modeling

```
library(GGally)
library(ggplot2)
library(lmtest)
library(lubridate)
library(moments)
library(sandwich)
library(stargazer)
library(tidyverse)
library(data.table)

source('./get_robust_se.R')
```

```
data <- read.csv('data/googleplaystore.csv')
summary(data)
```

```
##      App              Category            Rating         Reviews
##  Length:10841       Length:10841       Min.   : 1.000   Length:10841
##  Class :character   Class :character   1st Qu.: 4.000   Class :character
##  Mode  :character   Mode  :character   Median : 4.300   Mode  :character
##                                        Mean   : 4.193
##                                        3rd Qu.: 4.500
##                                        Max.   :19.000
##                                        NA's   :1474
##      Size              Installs            Type              Price
##  Length:10841       Length:10841       Length:10841       Length:10841
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Content.Rating        Genres          Last.Updated       Current.Ver
##  Length:10841       Length:10841       Length:10841       Length:10841
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Android.Ver
##  Length:10841
##  Class :character
##  Mode  :character
##
##
##
##
```

```r
clean_size <- function(s) {
  n <- nchar(s)
  last <- substr(s, n, n)
  if (last == "B") {
    k <- 10.0
  } else if (last == "M") {
    k <- 1.0
  } else if (last == "k") {
    k <- 0.1
  } else {
    return(NA)
  }
  return(as.numeric(substr(s, 1, n - 1)) * k)
}

clean_reviews <- function(s) {
  n = nchar(s)
  last <- substr(s, n, n + 1)
  if (last == "M") {
    return(as.numeric(substr(s, 1, n - 1)) * 1.0e6)
  } else {
    return(as.numeric(s))
  }
}

MAX_DATE = suppressWarnings(max(mdy(data$Last.Updated), na.rm = TRUE))
clean_date <- function(d) {
  return(interval(d, MAX_DATE) / years(1))
}

data_clean <- data %>%
  distinct() %>%
  mutate(
    installs            = suppressWarnings(as.numeric(gsub("[\\+,]", "", Installs))),
    size                = sapply(Size, clean_size),
    reviews             = sapply(Reviews, clean_reviews),
    rating              = Rating,
    price               = suppressWarnings(as.numeric(gsub("\\$", "", as.character(Price)))),
    is_free             = price == 0,
    last_updated        = suppressWarnings(mdy(Last.Updated)),
    last_updated        = interval(last_updated, MAX_DATE) / years(1),
    android_ver         = suppressWarnings(as.numeric(substr(gsub("Varies with device", NA, Android.Ver
    current_version     = suppressWarnings(as.numeric(substr(gsub("Varies with device", NA, Current.Ver
    category            = Category,
    is_family_category  = category == "FAMILY",
    is_game_category    = category == "GAME",
    is_tools_category   = category == "TOOLS",
    genre               = Genres,
    content_rating      = Content.Rating,
    is_content_everyone = content_rating == "Everyone"
  ) %>%
  select(., installs, size, reviews, rating, price, is_free, last_updated,
         android_ver, current_version, category, is_family_category,
```

```
        is_game_category, is_tools_category, genre, content_rating,
        is_content_everyone) %>%
  drop_na() %>%
  filter(., reviews >= 100, rating <= 5.0)

# bins for install count: install groups-- increments by 1, there are 19 of them.
install_groups <- data.table(table(data_clean$installs))
setnames(install_groups,c('V1','N'),c('installs','count'))
install_groups[, install_group := 1:.N]
install_groups <- install_groups[,c('installs','install_group')]
data_clean <- merge(data_clean,install_groups,by='installs')

glimpse(data_clean)
```

```
## Rows: 5,408
## Columns: 17
## $ installs            <dbl> 500, 500, 500, 1000, 1000, 1000, 1000, 1000, 1000,~
## $ size                <dbl> 22.0, 10.0, 2.5, 5.4, 2.4, 13.0, 6.5, 53.0, 15.0, ~
## $ reviews             <dbl> 156, 120, 124, 726, 112, 112, 123, 787, 141, 171, ~
## $ rating              <dbl> 4.6, 4.5, 4.8, 4.8, 4.5, 4.4, 4.8, 4.8, 5.0, 4.8, ~
## $ price               <dbl> 3.49, 0.00, 0.00, 3.99, 7.99, 0.00, 0.00, 1.99, 0.~
## $ is_free             <lgl> FALSE, TRUE, TRUE, FALSE, FALSE, TRUE, TRUE, FALSE~
## $ last_updated        <dbl> 0.402739726, 0.010958904, 1.958904110, 0.632876712~
## $ android_ver         <dbl> 4.4, 5.0, 4.2, 4.1, 4.0, 4.4, 3.0, 4.0, 4.1, 2.2, ~
## $ current_version     <dbl> 1.1, 1.5, 1.2, 2.3, 3.0, 4.5, 3.1, 1.4, 1.8, 1.7, ~
## $ category            <chr> "MEDICAL", "DATING", "FAMILY", "BUSINESS", "PRODUC~
## $ is_family_category  <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FA~
## $ is_game_category    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
## $ is_tools_category   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
## $ genre               <chr> "Medical", "Dating", "Entertainment", "Business", ~
## $ content_rating      <chr> "Everyone", "Mature 17+", "Everyone", "Everyone", ~
## $ is_content_everyone <lgl> TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, T~
## $ install_group       <int> 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
```

```
data_clean$log_installs        <- log10(data_clean$installs)
data_clean$log_size            <- log10(data_clean$size)
data_clean$log_current_version <- log10(data_clean$current_ver + 1)
data_clean$log_last_updated    <- log10(data_clean$last_updated + 1)
```

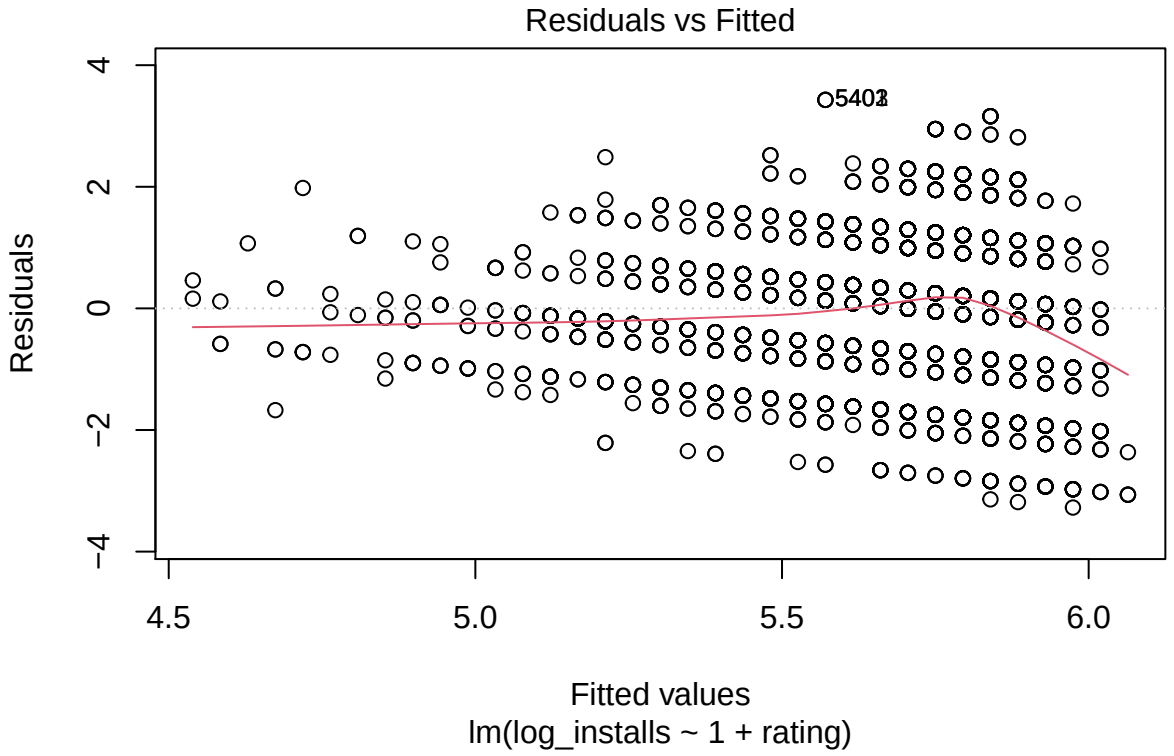
```
model_small  <- lm(log_installs ~ 1 + rating, data = data_clean)
model_medium <- lm(log_installs ~ 1 + rating + log_size + log_current_version +
                     log_last_updated + is_free + is_family_category +
                     is_game_category + is_tools_category + is_content_everyone,
                   data = data_clean)
model_large  <- lm(log_installs ~ 1 + rating + log_size + log_current_version +
                     log_last_updated + is_free + is_content_everyone +
                     rating * is_family_category + rating * is_game_category +
                     rating * is_tools_category,
                   data = data_clean)

stargazer(
  model_small,
  model_medium,
  model_large,
```

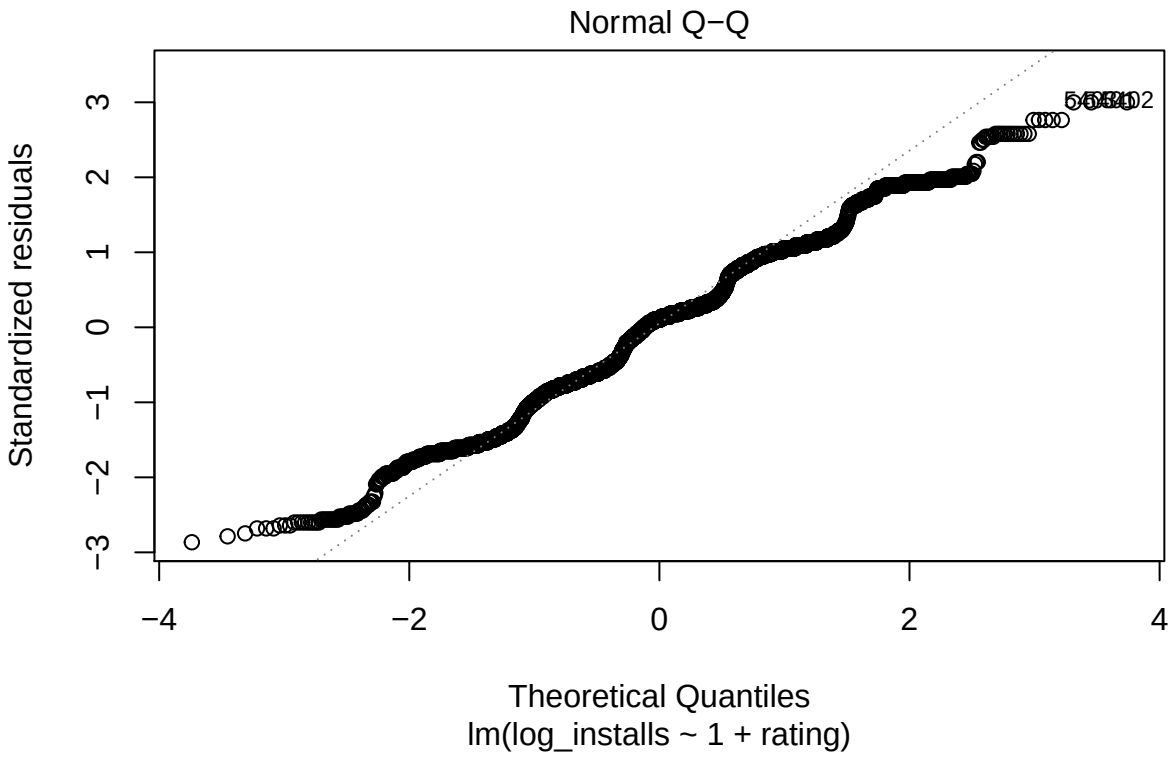
```r
  type = "text",
  se = list(get_robust_se(model_small), get_robust_se(model_medium),
            get_robust_se(model_large))
)
```

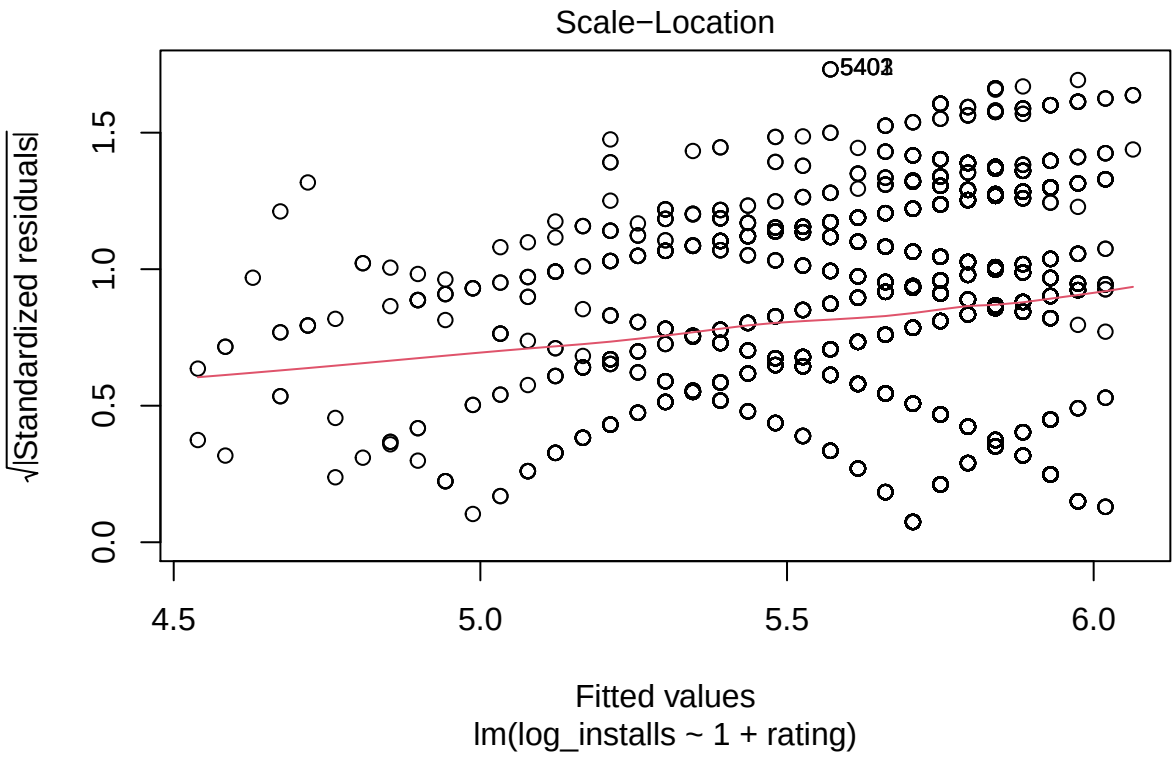
```
##
## ===========================================================================================================
##                                                     Dependent variable:
##                             -------------------------------------------------------------------------------
##                                                        log_installs
##                                         (1)                        (2)                        (3)
## -----------------------------------------------------------------------------------------------------------
## rating                               0.448***                   0.338***                   0.288***
##                                      (0.032)                    (0.031)                    (0.039)
##
## log_size                                                        0.432***                   0.429***
##                                                                 (0.033)                    (0.033)
##
## log_current_version                                             0.430***                   0.428***
##                                                                 (0.057)                    (0.057)
##
## log_last_updated                                               -0.909***                  -0.879***
##                                                                 (0.071)                    (0.071)
##
## is_free                                                         1.418***                   1.423***
##                                                                 (0.050)                    (0.051)
##
## is_family_category                                               0.032                      0.384
##                                                                 (0.037)                    (0.333)
##
## is_game_category                                                0.624***                  -2.374***
##                                                                 (0.046)                    (0.450)
##
## is_tools_category                                               0.272***                   -0.125
##                                                                 (0.054)                    (0.423)
##
## rating:is_family_category                                                                  -0.085
##                                                                                            (0.080)
##
## rating:is_game_category                                                                    0.702***
##                                                                                            (0.106)
##
## rating:is_tools_category                                                                    0.095
##                                                                                            (0.104)
##
## is_content_everyone                                              -0.067*                    -0.066*
##                                                                 (0.035)                    (0.035)
##
## Constant                             3.822***                   2.261***                   2.469***
##                                      (0.133)                    (0.154)                    (0.181)
##
## -----------------------------------------------------------------------------------------------------------
## Observations                           5,408                      5,408                      5,408
## R2                                     0.028                      0.265                      0.271
```

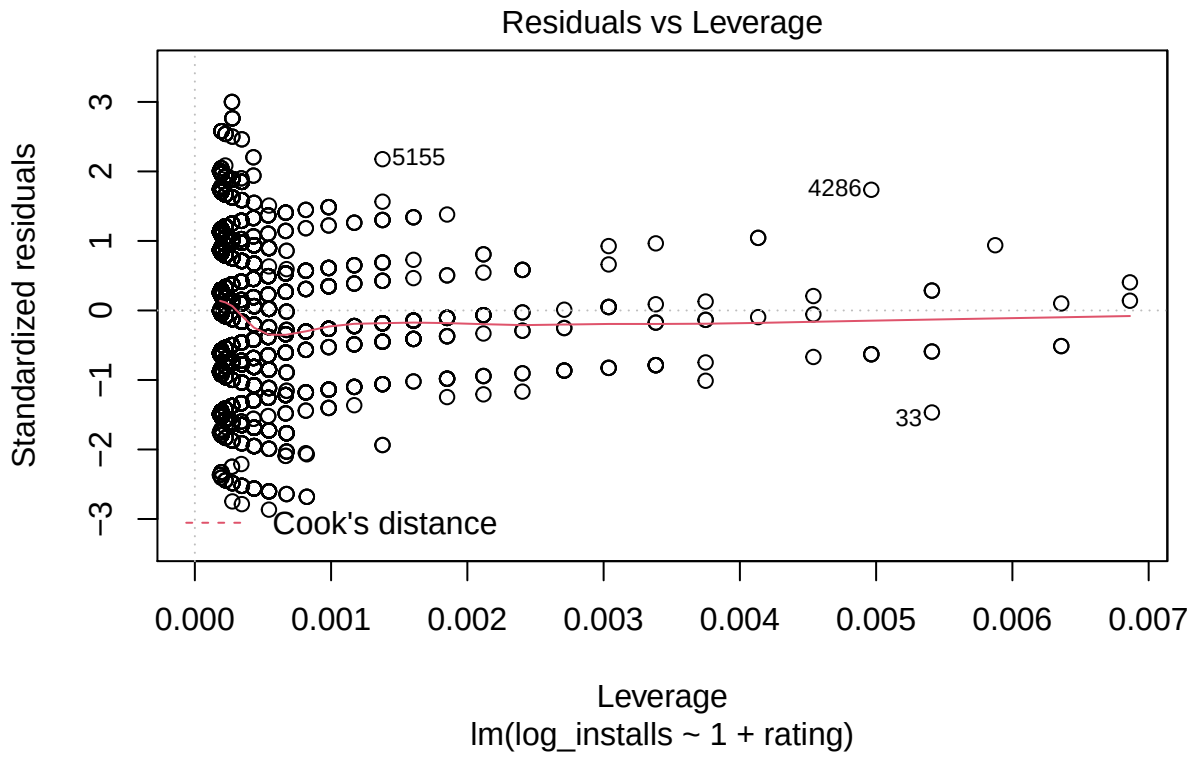
```
## Adjusted R2                          0.028                    0.264                    0.269
## Residual Std. Error            1.143 (df = 5406)        0.995 (df = 5398)        0.991 (df = 5395)
## F Statistic                155.421*** (df = 1; 5406) 216.073*** (df = 9; 5398) 166.929*** (df = 12; 5
## ==============================================================================================
## Note:                                                                        *p<0.1; **p<0.05; ***p<
```

```
plot(model_small)
```

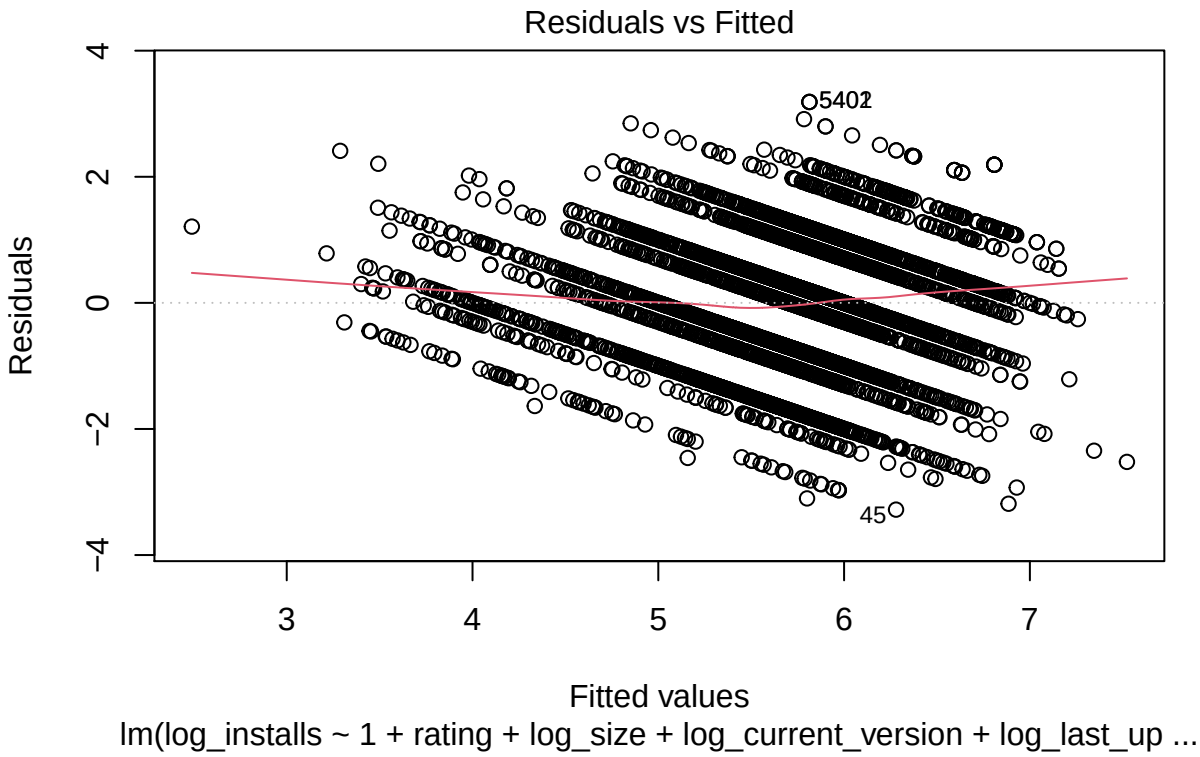


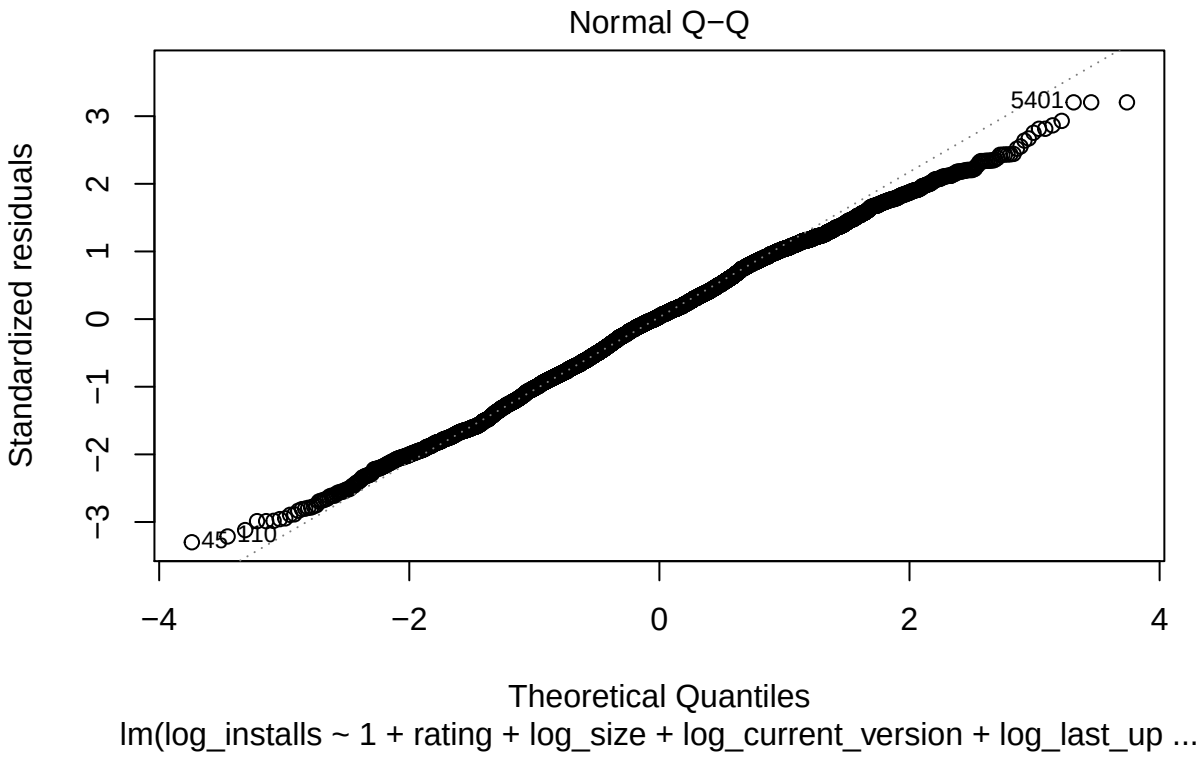

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_installs ~ 1 + rating)

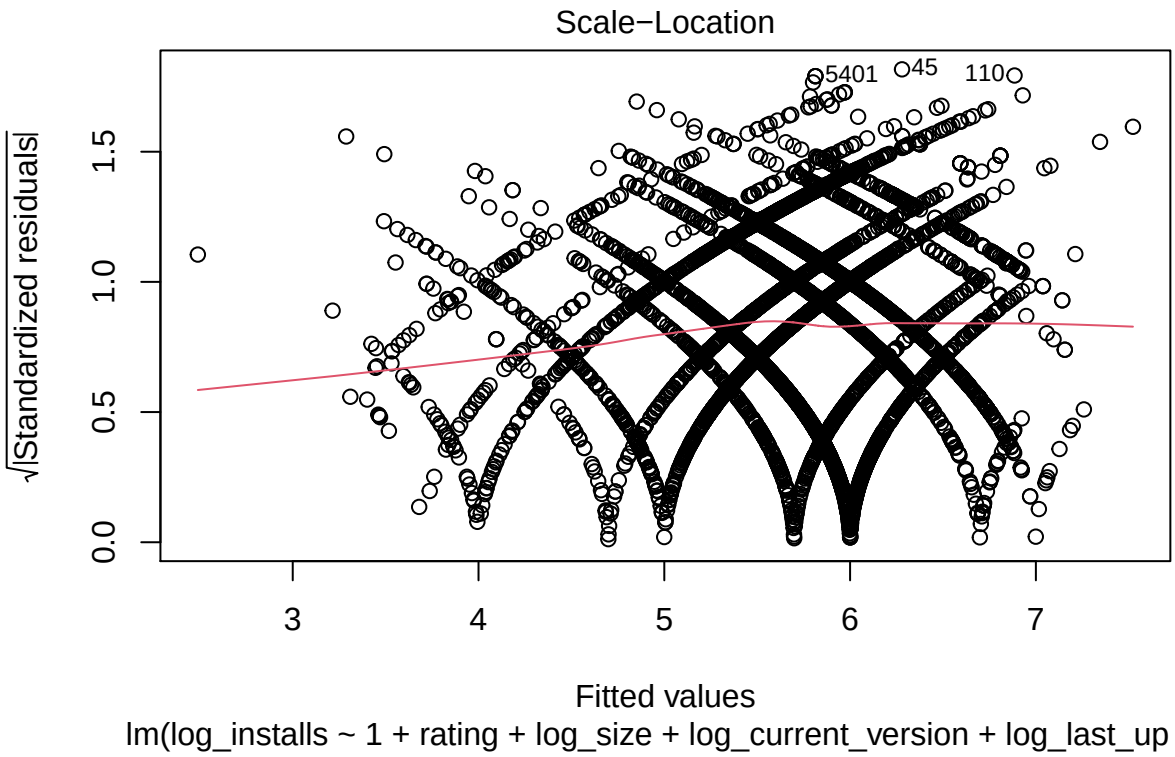

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
4.5
5.0
5.5
6.0
Fitted values
lm(log_installs ~ 1 + rating)

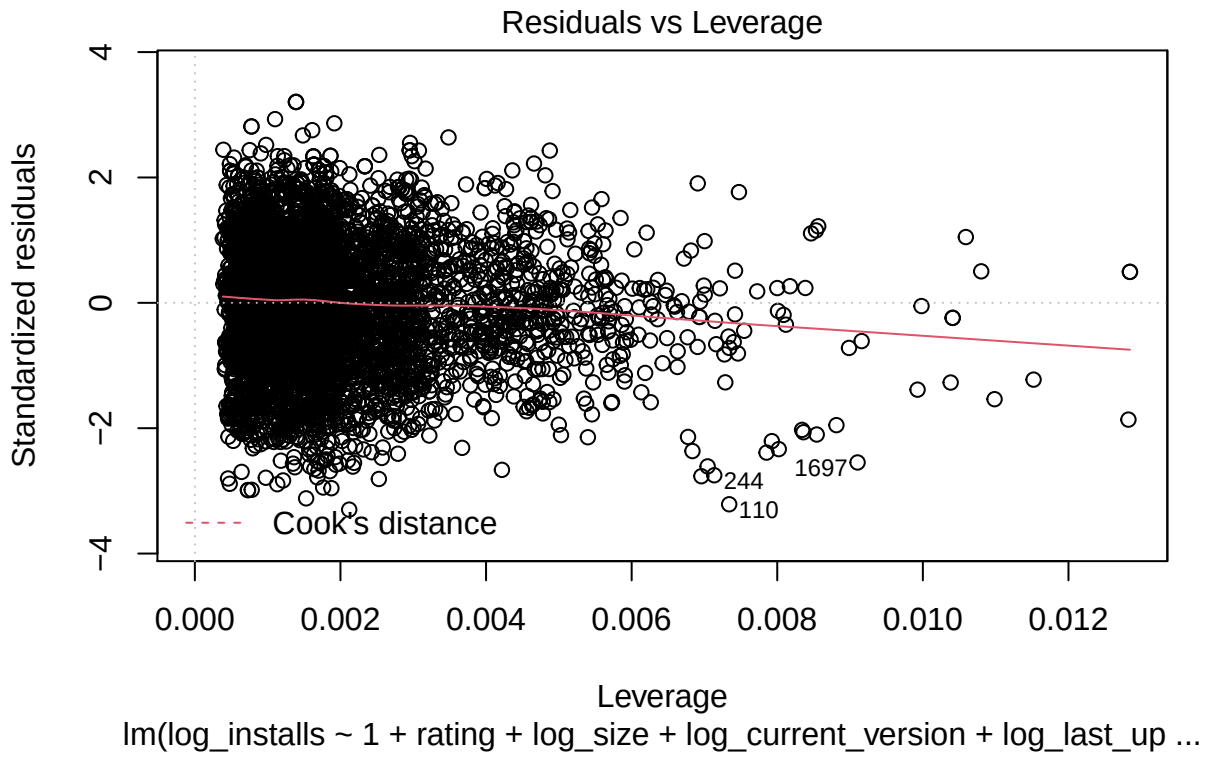



```
plot(model_medium)
```

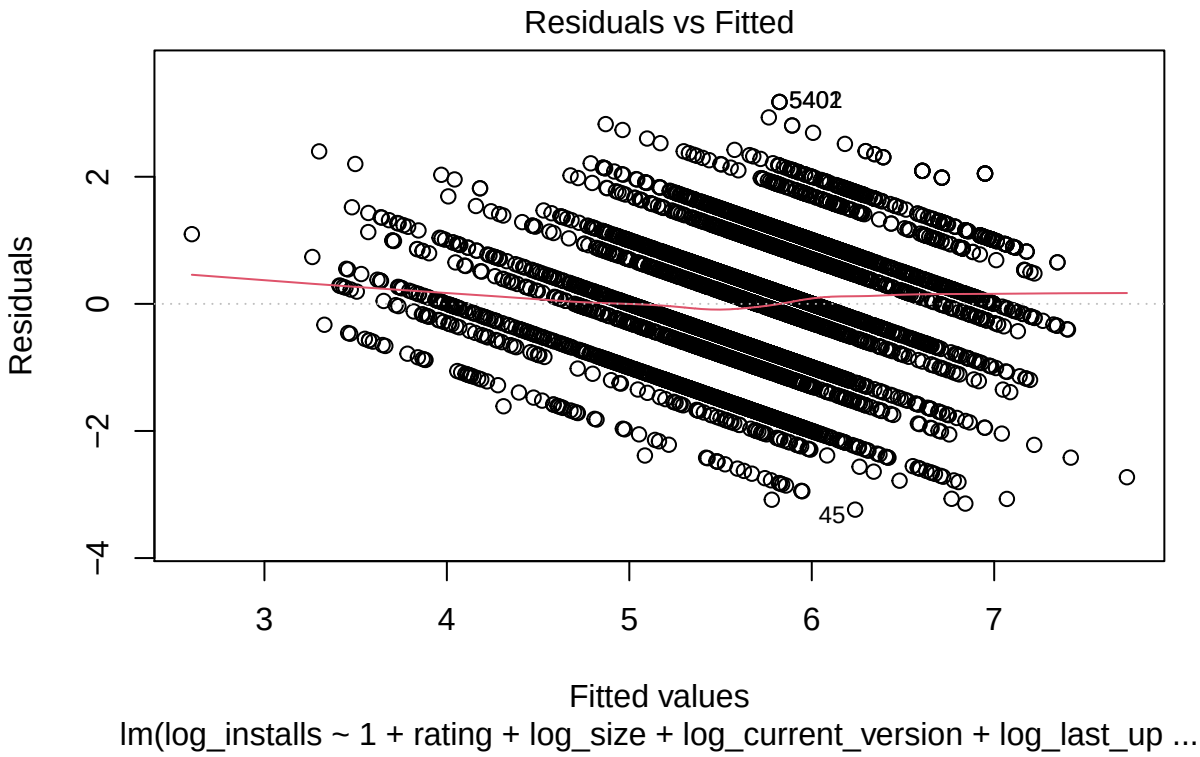

Residuals vs Fitted
Residuals
Fitted values
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...
5402
45
4
2
0
−2
−4
3
4
5
6
7

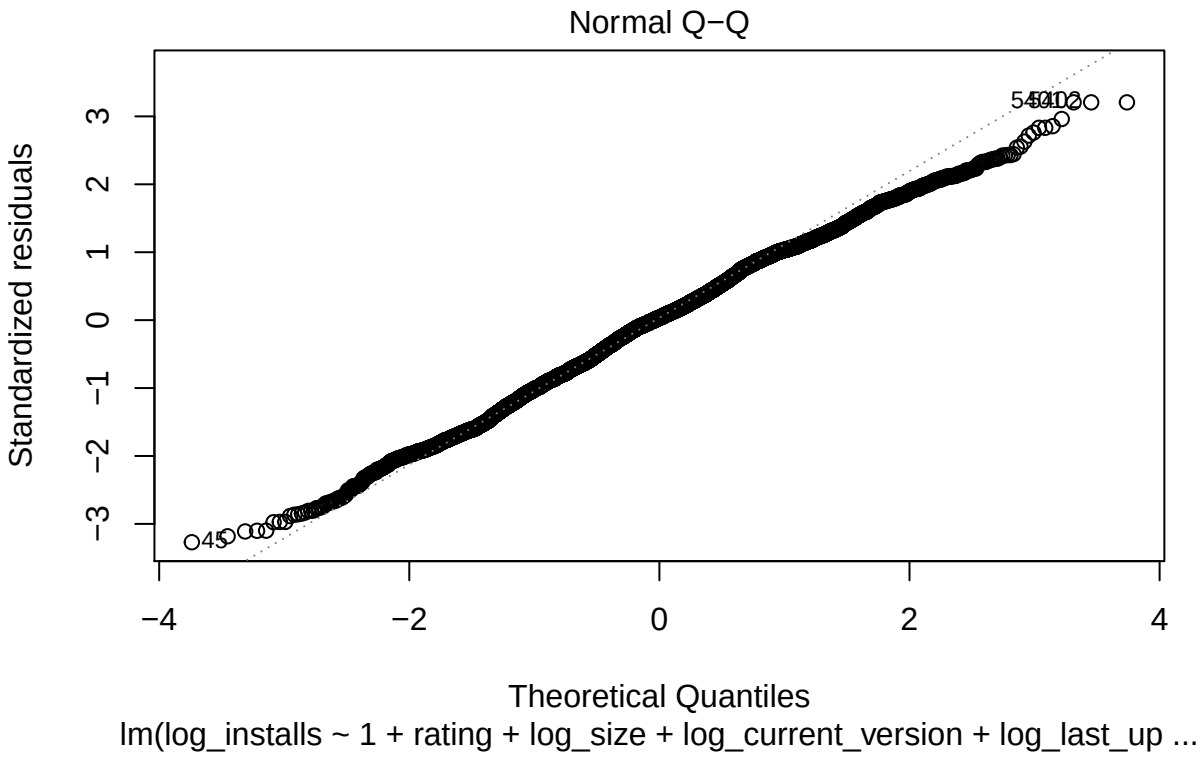

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...
5401
45
110
−3
−2
−1
0
1
2
3
−4
−2
0
2
4

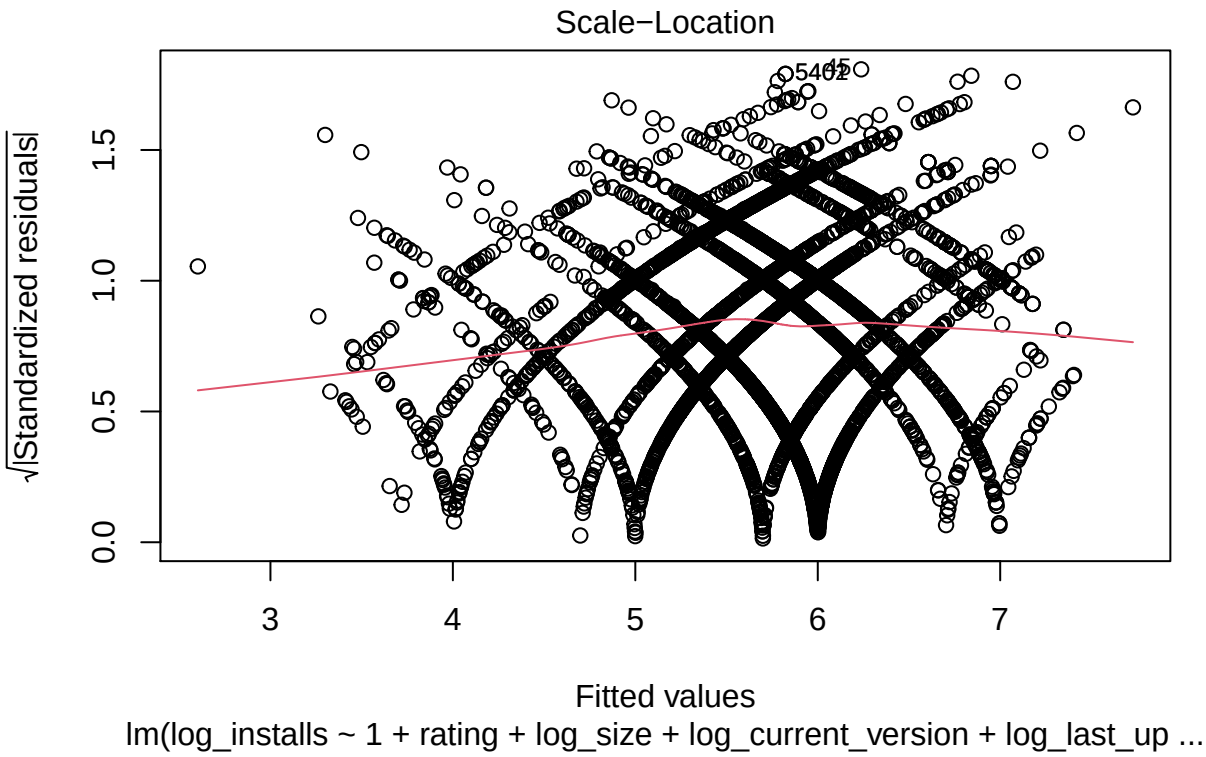

Scale−Location
5401
45
110
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
Fitted values
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...

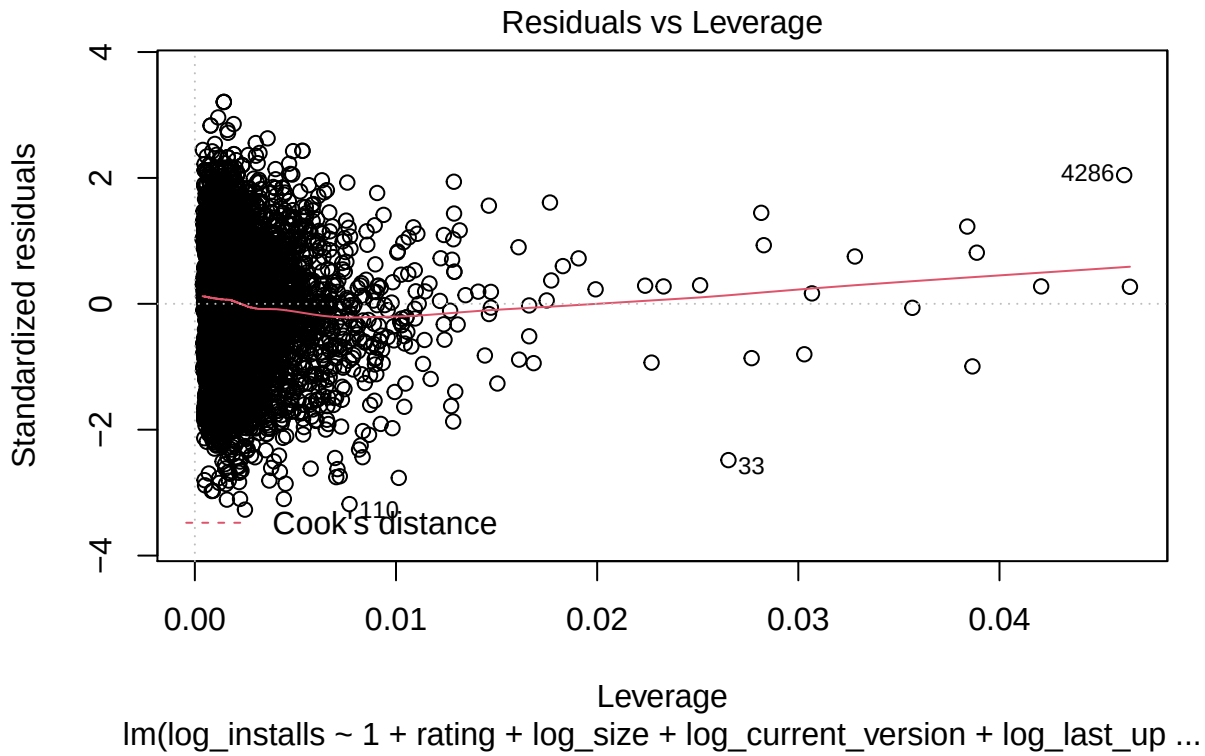



```
plot(model_large)
```

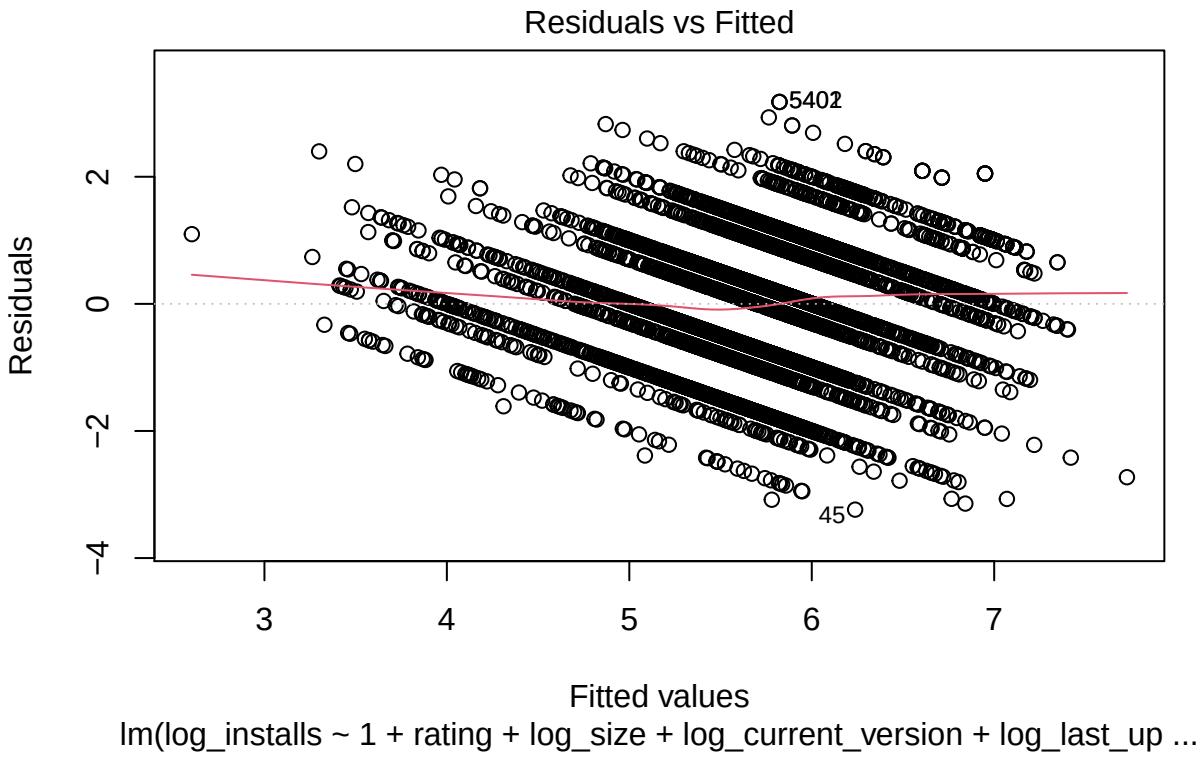

Residuals vs Fitted
Residuals
Fitted values
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...
5402
45
−4
−2
0
2
3
4
5
6
7

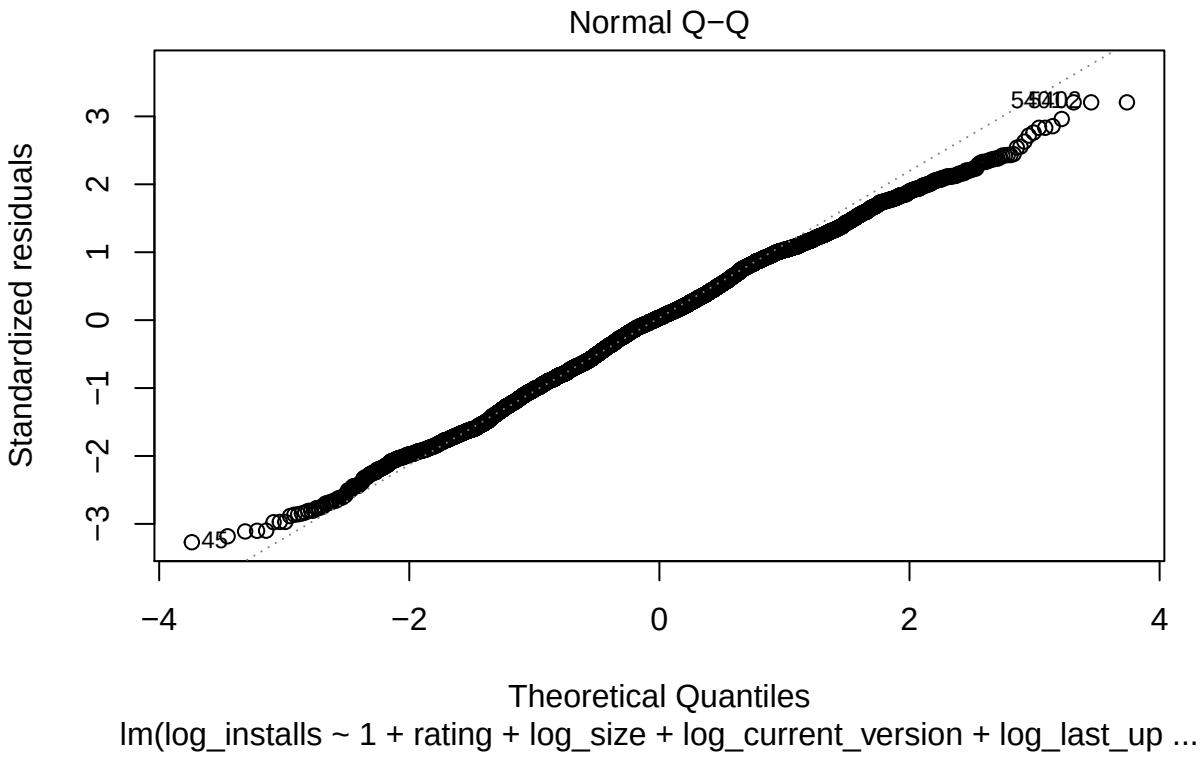

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...

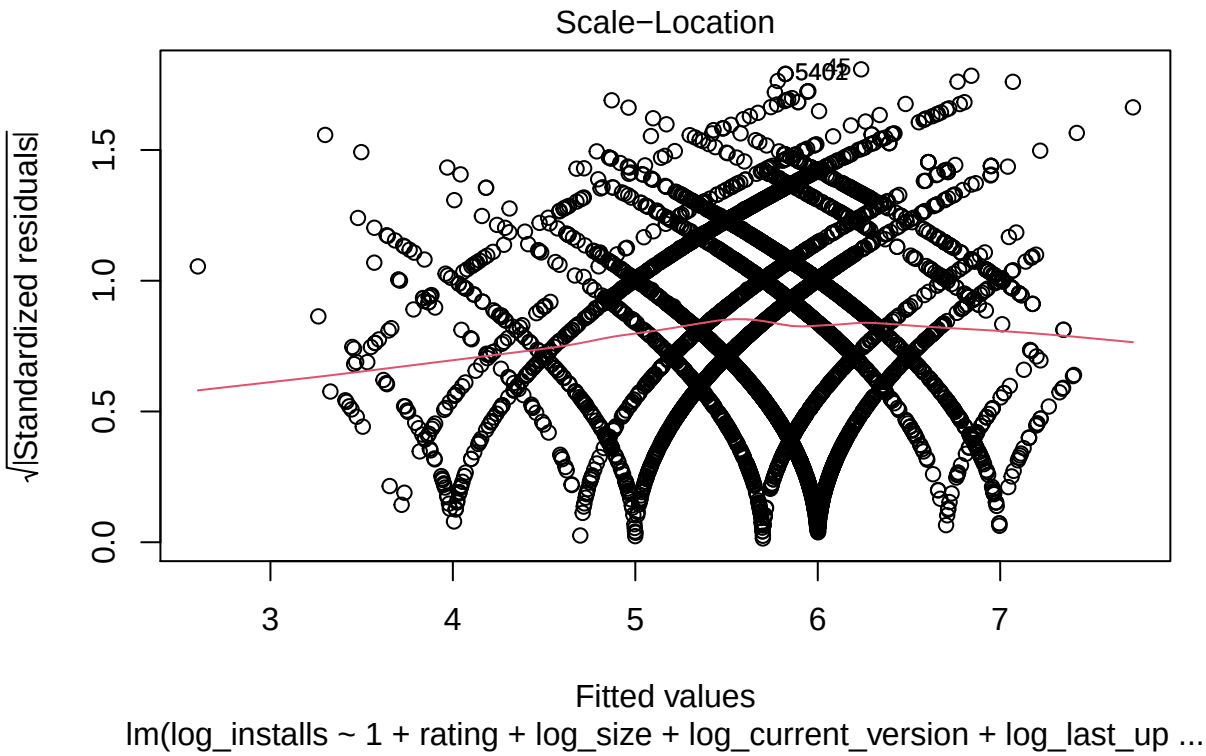

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
Fitted values
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...

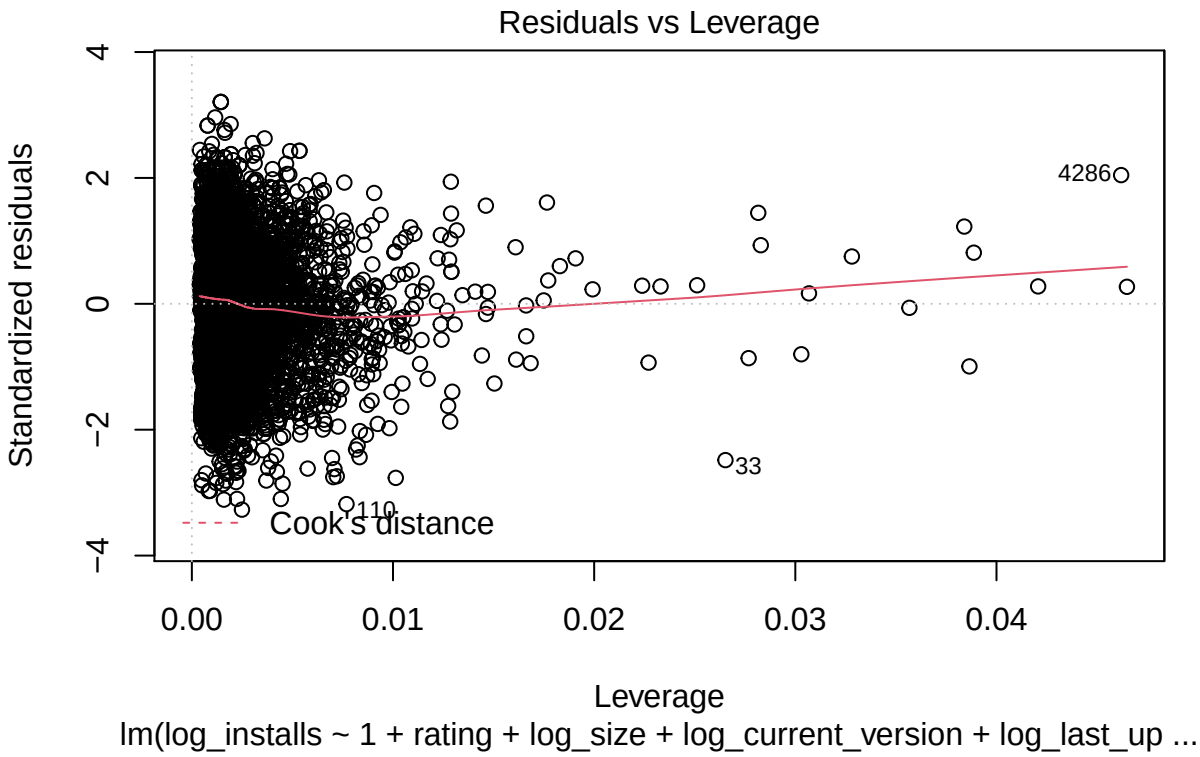



```
plot(model_large)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...
5402
45
−4
−2
0
2
3
4
5
6
7


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...
−4
−2
0
2
4
−3
−2
−1
0
1
2
3
45


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
Fitted values
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...


Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
4286
33
110
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(log_installs ~ 1 + rating + log_size + log_current_version + log_last_up ...