Stats_101A_hw5_anna_piskun

Anna Piskun

2/7/2020

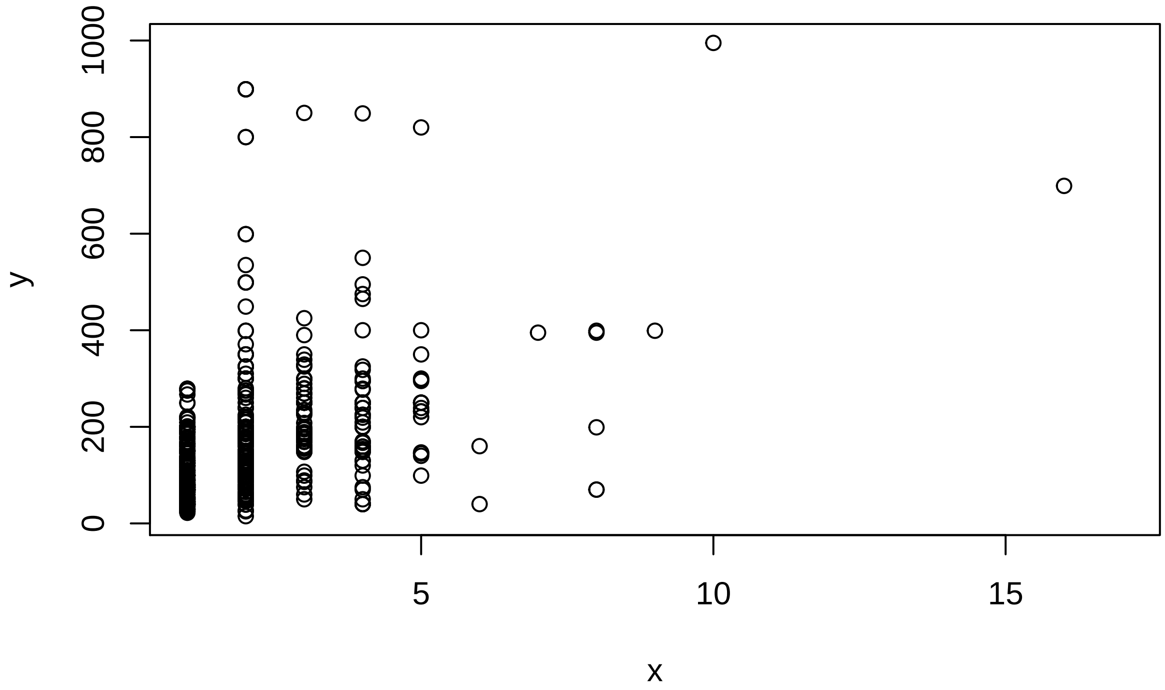
Cleaning Data/Testing Multiple Models

```
setwd("~/Desktop")

library(readr)
airbnb <- read_csv("la_airbnb_detailed.csv")
```

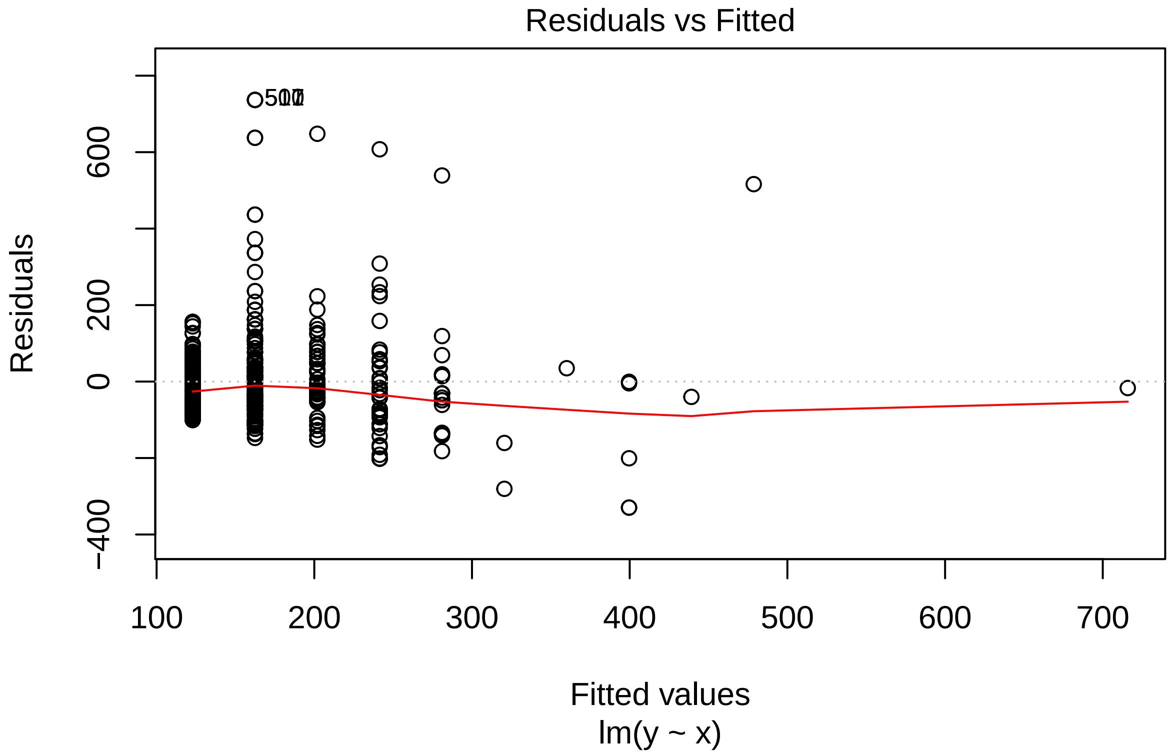
```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   id = col_double(),
##   scrape_id = col_double(),
##   last_scraped = col_date(format = ""),
##   thumbnail_url = col_logical(),
##   medium_url = col_logical(),
##   xl_picture_url = col_logical(),
##   host_id = col_double(),
##   host_since = col_date(format = ""),
##   host_is_superhost = col_logical(),
##   host_listings_count = col_double(),
##   host_total_listings_count = col_double(),
##   host_has_profile_pic = col_logical(),
##   host_identity_verified = col_logical(),
##   neighbourhood_group_cleansed = col_logical(),
##   latitude = col_double(),
##   longitude = col_double(),
##   is_location_exact = col_logical(),
##   accommodates = col_double(),
##   bathrooms = col_double(),
##   bedrooms = col_double()
##   # ... with 34 more columns
## )

## See spec(...) for full column specifications.
```

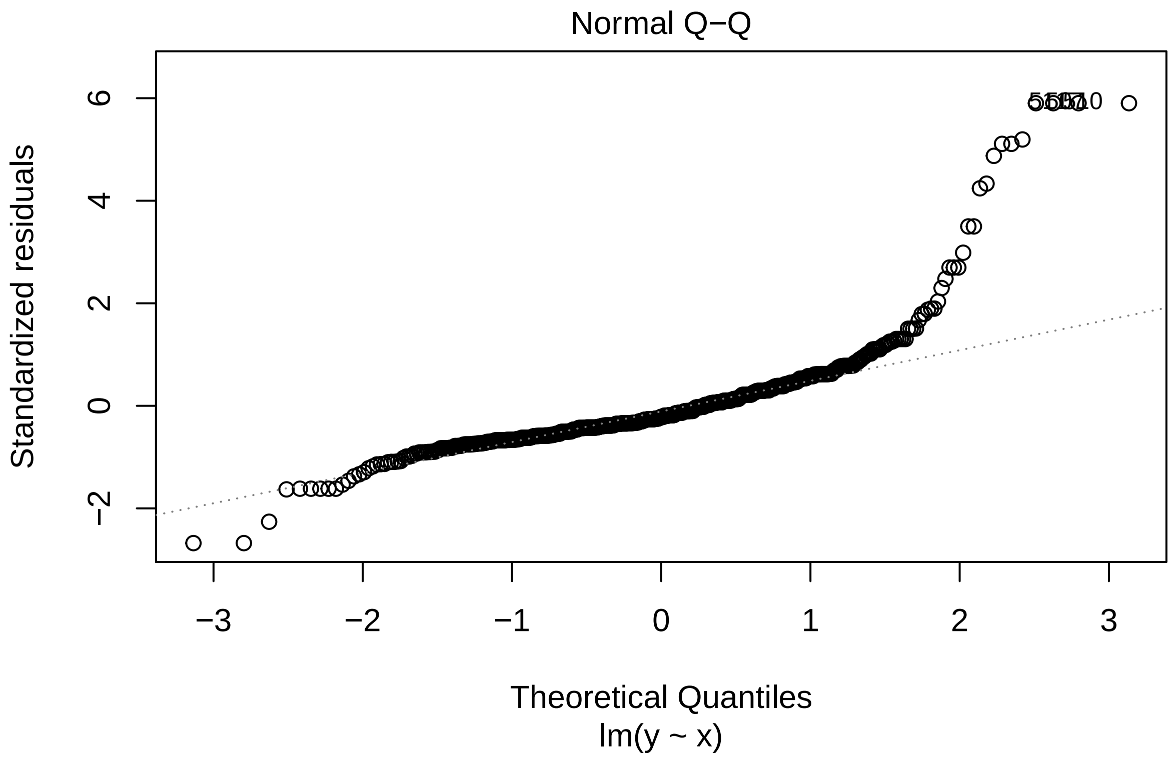
```
airbnb_westwood <- dplyr::filter(airbnb, neighbourhood =="Westwood")

airbnb_westwood_clean <- dplyr::select(airbnb_westwood, id, name, description, property_type, room_type

airbnb_beds <- dplyr::filter(airbnb_westwood_clean, beds > 0)
#having 0 beds is a nonsensical value, so we remove it to avoid creating pseudo influential points and
airbnb_final <- dplyr::filter(airbnb_beds, beds < 50)
#after looking at the description a little closer, the airbnb with 50 beds was for a hostel. By limitin
```

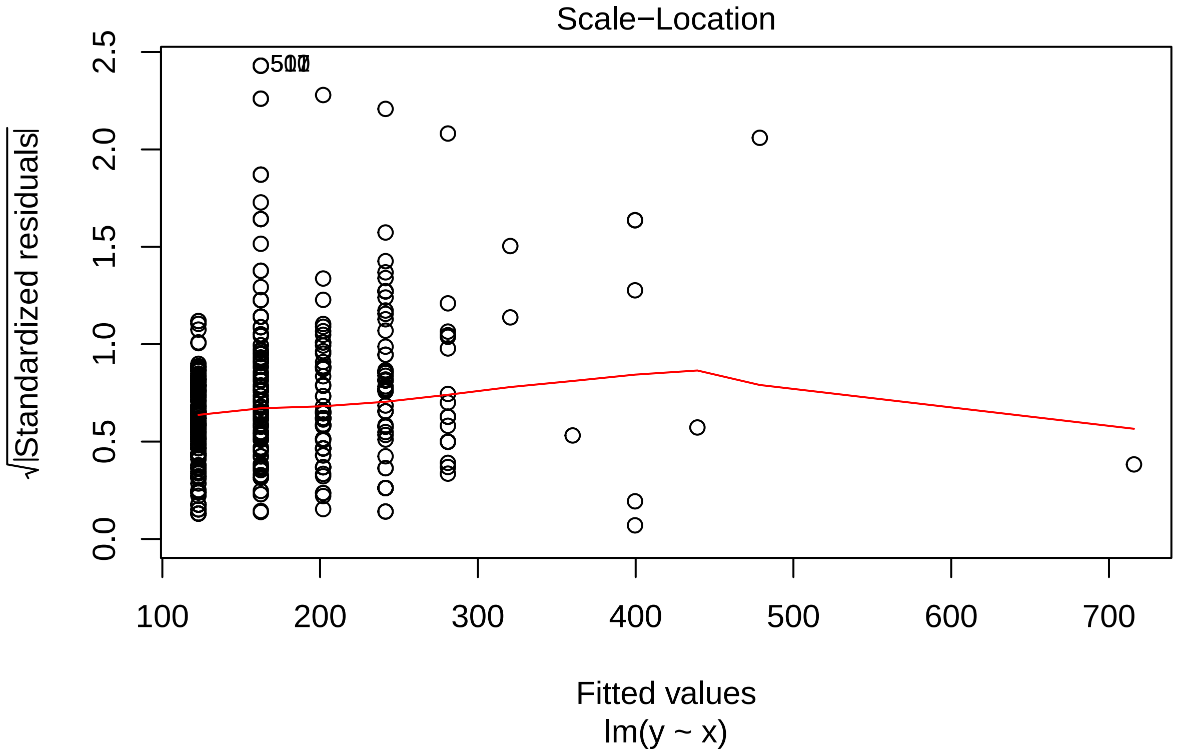
```r
x <- airbnb_final$beds
y <- airbnb_final$price
plot(y ~ x)
```

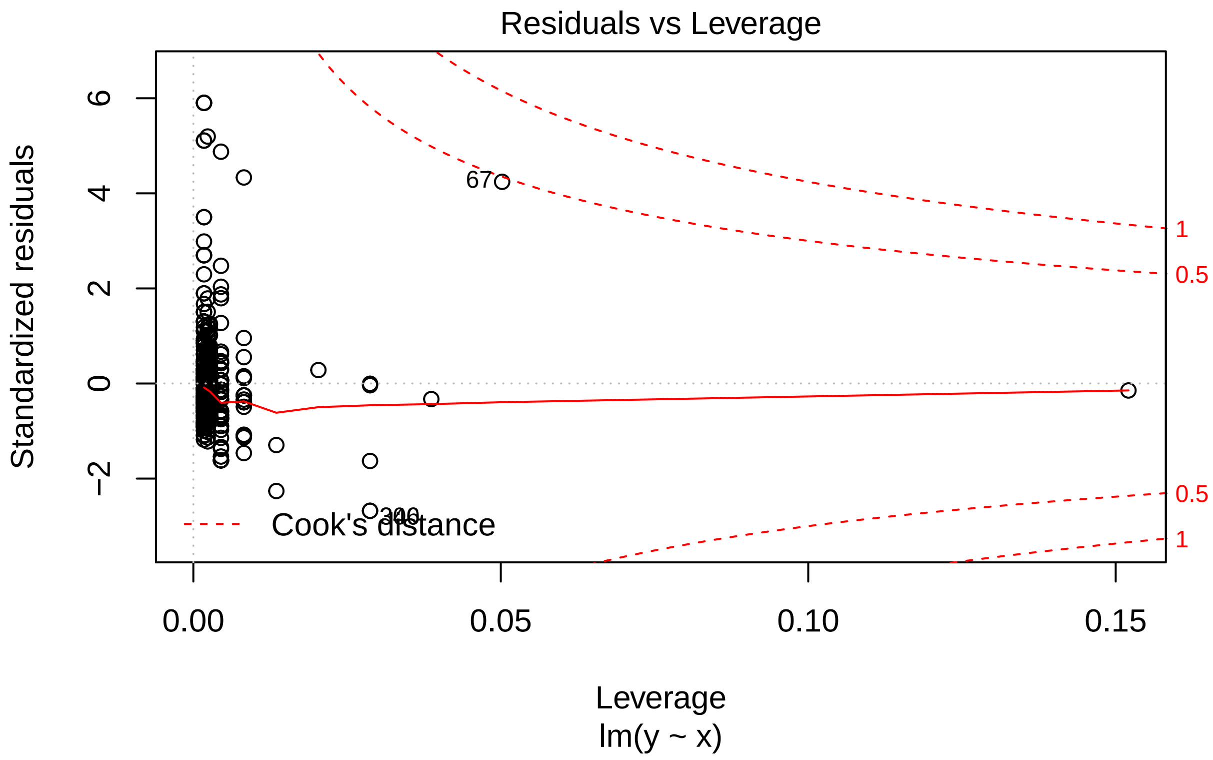


```r
airbnb_lm <- lm(y~x)
plot(airbnb_lm)
```

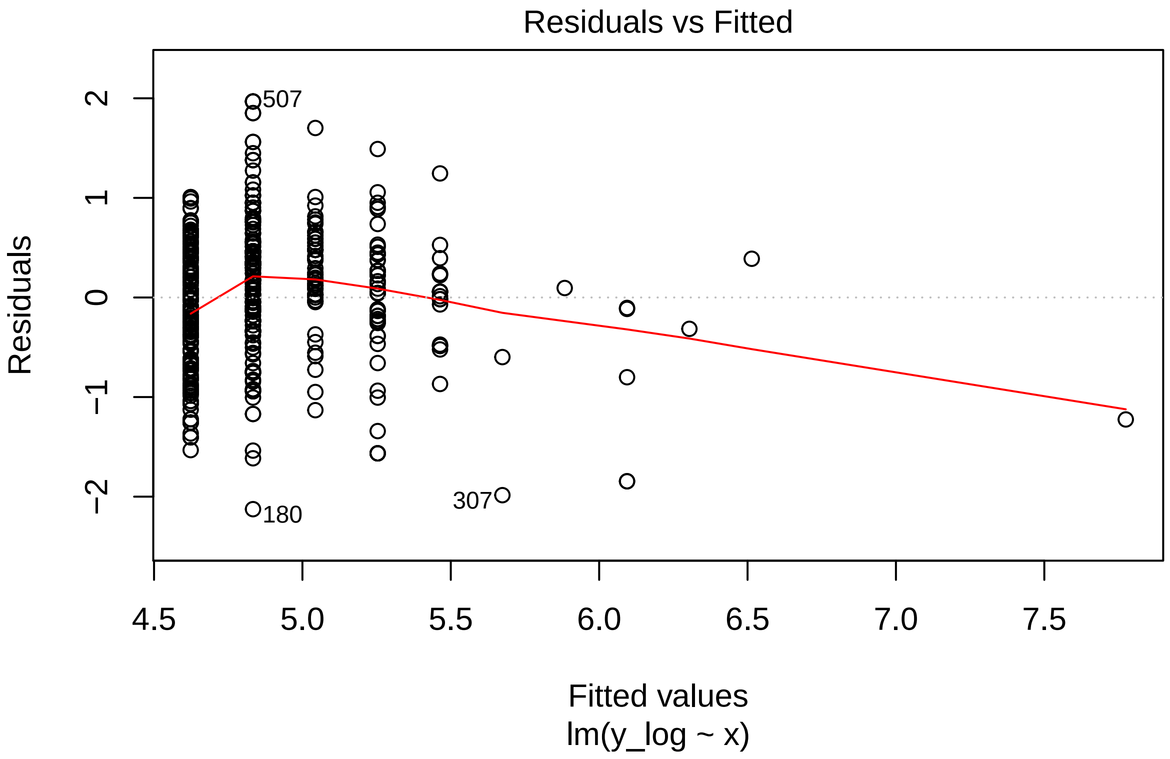


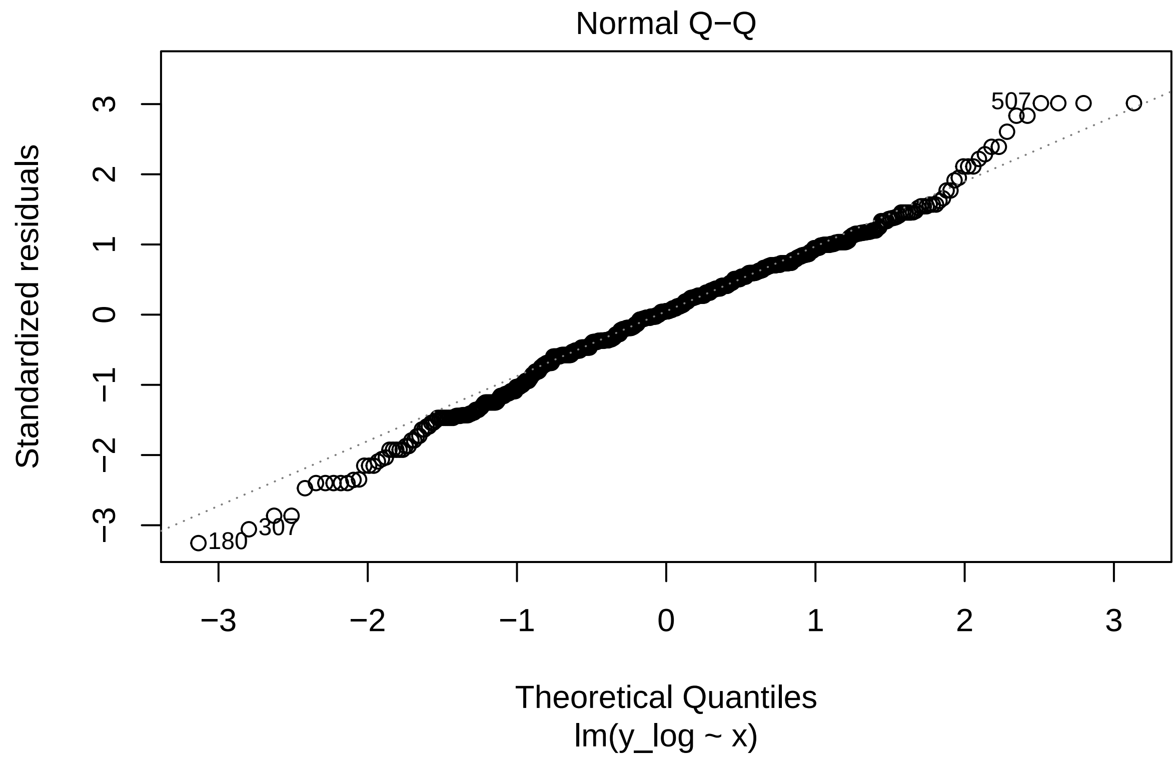

Normal Q−Q
Standardized residuals
6
4
2
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y ~ x)

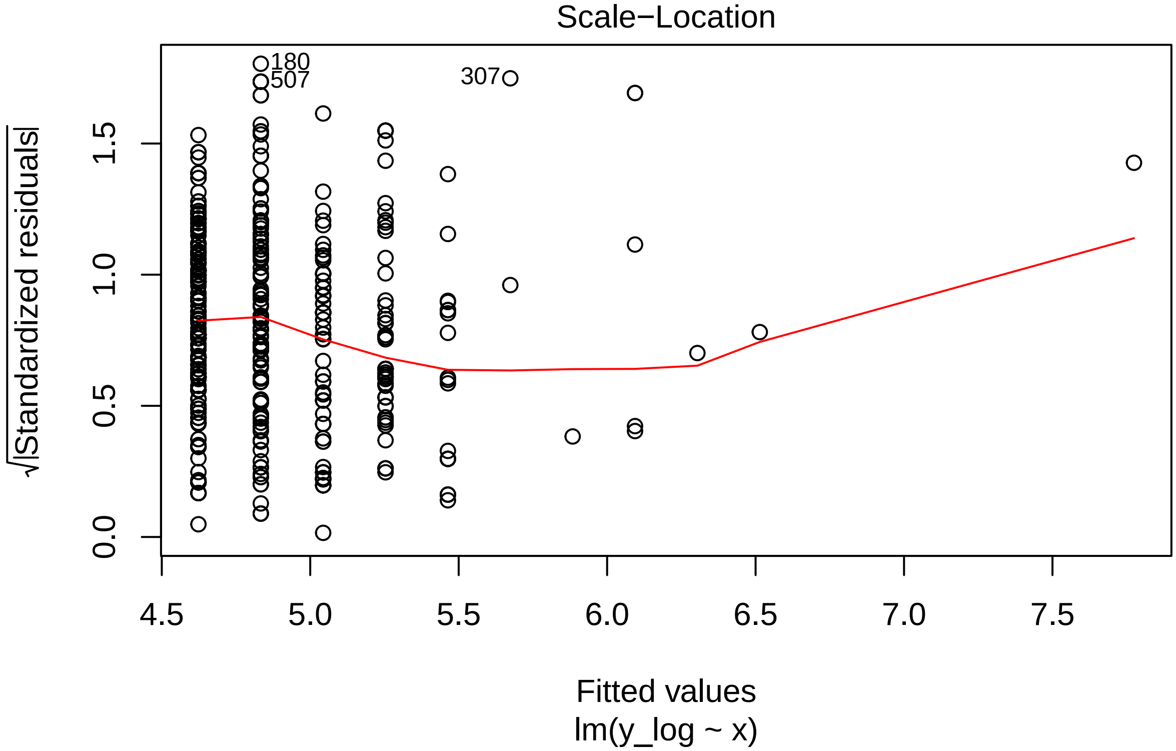

Scale−Location
√|Standardized residuals|
2.5
2.0
1.5
1.0
0.5
0.0
100
200
300
400
500
600
700
Fitted values
lm(y ~ x)

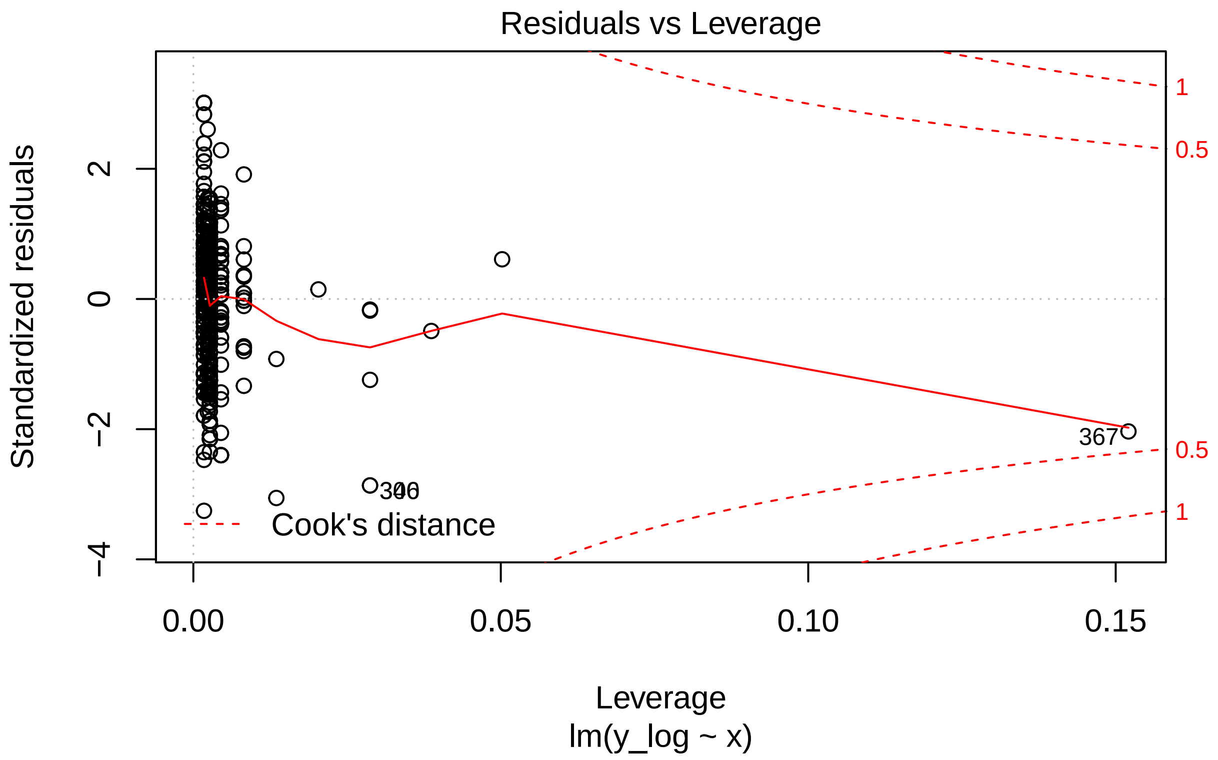


```r
x_log <- log(x)
y_log <- log(y)

airbnb_log_lm <- lm(y_log ~ x)
plot(airbnb_log_lm)
```

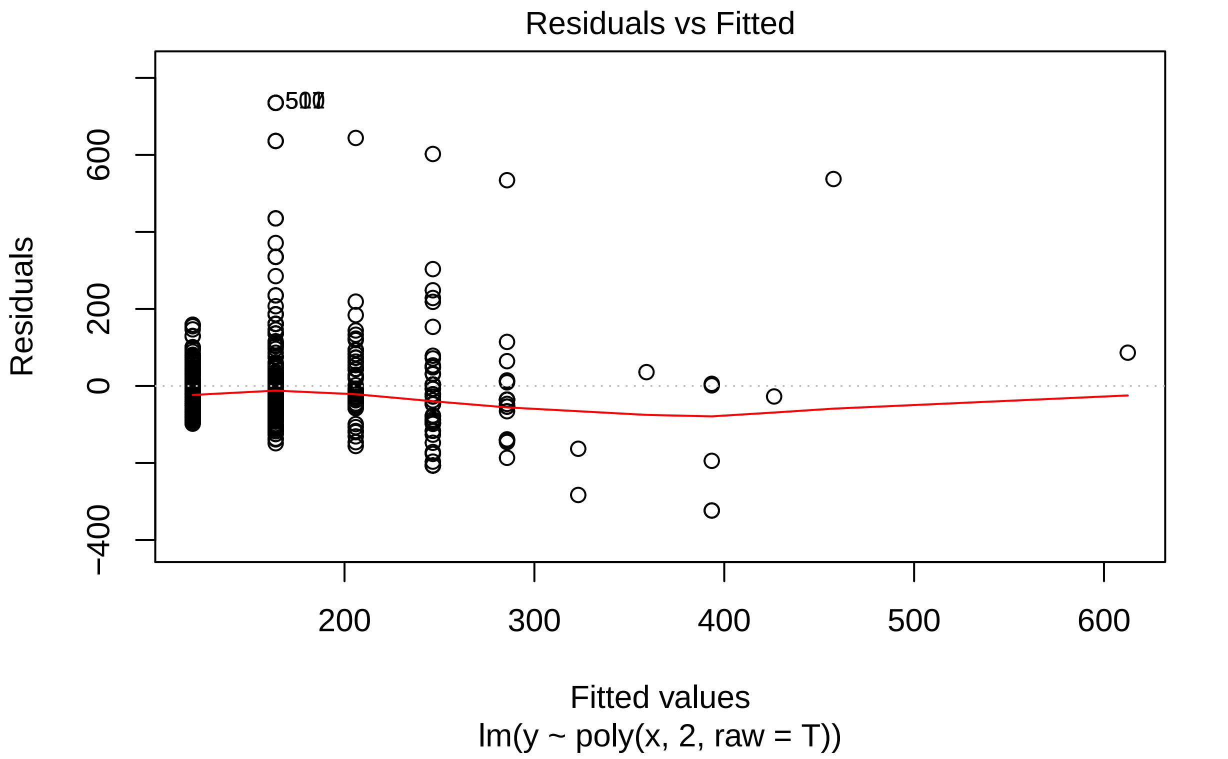


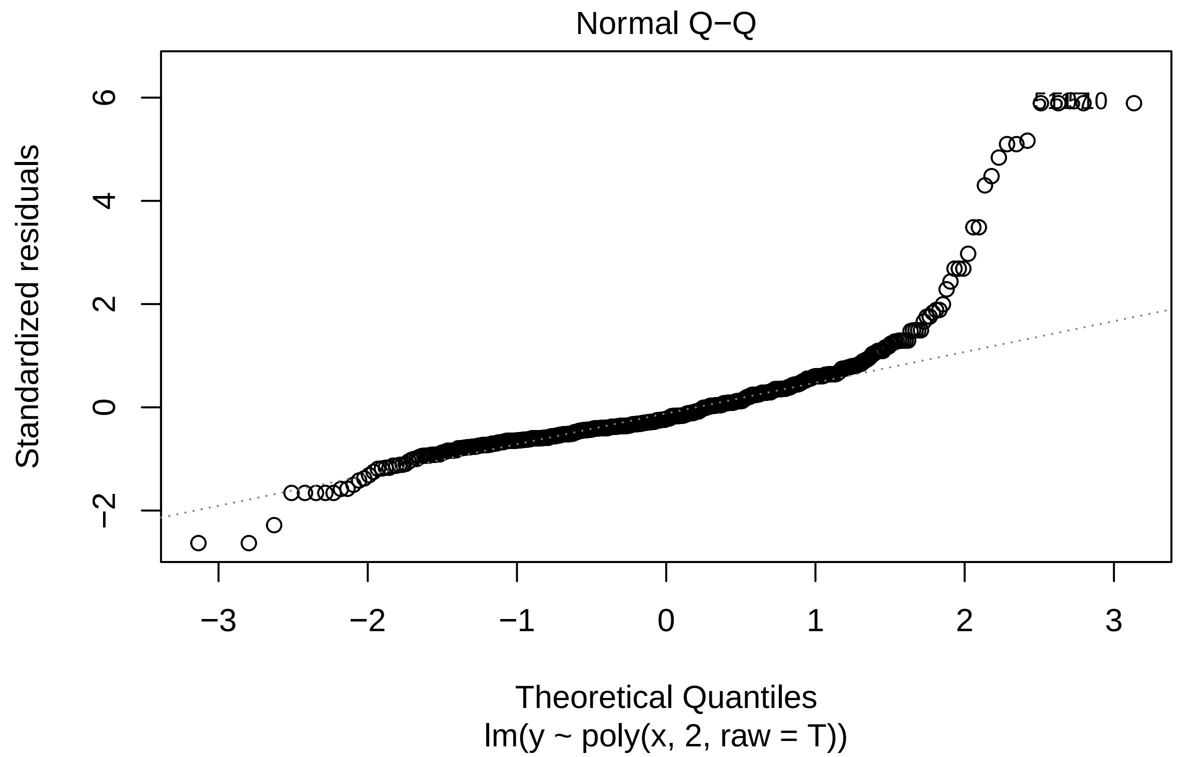

Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
507
180
307
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y_log ~ x)

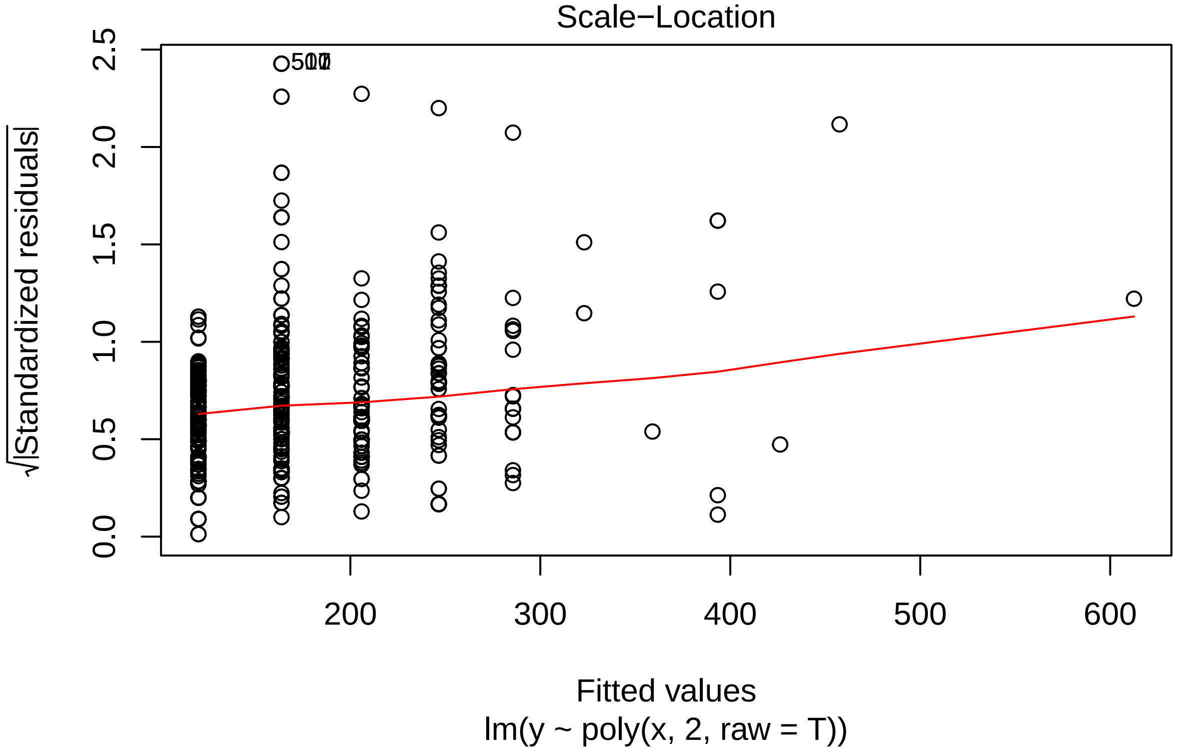

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
180
507
307
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(y_log ~ x)

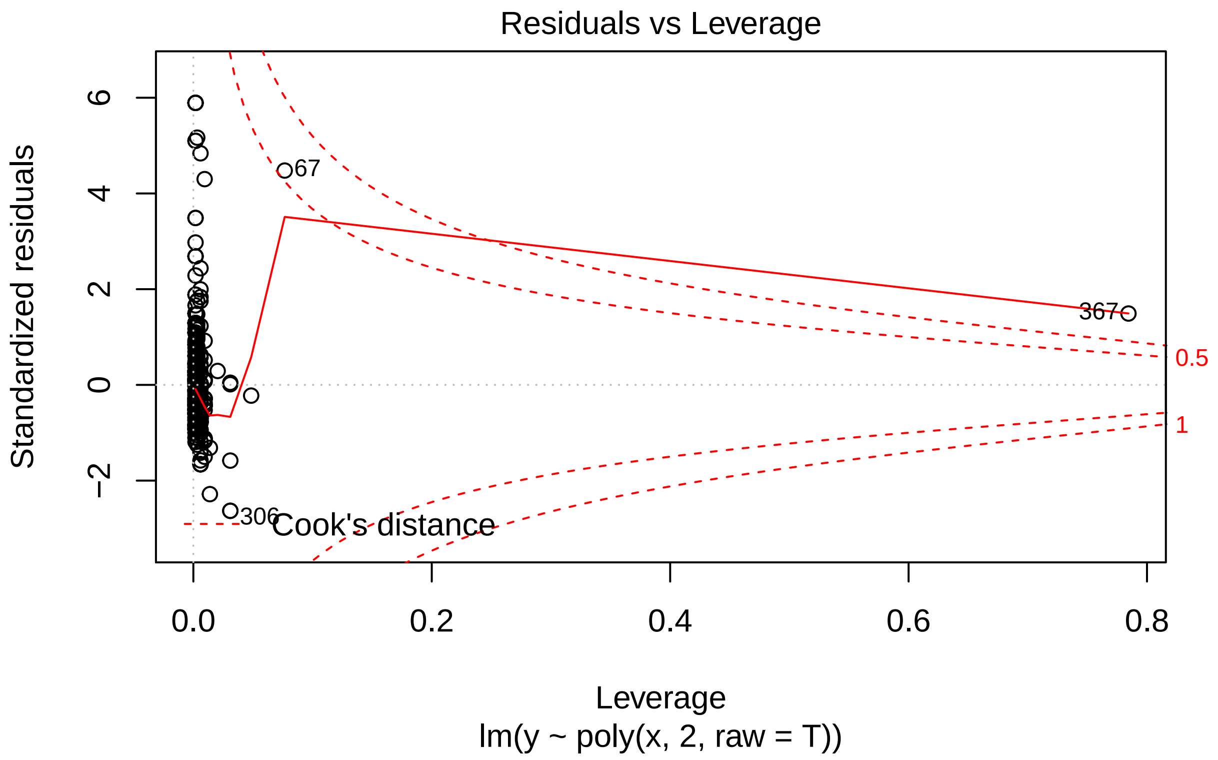


```
airbnb_quad <- lm(y~poly(x,2,raw = T), data = airbnb_final)
plot(airbnb_quad)
```

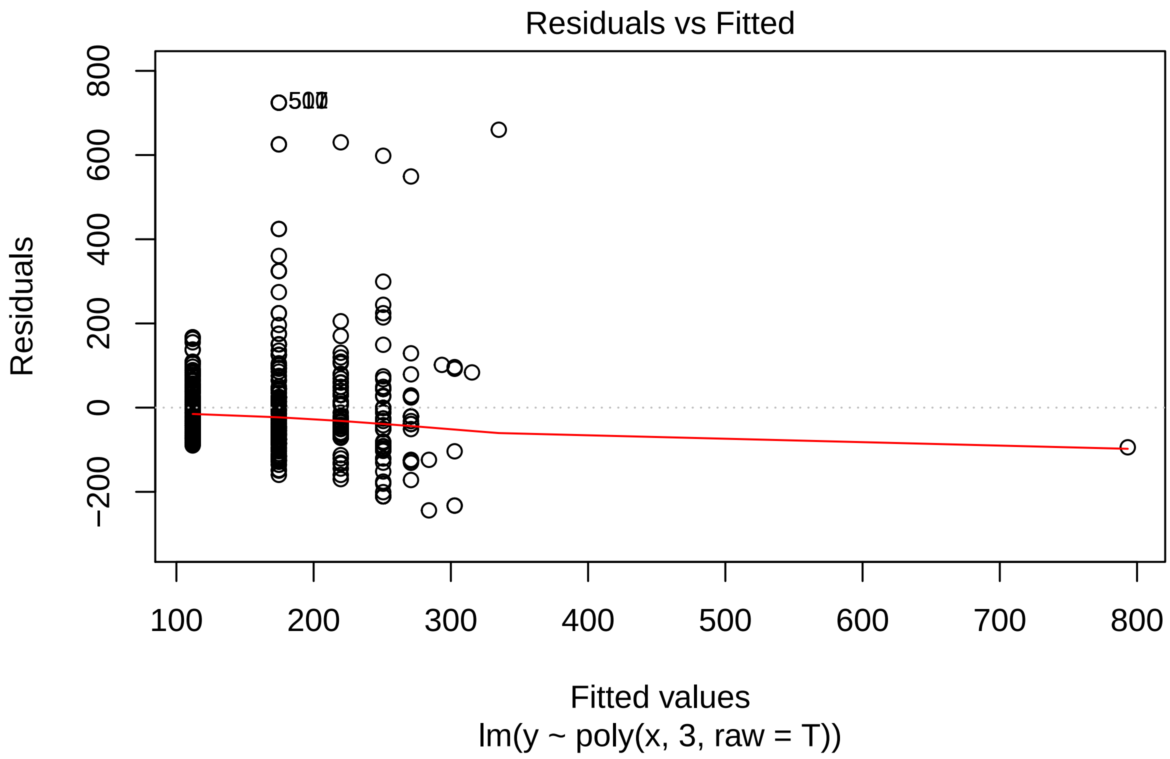


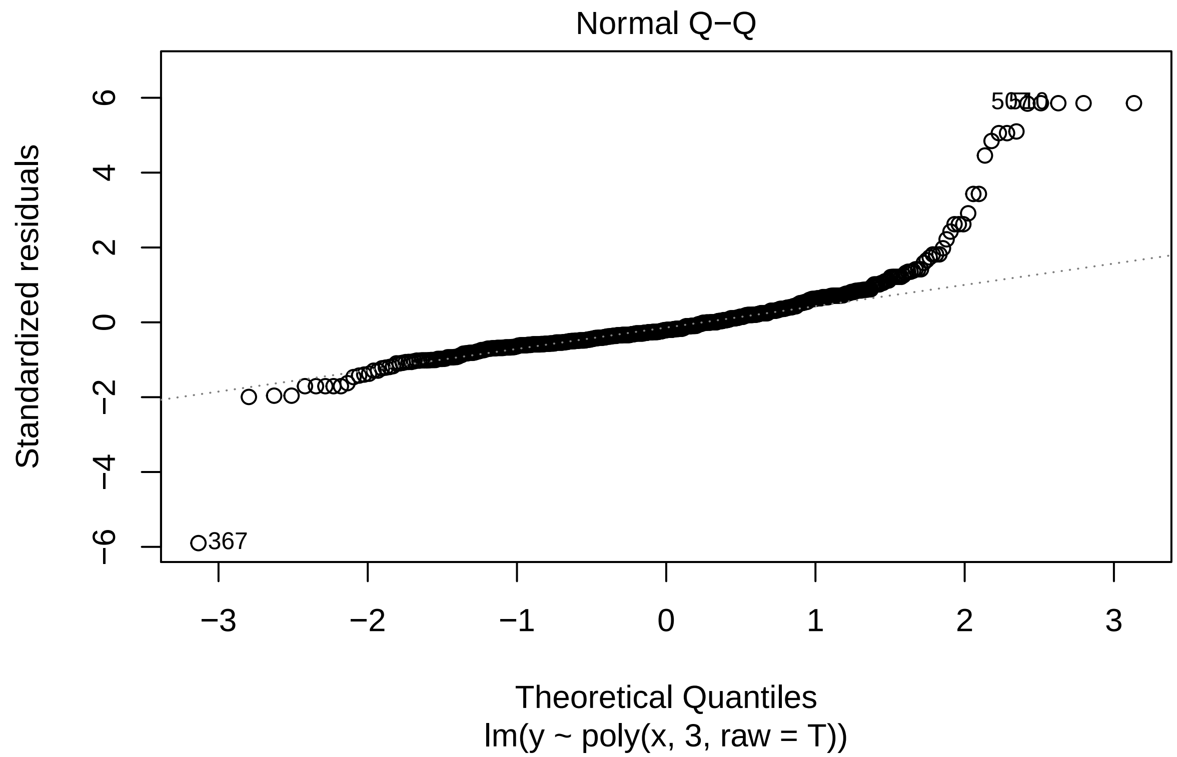

Normal Q−Q
Standardized residuals
6
4
2
0
−2
5557510
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y ~ poly(x, 2, raw = T))

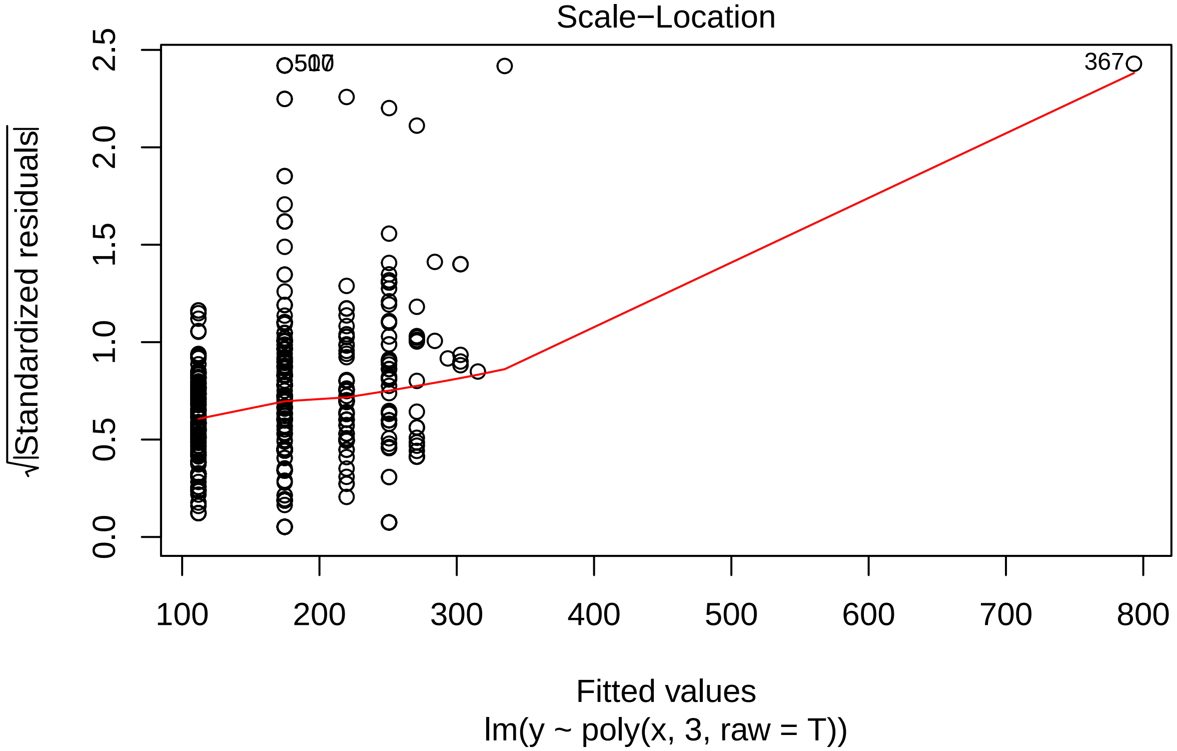

Scale−Location
√|Standardized residuals|
2.5
2.0
1.5
1.0
0.5
0.0
510
200
300
400
500
600
Fitted values
lm(y ~ poly(x, 2, raw = T))



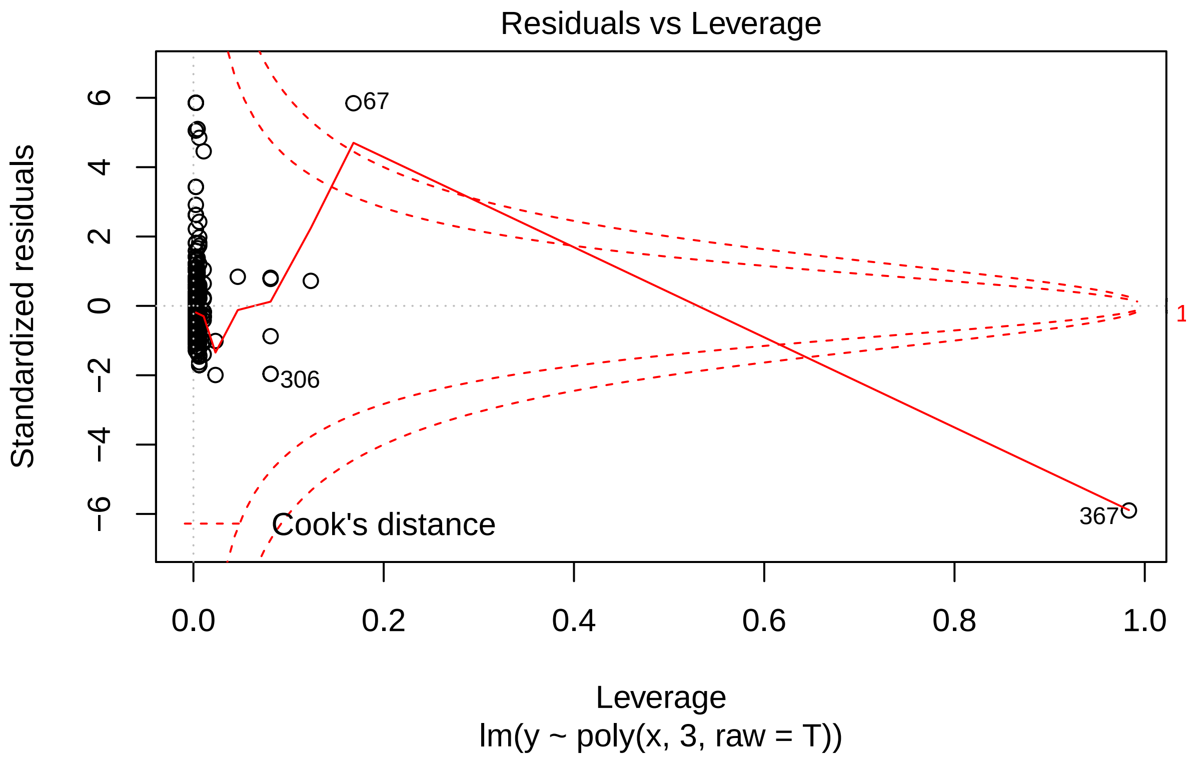
```
airbnb_cub <- lm(y~poly(x,3,raw = T), data = airbnb_final)
plot(airbnb_cub)
```



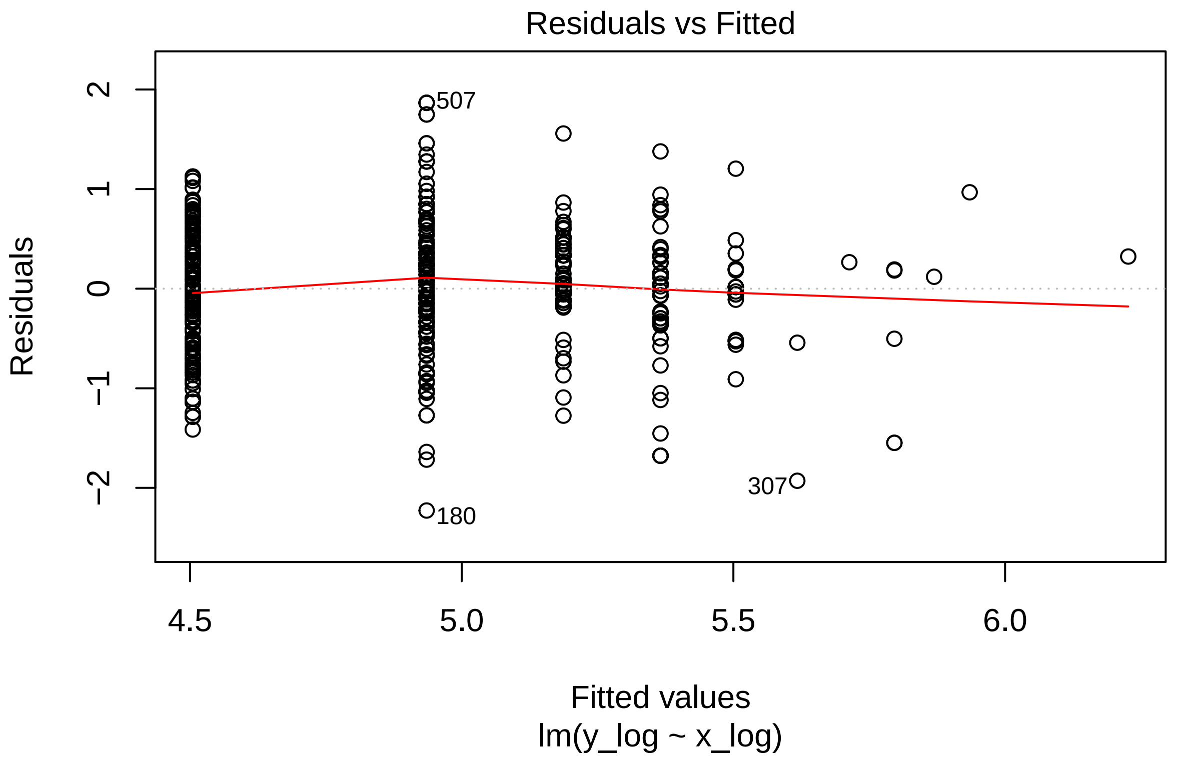


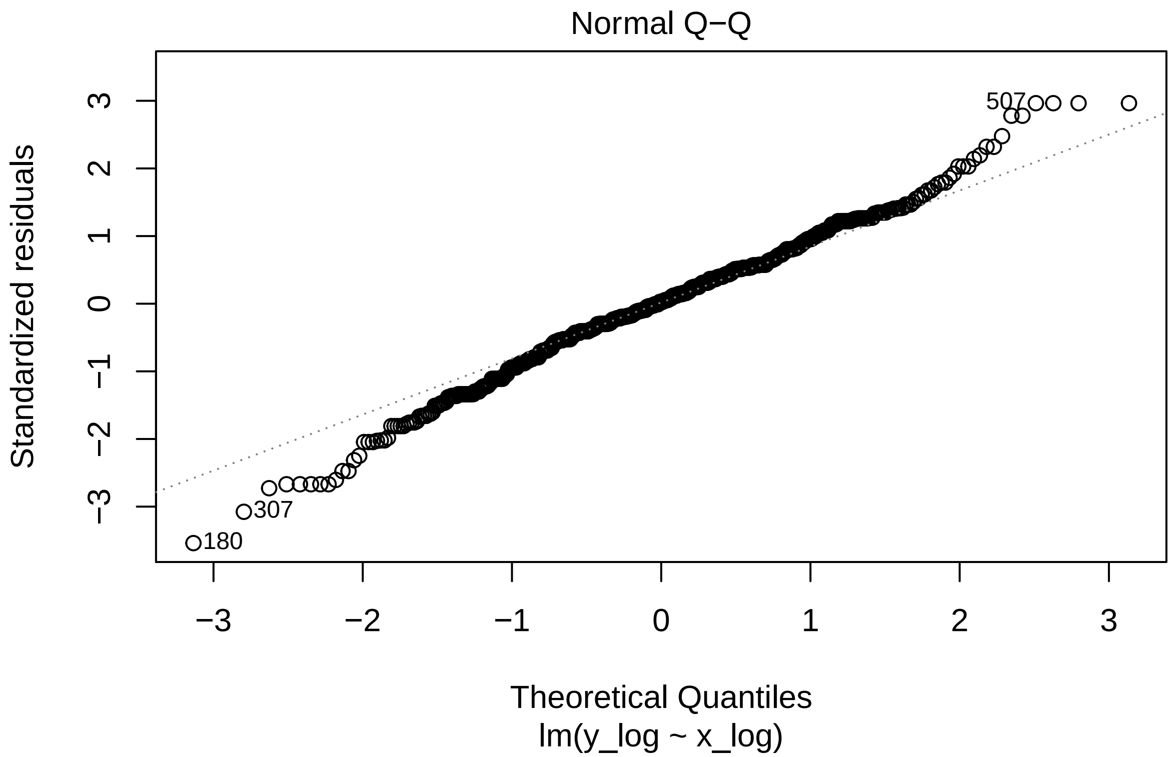


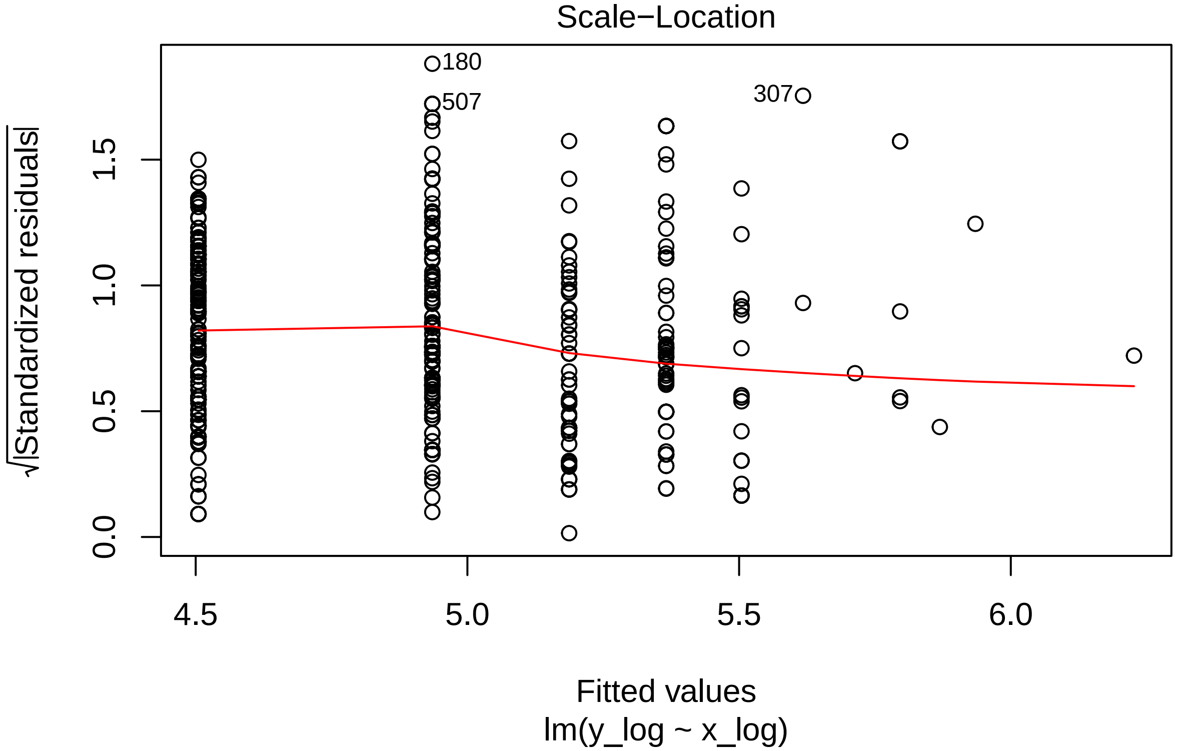
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

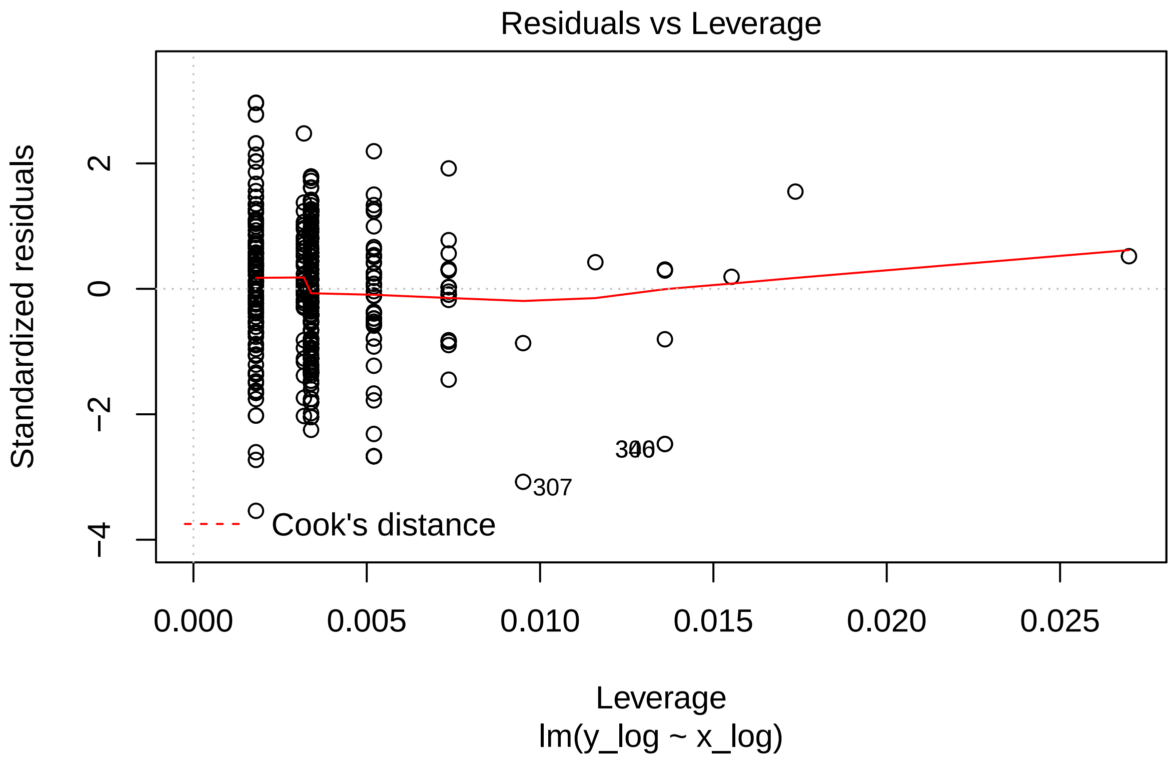


```
airbnb_log_log_lm <- lm(y_log~x_log)
plot(airbnb_log_log_lm)
```




Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
3
507
307
180
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y_log ~ x_log)

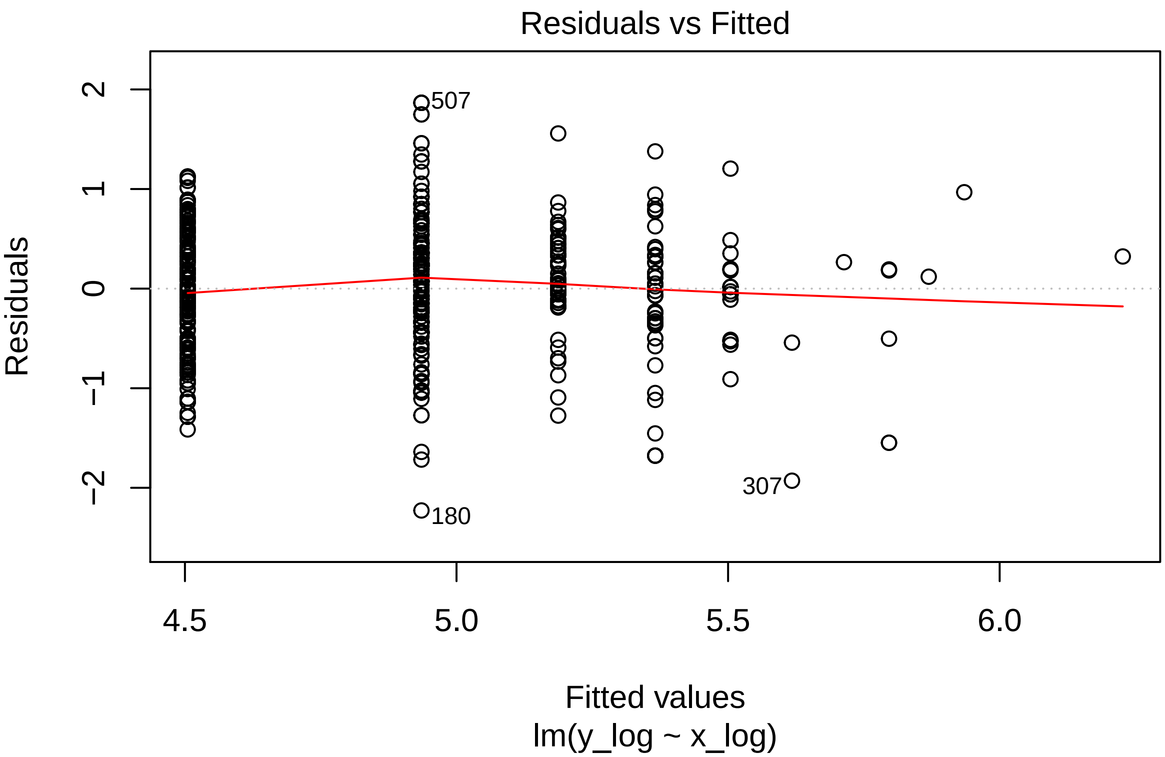

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
180
507
307
4.5
5.0
5.5
6.0
Fitted values
lm(y_log ~ x_log)



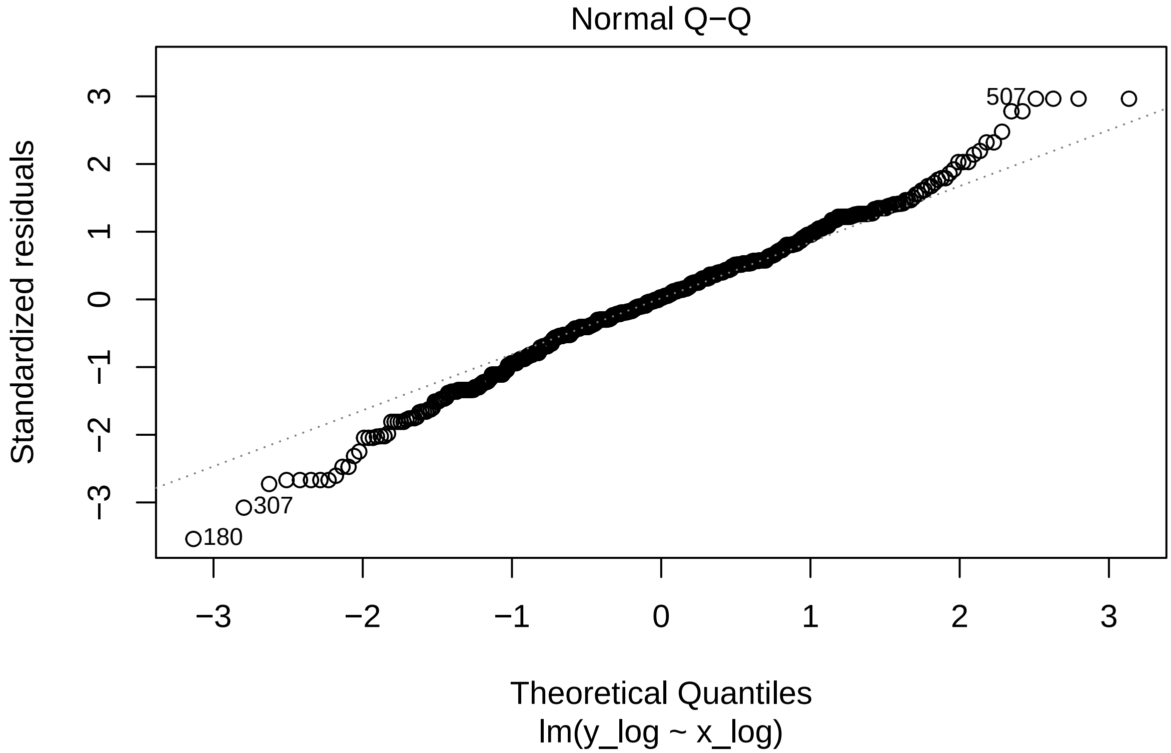
```
#determine that using a log/log transformation gives us the best fitting model
```

Part A and B: summary output of final model and diagnostic plots

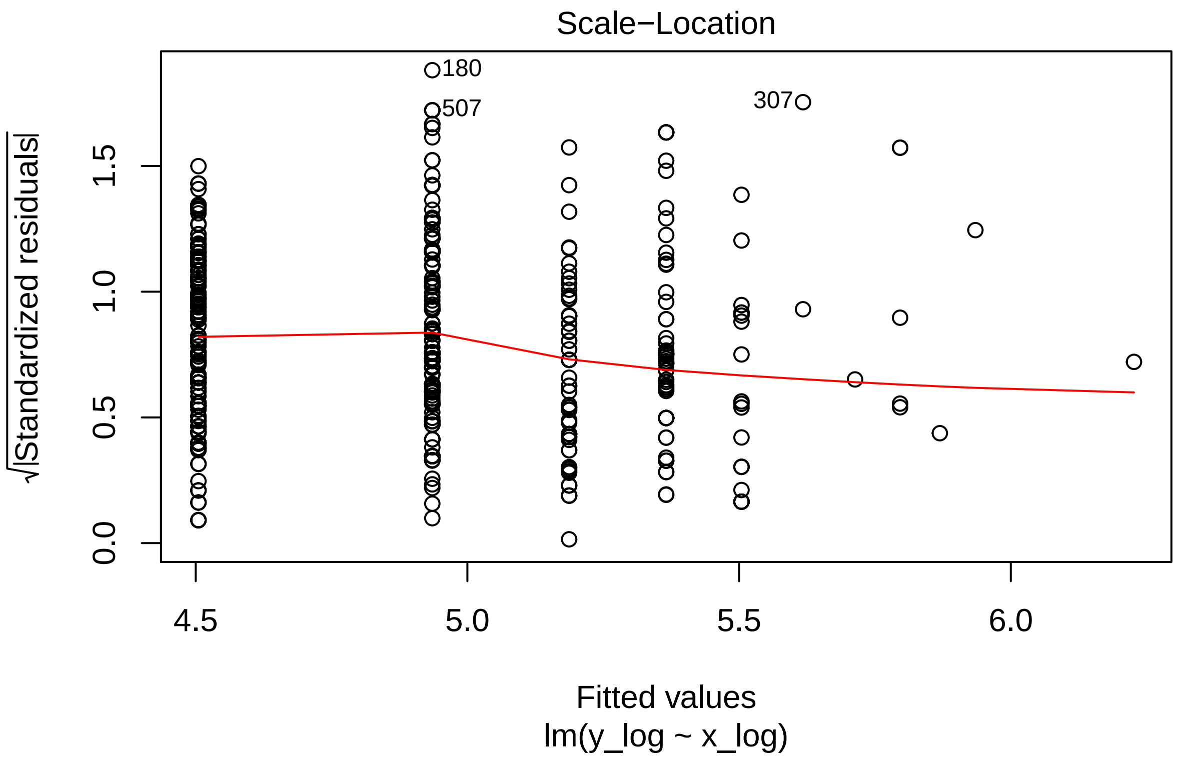
```
airbnb_log_log_lm <- lm(y_log~x_log)
plot(airbnb_log_log_lm)
```

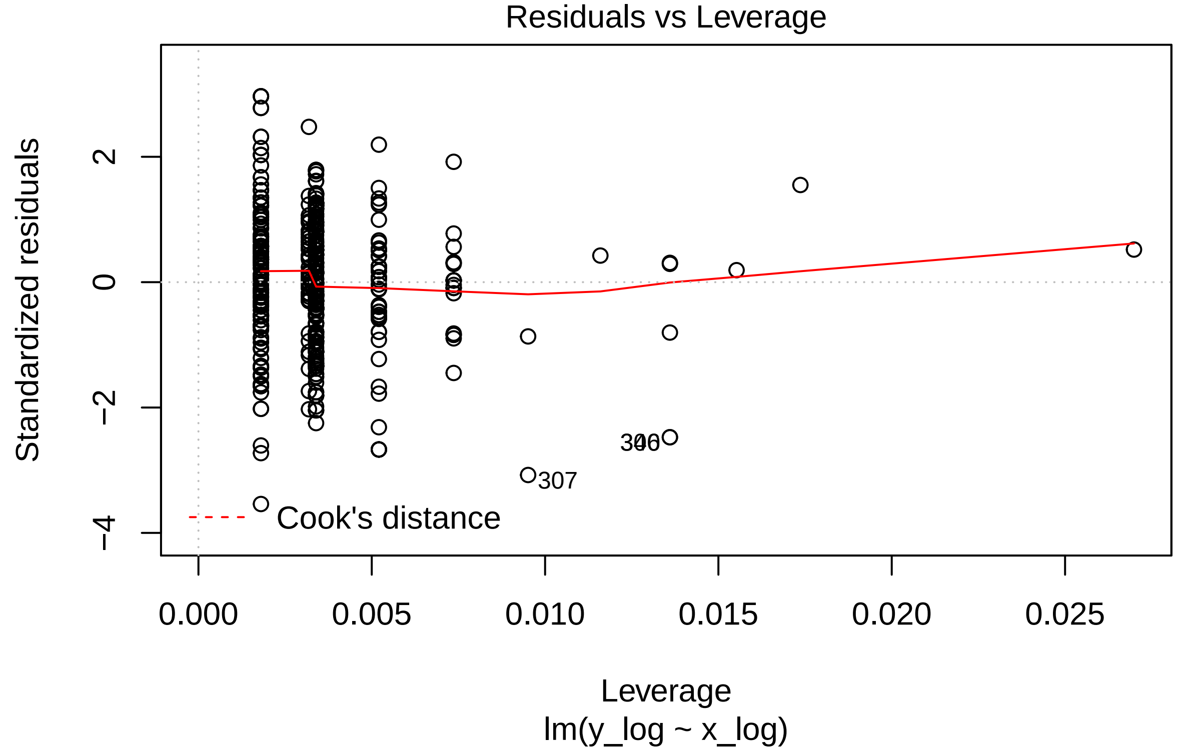


Residuals vs Fitted
Residuals
2
1
0
−1
−2
507
180
307
4.5
5.0
5.5
6.0
Fitted values
lm(y_log ~ x_log)



Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
507
307
180
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y_log ~ x_log)





```r
summary(airbnb_log_log_lm)
```

```
##
## Call:
## lm(formula = y_log ~ x_log)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.22741 -0.34034  0.01702  0.36249  1.86582 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.50505    0.03670  122.76   <2e-16 ***
## x_log        0.62096    0.04541   13.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6299 on 579 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.2441, Adjusted R-squared:  0.2428 
## F-statistic:   187 on 1 and 579 DF,  p-value: < 2.2e-16
```

```r
# gives us the following formula for our model: Price = 4.50505 + 0.62096 (Beds)
```

Part C: An explanation of why you made any transformations you might have made and an explanation of why you removed any points you might have removed

Going through the original airbnb dataset it becomes clear that many of the variables included were not needed to create a model that would explain the relationship between price and number of beds for listings in Westwood. When deciding which variables to keep I obviously chose beds and prices, but also included additional variables such as property type and description to help explain outliers. This came into play when deciding which values to exclude. I removed any rows with a bed value of 0, since this is clearly nonsensical data that was probably due to error. On the other end, I removed rows with beds > 50 since when looking at the row with 50 beds its property type said that the building was a hostel. By doing so, I narrowed the scope of the problem to all listings of single units instead of just all listings on Airbnb. When determining which model fit the best I considered multiple transformations. Eventually landing on a log/log transformation, it became clear that this was the best fit by analyzing the diagnostic plots. First looking at the residual plot we see that there is little to no trend so the constant variance assumption is satisfied. Next looking at the QQ plot, the log/log transformation produced the straightest fit when compared to the other tested models (no tranformation, quadratic tranform, cubic transform, and log(y) transform) which indicates that our data does in fact follow a normal distribution. The scale-location plot again shows pretty much no clear trend or pattern which supports our earlier notion that the constant variance condition is met. Lastly, the residuals vs. leverage plot shows no high leverage points (there is one potential influential point, but compared to the other models it is very mild) thus confirming that the log/log model is the best fit for the data. Additionally, the model gives us an r-squared value of 0.2441, which when dealing with real datasets and the random variability that exists in the real world, is quite good and allows us to explain a significant portion of the variability with our model.

Part D: what does your model estimate is the price of a rental in Westwood with 3 beds? Give an interval and indicate whether it is a prediction or confidence interval.

```r
exp(predict(airbnb_log_log_lm, data.frame(x_log = log(3)), interval = "prediction", level = 0.95))
```

```
##        fit      lwr      upr
## 1 178.9742 51.83346 617.9743
```

The model estimates that the price of a rental in Westwood with 3 beds is around $178.97 with a prediction interval of 51.83346 to 617.9743.

Part E: A discussion of any shortcomings of your model and their affect on your interval in (d)

Specific to the log/log transformation, the tails of the QQ plot veer off the straight line which shows some weakness in the assumption that our data is normal which could affect our prediction interval by giving skewed lower and upper bounds. Likewise, the existence of outliers (such as the unit with 17 beds) can affect our model and shift our prediction interval higher. Another major shortcoming of my model, and just a simple linear regression in general, is that it does not take into consideration the many other variables that may affect the relationship between the number of beds and price. For example, one of the variables I included in my cleaned dataset was bed type. This could affect price in that a twin bed and a queen bed can accomodate different numbers of guests and would result in there being a price difference. However, when just looking at the number of beds we don't see the added layers that could better explain our data and create a better fitting model.