Regression Analysis Part 3

Cody

2024-06-07

1 Fit the multiple regression model, using two of your categorical predictors and 2-3 numerical predicitors, and include all pairwise interaction terms

Initial linear model summary:

```
Call:
lm(formula = reviews_per_year ~ (minimum_nights + number_of_reviews + 
    neighbourhood_group + room_type)^2, data = NYC)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.127  -8.625  -5.244   3.982  71.835 

Coefficients:
                                                  Estimate Std. Error
(Intercept)                                      3.563e+01  5.913e+00
minimum_nights                                  -4.806e-01  1.780e-01
number_of_reviews                                2.933e-01  1.034e-01
neighbourhood_groupBrooklyn                     -2.587e+01  6.035e+00
neighbourhood_groupManhattan                    -2.517e+01  6.013e+00
neighbourhood_groupQueens                       -1.935e+01  6.513e+00
room_typePrivate room                           -2.759e+01  6.424e+00
room_typeShared room                            -2.351e+01  1.209e+01
minimum_nights:number_of_reviews                 4.519e-04  4.166e-04
minimum_nights:neighbourhood_groupBrooklyn       1.827e-01  2.014e-01
minimum_nights:neighbourhood_groupManhattan      2.528e-01  1.905e-01
minimum_nights:neighbourhood_groupQueens        -1.606e-01  2.773e-01
minimum_nights:room_typePrivate room             2.022e-01  7.518e-02
minimum_nights:room_typeShared room             -1.086e-01  3.719e-01
number_of_reviews:neighbourhood_groupBrooklyn   -5.644e-02  1.034e-01
number_of_reviews:neighbourhood_groupManhattan  -6.161e-02  1.032e-01
number_of_reviews:neighbourhood_groupQueens     -1.866e-02  1.046e-01
number_of_reviews:room_typePrivate room         -2.872e-02  2.122e-02
number_of_reviews:room_typeShared room           2.145e-01  1.421e-01
neighbourhood_groupBrooklyn:room_typePrivate room  2.805e+01  6.568e+00
neighbourhood_groupManhattan:room_typePrivate room 2.529e+01  6.597e+00
neighbourhood_groupQueens:room_typePrivate room    2.612e+01  7.082e+00
neighbourhood_groupBrooklyn:room_typeShared room   2.502e+01  1.333e+01
neighbourhood_groupManhattan:room_typeShared room  2.507e+01  1.343e+01
neighbourhood_groupQueens:room_typeShared room     2.326e+01  1.411e+01
                                                 t value Pr(>|t|)
(Intercept)                                        6.025 2.40e-09 ***
```

```
minimum_nights                                        -2.700 0.007051 **
number_of_reviews                                      2.838 0.004641 **
neighbourhood_groupBrooklyn                           -4.287 1.99e-05 ***
neighbourhood_groupManhattan                          -4.186 3.10e-05 ***
neighbourhood_groupQueens                             -2.971 0.003037 **
room_typePrivate room                                 -4.295 1.92e-05 ***
room_typeShared room                                  -1.945 0.052096 .
minimum_nights:number_of_reviews                       1.085 0.278340
minimum_nights:neighbourhood_groupBrooklyn             0.907 0.364599
minimum_nights:neighbourhood_groupManhattan            1.327 0.184712
minimum_nights:neighbourhood_groupQueens              -0.579 0.562723
minimum_nights:room_typePrivate room                   2.689 0.007280 **
minimum_nights:room_typeShared room                   -0.292 0.770233
number_of_reviews:neighbourhood_groupBrooklyn         -0.546 0.585214
number_of_reviews:neighbourhood_groupManhattan        -0.597 0.550584
number_of_reviews:neighbourhood_groupQueens           -0.178 0.858395
number_of_reviews:room_typePrivate room               -1.354 0.176201
number_of_reviews:room_typeShared room                 1.510 0.131462
neighbourhood_groupBrooklyn:room_typePrivate room      4.270 2.14e-05 ***
neighbourhood_groupManhattan:room_typePrivate room     3.833 0.000135 ***
neighbourhood_groupQueens:room_typePrivate room        3.689 0.000238 ***
neighbourhood_groupBrooklyn:room_typeShared room       1.876 0.060950 .
neighbourhood_groupManhattan:room_typeShared room      1.867 0.062227 .
neighbourhood_groupQueens:room_typeShared room         1.648 0.099655 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.03 on 967 degrees of freedom
Multiple R-squared:  0.4254,    Adjusted R-squared:  0.4112
F-statistic: 29.83 on 24 and 967 DF,  p-value: < 2.2e-16
```

2 Assess whether or not any of the interaction terms should be included in the model by investigating the p-values associated with each interaction term

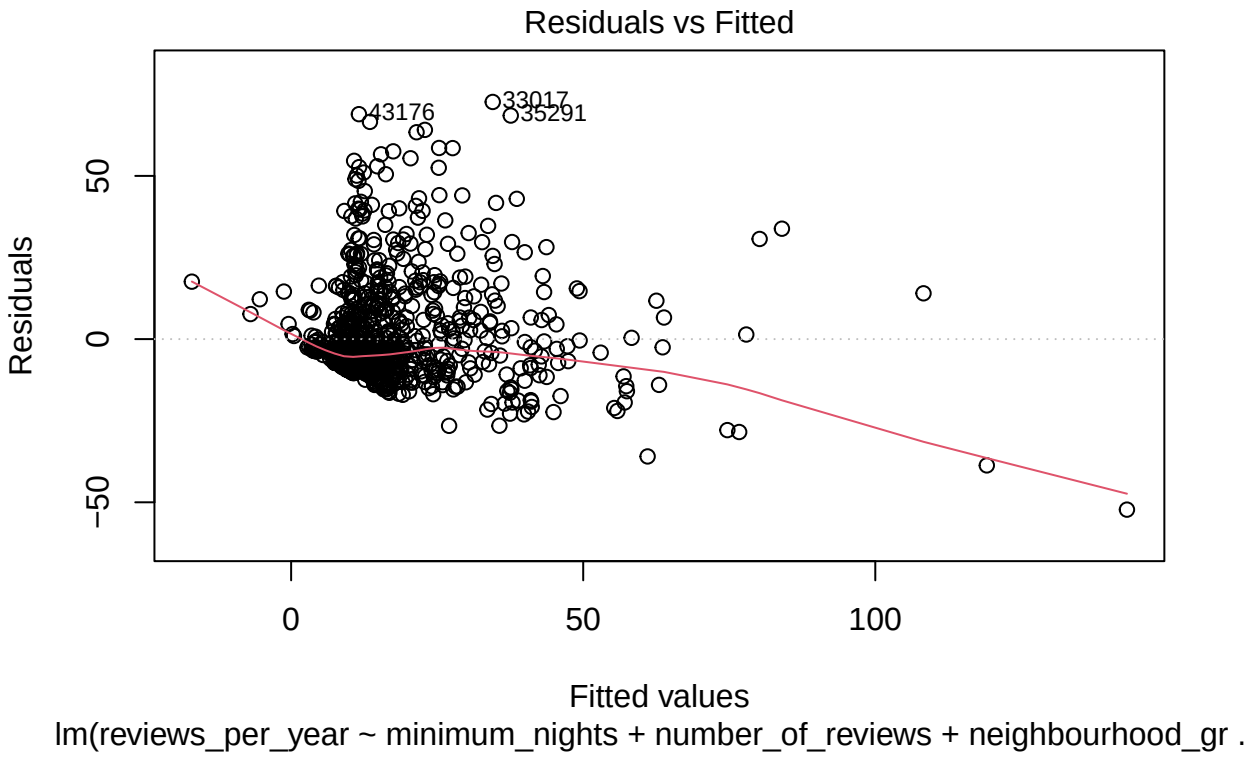
Final linear model summary:

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

Residuals:
    Min      1Q  Median      3Q     Max
-52.240  -8.673  -5.089   4.193  72.649

Coefficients:
                                                Estimate Std. Error
(Intercept)                                    33.833017   4.814931
minimum_nights                                 -0.233935   0.047541
number_of_reviews                               0.228648   0.009212
neighbourhood_groupBrooklyn                   -24.010298   4.889548
neighbourhood_groupManhattan                  -23.213259   4.876184
neighbourhood_groupQueens                     -17.496177   5.339043
```

```
room_typePrivate room                                -25.262136   6.113565
room_typeShared room                                 -21.298280  11.706250
minimum_nights:room_typePrivate room                   0.198620   0.060099
minimum_nights:room_typeShared room                   -0.218098   0.362444
neighbourhood_groupBrooklyn:room_typePrivate room     24.884467   6.260990
neighbourhood_groupManhattan:room_typePrivate room    21.941027   6.279084
neighbourhood_groupQueens:room_typePrivate room       23.012767   6.800664
neighbourhood_groupBrooklyn:room_typeShared room      24.379675  12.958958
neighbourhood_groupManhattan:room_typeShared room     27.713958  12.761505
neighbourhood_groupQueens:room_typeShared room        23.088923  13.709781
                                                     t value Pr(>|t|)
(Intercept)                                            7.027 3.97e-12 ***
minimum_nights                                        -4.921 1.01e-06 ***
number_of_reviews                                     24.821  < 2e-16 ***
neighbourhood_groupBrooklyn                           -4.911 1.06e-06 ***
neighbourhood_groupManhattan                          -4.761 2.22e-06 ***
neighbourhood_groupQueens                             -3.277 0.001086 **
room_typePrivate room                                 -4.132 3.90e-05 ***
room_typeShared room                                  -1.819 0.069158 .
minimum_nights:room_typePrivate room                   3.305 0.000985 ***
minimum_nights:room_typeShared room                   -0.602 0.547485
neighbourhood_groupBrooklyn:room_typePrivate room      3.975 7.57e-05 ***
neighbourhood_groupManhattan:room_typePrivate room     3.494 0.000497 ***
neighbourhood_groupQueens:room_typePrivate room        3.384 0.000743 ***
neighbourhood_groupBrooklyn:room_typeShared room       1.881 0.060229 .
neighbourhood_groupManhattan:room_typeShared room      2.172 0.030120 *
neighbourhood_groupQueens:room_typeShared room         1.684 0.092478 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.08 on 976 degrees of freedom
Multiple R-squared:  0.4158,    Adjusted R-squared:  0.4068
F-statistic: 46.31 on 15 and 976 DF,  p-value: < 2.2e-16
```

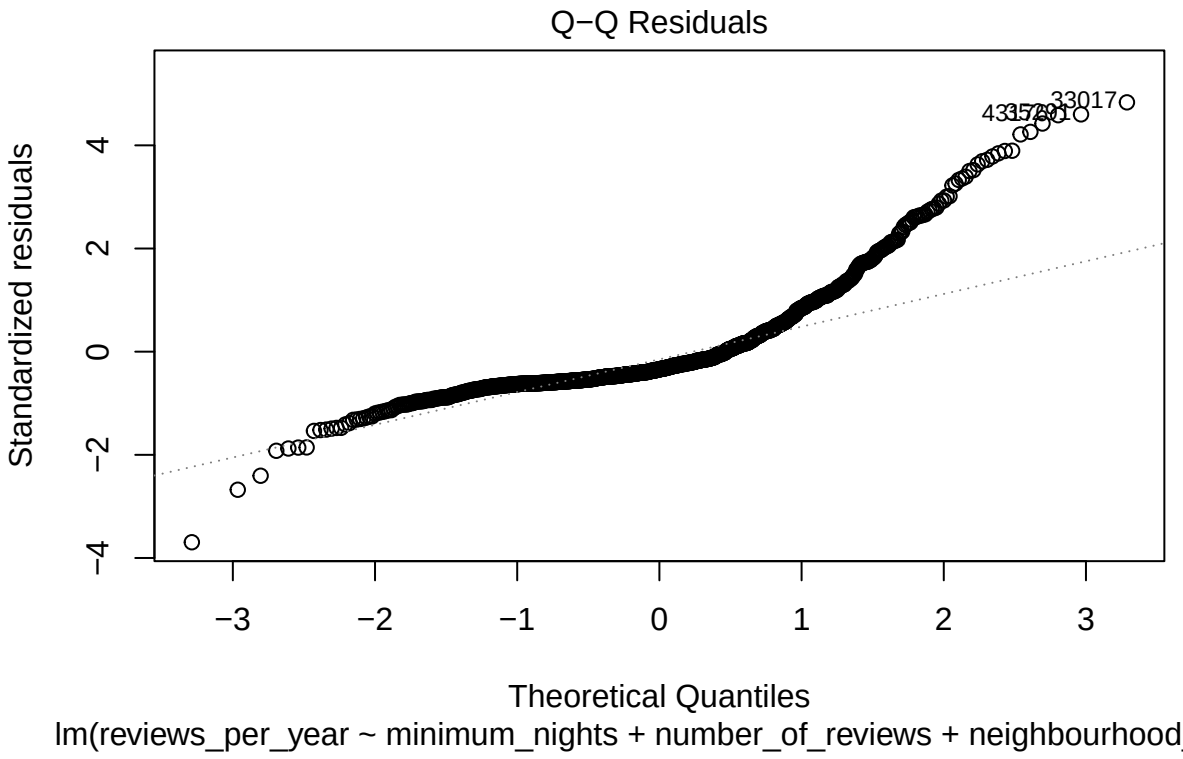
3 Create a set of diagnostic plots and interpret them

How does the linearity assumption appear to be met?



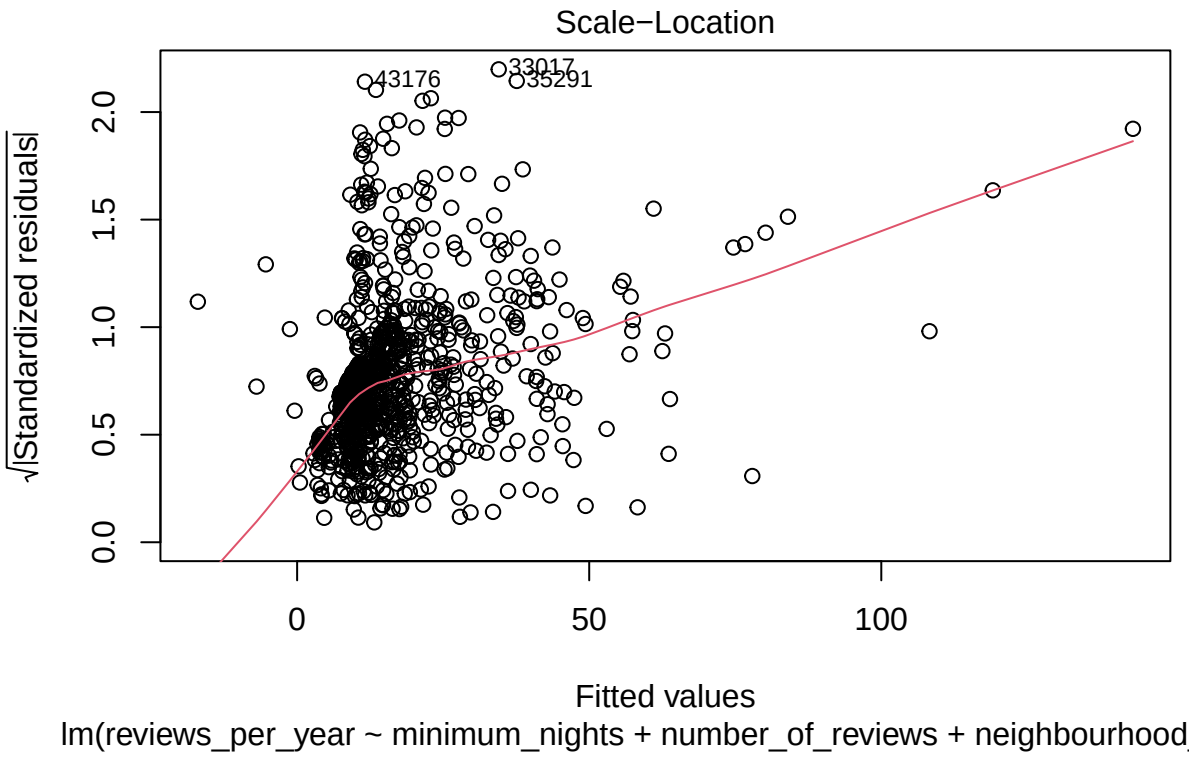
- The linearity assumption does not appear to be a perfect assumption as the residuals are not evenly scattered around the fitted line.

How does the normality assumption appear to be met?



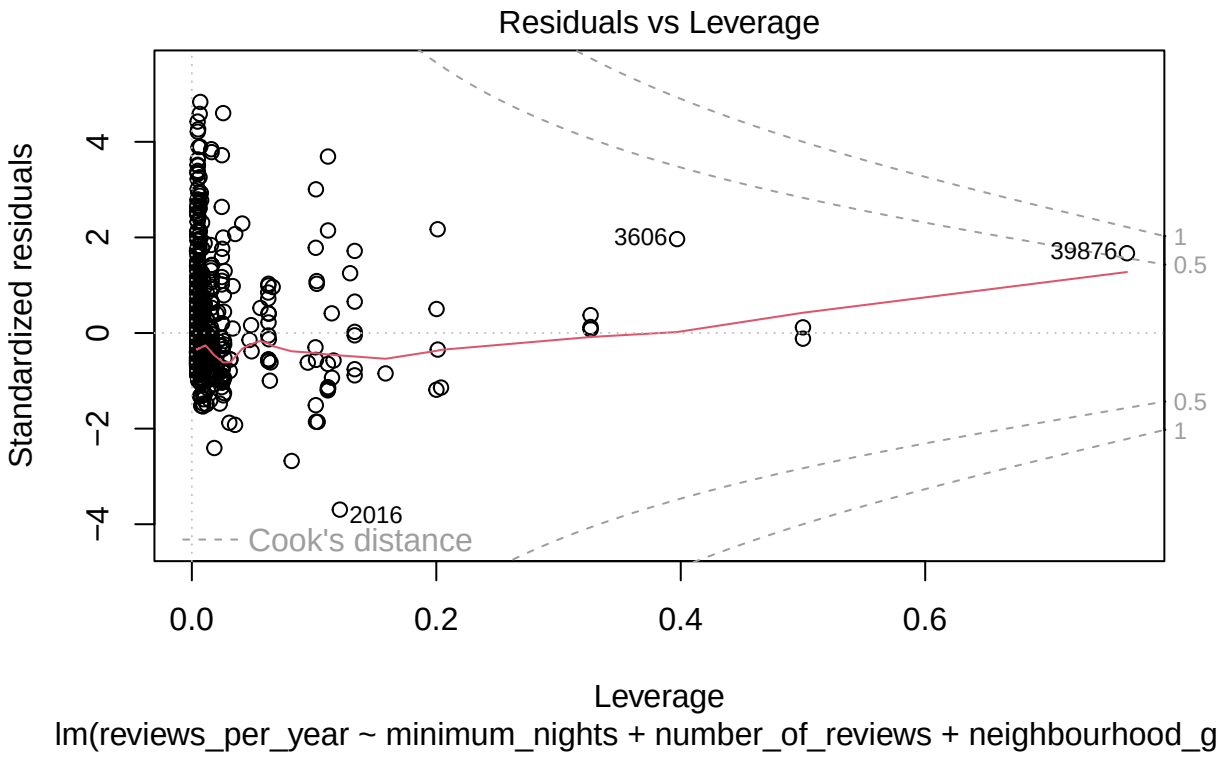
- The data skews towards extremes at the higher end of the theoretical quantiles, normality is not met.

How does the homoscedasticity assumption appear to be met?



- Homoscedasticity does not appear to be met as the scale-location plot trend line increases with increasing fitted values, which indicates a increasingly large variance with increasingly large predictions.

Are there any influential observations in your model? How do you know they are influential?



- I consider Point 39876 an influential point because the cook's distance is greater than 0.5, there are no other points with a cook's distance greater than this.

4 Interpret the presence of any interaction terms in the model:

If there is an interaction between a numerical predictor and a categorical predictor, provide a verbal interpretation of the associated regression coefficients.

- There is an interaction between the room type and the minimum nights of stay. For a stay in a private room you, if you increase the minimum nights required for a stay there is a positive impact on the review frequency. This association is less clear for shared rooms.

If there is an interaction between two categorical predictors, provide verbal interpretations of the associated regression coefficients.

- There is an interaction between neighborhood group and room type. The burrough the Airbnb is located in combined with whether the room is shared or private has an impact on the frequency of reviews. A shared room in Manhattan for example will raise the review frequency more than a private room in Manhattan. And a private room in Brooklyn will raise the review frequency more than a private room in Manhattan.

5 Are there any categories in one of your categorical variables which could be combined together?

Perform a set of simultaneous hypothesis tests associated with the different levels of one of your categorical variables (the multcomp or car package will be very useful in this regard).

```
Loading required package: mvtnorm
```

```
Warning: package 'mvtnorm' was built under R version 4.3.1

Loading required package: survival

Loading required package: TH.data

Loading required package: MASS


Attaching package: 'TH.data'

The following object is masked from 'package:MASS':

    geyser


     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = reviews_per_year ~ minimum_nights + number_of_reviews +
    neighbourhood_group + room_type + minimum_nights:room_type +
    neighbourhood_group:room_type, data = NYC)

Linear Hypotheses:
                                                                 Estimate
neighbourhood_groupManhattan - neighbourhood_groupBrooklyn == 0    0.797
                                                                 Std. Error
neighbourhood_groupManhattan - neighbourhood_groupBrooklyn == 0      1.422
                                                                 t value
neighbourhood_groupManhattan - neighbourhood_groupBrooklyn == 0   0.561
                                                                 Pr(>|t|)
neighbourhood_groupManhattan - neighbourhood_groupBrooklyn == 0    0.575
(Adjusted p values reported -- single-step method)
```

- Combining burroughs Manhattan and Brooklyn together does not appear to have a significant impact on the model fit.

If you get insignificant p-values associated with combining two (or more) categories together, refit your model and perform diagnostics to see what assumptions are still met (or not met)

```
Analysis of Variance Table

Model 1: reviews_per_year ~ minimum_nights + number_of_reviews + neighbourhood_group +
    room_type + minimum_nights:room_type + neighbourhood_group:room_type
Model 2: reviews_per_year ~ minimum_nights + number_of_reviews + room_type +
    new_neighbourhood_group + minimum_nights:room_type + room_type:new_neighbourhood_group
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    976 222002
2    979 222567 -3   -565.48 0.8287 0.4782

adjusted r squared for old model:

0.4068291

adjusted r squared for new model with brooklyn and manhattan combined:

0.4071405
```

- This combination increases the correlation, as the adjusted R squared value increases in the updated model that combines Brooklyn and Manhattan.
- There is not a significant difference between the old model and the newly refit model, showing these variables could be combined.