up this claim, showing a very high p value that points towards a valid homoscedacity assumption.
- Weighting the residuals also appeared to improve some of the positive skewness seen in the base model's standardized residuals vs fitted values plot.

Summary of transformations done to arrive at final model

- Removing outlier. The removal of the very influential point didn't have as big of an impact as I was hoping for on the model fit, the effect was likely diluted by having such a large sample size.
- Weighted Least Squares approach to residuals. I wanted to avoid using a transformation on the response variable if possible to reduce any potential impacts on the ability to interpret the model. This made the weighted least squares approach make sense. I decided to weight the model residuals based on the predictor that indicates the required minimum nights of stay in an airbnb.

Final assessment

Summary of base model:

```
Call:
lm(formula = reviews_per_year ~ minimum_nights + availability_365 +
    number_of_reviews + neighbourhood_group, data = NYC)

Residuals:
    Min      1Q  Median      3Q     Max 
-52.090  -8.050  -5.504   3.479  72.115 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  15.547899   2.964978   5.244 1.92e-07 ***
minimum_nights               -0.111358   0.029105  -3.826 0.000138 ***
availability_365              0.010523   0.003912   2.690 0.007263 ** 
number_of_reviews             0.220832   0.009313  23.713  < 2e-16 ***
neighbourhood_groupBrooklyn  -6.997951   2.979934  -2.348 0.019053 *  
neighbourhood_groupManhattan -7.247964   2.993194  -2.421 0.015637 *  
neighbourhood_groupQueens    -1.881665   3.223636  -0.584 0.559550    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.19 on 985 degrees of freedom
Multiple R-squared:  0.4015,    Adjusted R-squared:  0.3979 
F-statistic: 110.2 on 6 and 985 DF,  p-value: < 2.2e-16
```

Summary of modified model:

```

Call:
lm(formula = reviews_per_year ~ minimum_nights + availability_365 + 
    number_of_reviews + neighbourhood_group, data = NYC, subset = -influentialIndicies, 
    weights = minimum_nights^2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-281.08  -13.47   -3.11   20.82  732.71 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   4.630667   0.843456   5.490 5.11e-08 ***
minimum_nights               -0.030308   0.005180  -5.851 6.66e-09 ***
availability_365             -0.004874   0.001559  -3.126  0.00182 ** 
number_of_reviews             0.183289   0.002945  62.235  < 2e-16 ***
neighbourhood_groupBrooklyn   2.214977   0.788344   2.810  0.00506 ** 
neighbourhood_groupManhattan  0.392549   0.742566   0.529  0.59717    
neighbourhood_groupQueens     1.744435   1.089266   1.601  0.10959    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.03 on 984 degrees of freedom
Multiple R-squared:  0.9532,    Adjusted R-squared:  0.9529 
F-statistic:  3340 on 6 and 984 DF,  p-value: < 2.2e-16
```

Comparison of other model parameters:

```
The adjusted R squared value of the base model is 0.3979 and the adjusted R squared value of the modifie
```

```
The R squared value of the base model is 0.4015 and the R squared value of the modified model is: 0.9532
```

```
The standard error of the base model is 15.1949 and the standard error of the modified model is: 66.0302
```

Some final comments:

- Weighting residuals based on the minimum nights required for a stay had more of an impact on the model fit and heteroscedacity than I was originally anticipating. There was a drastic improvement to various model fitting parameters (54.48% increase in R squared values). One hypothesis i have for why this is the case is because some places that require long term stays for guests might have more of an ability to make a large impression on a guest than quick stays, resulting in a very high hit rate for reviews.

- The clear sacrifice made here is a standard error that is approximately 4.5 times bigger than the base model's standard error. This is a worthwhile sacrifice as the new model does a much better job at meeting the assumptions for successful regression while still maintaining ease of interpretation.