# Project Part 2

```
library(readr)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v purrr     1.0.2
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(lme4)
```

```
Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
library(nlme)
```

```
Attaching package: 'nlme'
```

```
The following object is masked from 'package:lme4':

    lmList

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

    collapse
```

```r
library(performance)
library(ggplot2)

df_baseball <- read_csv("https://www.dropbox.com/scl/fi/2bcvc8eabdinum2e3r9oj/statcast_pit
```

```
Rows: 346250 Columns: 94
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (16): pitch_type, player_name, events, description, des, game_type, sta...
dbl  (69): release_speed, release_pos_x, release_pos_z, batter, pitcher, zon...
lgl   (8): spin_dir, spin_rate_deprecated, break_angle_deprecated, break_len...
date  (1): game_date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# only keep variables of interest, drop NA obs.
df_baseball_clean <- df_baseball |>
  mutate(dist_from_cen = sqrt((plate_x^2) + (plate_z^2)),
         pitch_type = factor(pitch_type),
         game_type = factor(game_type),
         batter = factor(batter),
         stand = factor(stand),
         type = factor(type)) |>
  select(type, hit_distance_sc, pitch_type, dist_from_cen, launch_angle, launch_speed, gam
  filter(type == "X") |>
  drop_na() |>
  mutate(dist_from_cen_gmc = dist_from_cen - mean(dist_from_cen),
         hit_distance_gmc = hit_distance_sc - mean(hit_distance_sc),
         launch_angle_gmc = launch_angle - mean(launch_angle),
         launch_speed_gmc = launch_speed - mean(launch_speed))
```

```
df_baseball_clean
```

```
# A tibble: 60,811 x 15
   type  hit_distance_sc pitch_type dist_from_cen launch_angle launch_speed
   <fct>           <dbl> <fct>              <dbl>        <dbl>        <dbl>
 1 X                 180 FF                 3.25           28         61.9
 2 X                   1 KC                 1.52          -62         31.7
 3 X                  30 SI                 2.85           -4        105
 4 X                  51 FC                 2.85            0        107
 5 X                 279 ST                 2.39           38         84.6
 6 X                  88 SI                 2.10            3        106.
 7 X                 352 FF                 3.42           25         99.4
 8 X                 110 SL                 2.17            5         93
 9 X                 282 FF                 3.43           45         90.2
10 X                 379 FF                 2.73           27         97.6
# i 60,801 more rows
# i 9 more variables: game_type <fct>, batter <fct>, stand <fct>,
#   plate_x <dbl>, plate_z <dbl>, dist_from_cen_gmc <dbl>,
#   hit_distance_gmc <dbl>, launch_angle_gmc <dbl>, launch_speed_gmc <dbl>
```

```
save(df_baseball_clean, file = "df_baseball_clean.RData")
```

'''modelgt <- lm(hit_distance_sc ~ game_type, data = df_baseball_clean) summary(modelgt)

modelgt |> ggplot(aes(x=game_type, y=hit_distance_sc)) + geom_boxplot()'''

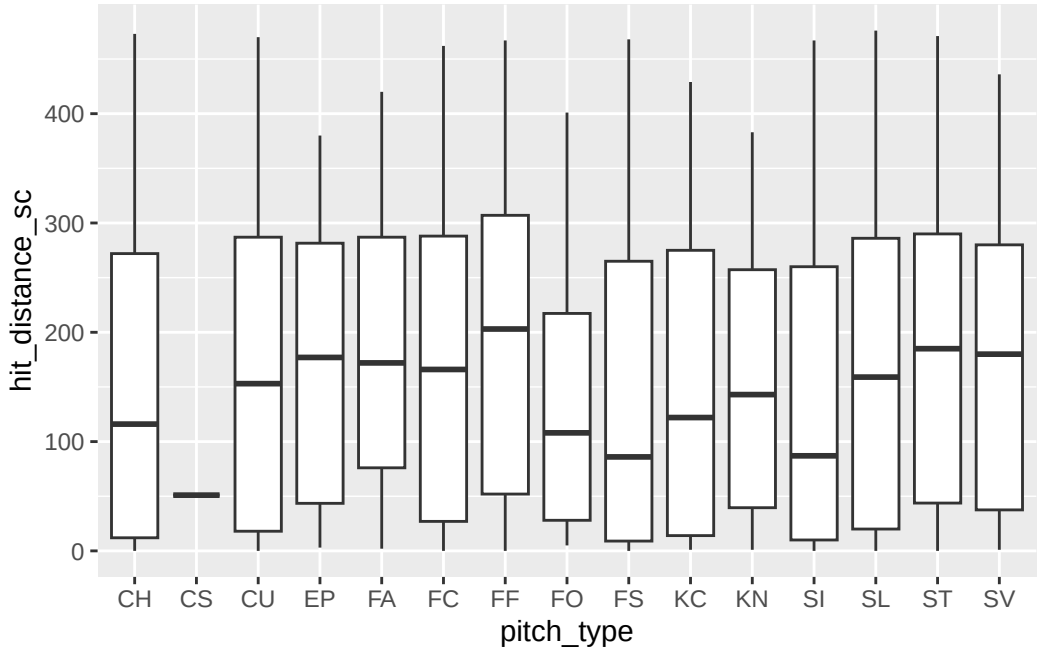
This is how we realized that game_type wouldn't do anything for us

```
modelpt <-  lm(hit_distance_sc ~ pitch_type, data = df_baseball_clean)
summary(modelpt)
```

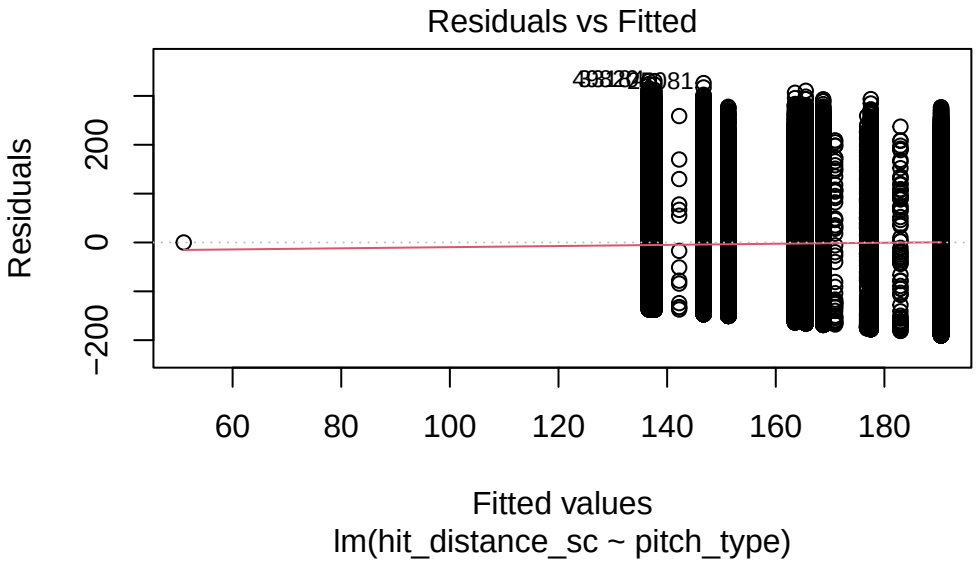
```
Call:
lm(formula = hit_distance_sc ~ pitch_type, data = df_baseball_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-190.43 -131.43  -10.43  119.50  330.36

Coefficients:
```

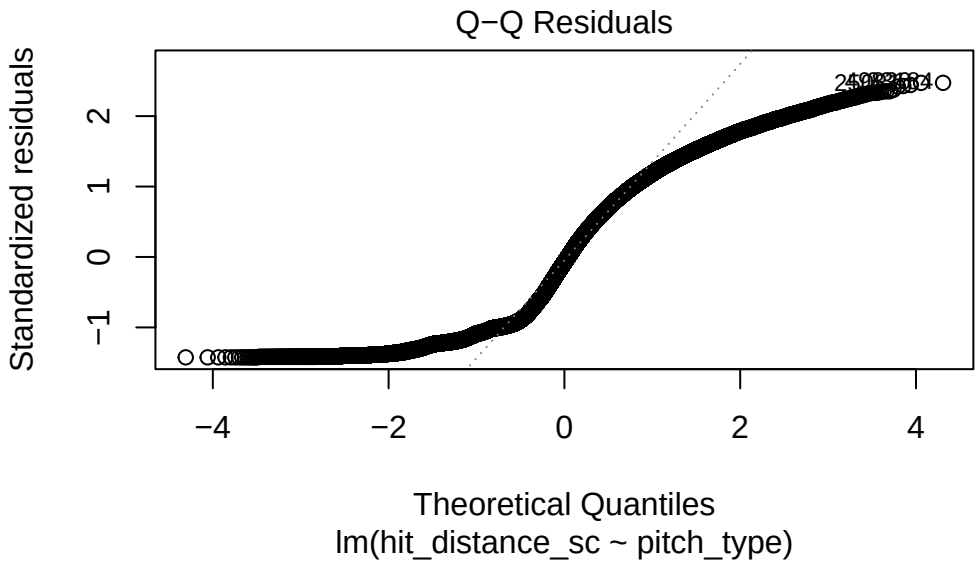
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   146.711      1.603  91.540  < 2e-16 ***
pitch_typeCS  -95.711    133.650  -0.716  0.47391
pitch_typeCU   16.830      2.817   5.974 2.33e-09 ***
pitch_typeEP   24.226     19.356   1.252  0.21072
pitch_typeFA   36.246     15.941   2.274  0.02298 *
pitch_typeFC   22.026      2.457   8.964  < 2e-16 ***
pitch_typeFF   43.722      1.888  23.161  < 2e-16 ***
pitch_typeFO   -4.497     35.753  -0.126  0.89990
pitch_typeFS   -9.051      3.444  -2.628  0.00859 **
pitch_typeKC    4.542      4.998   0.909  0.36346
pitch_typeKN    4.639     15.027   0.309  0.75757
pitch_typeSI  -10.075      2.022  -4.982 6.31e-07 ***
pitch_typeSL   18.811      2.117   8.886  < 2e-16 ***
pitch_typeST   30.787      2.846  10.816  < 2e-16 ***
pitch_typeSV   30.105      9.802   3.071  0.00213 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.6 on 60796 degrees of freedom
Multiple R-squared:  0.02307,   Adjusted R-squared:  0.02285
F-statistic: 102.6 on 14 and 60796 DF,  p-value: < 2.2e-16
```

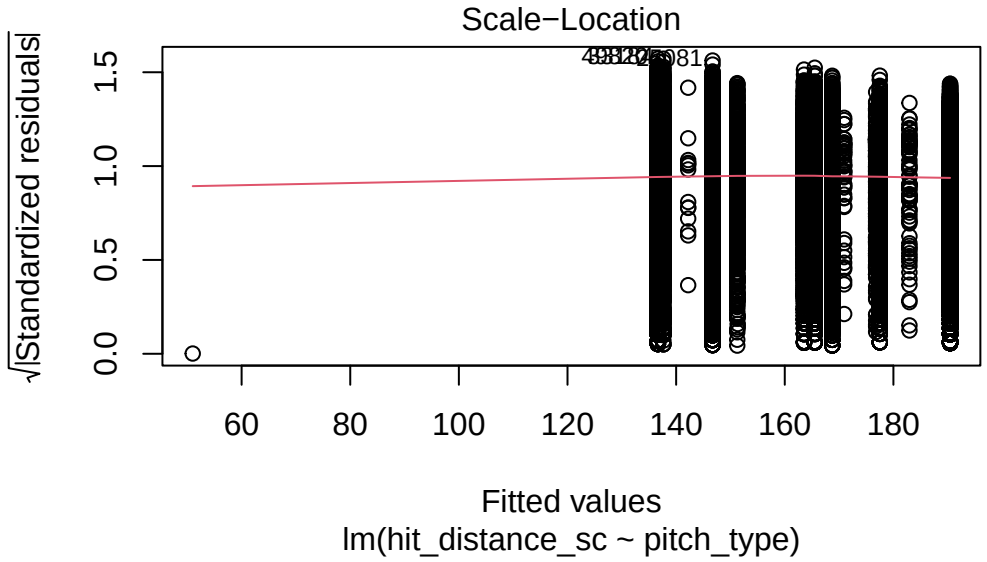
```r
modelpt |> ggplot(aes(x=pitch_type, y=hit_distance_sc)) +
  geom_boxplot()
```



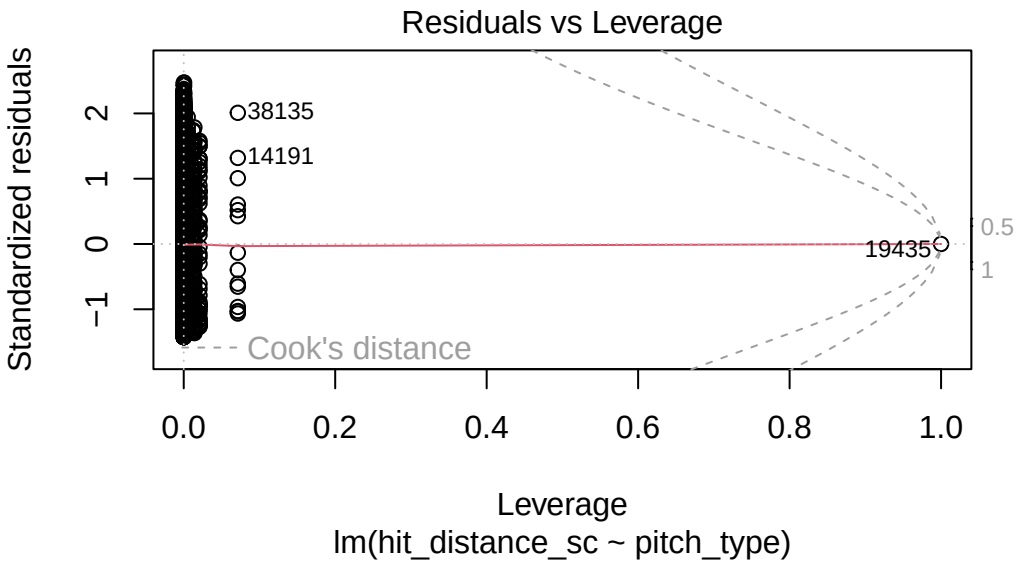

```
plot(modelpt)
```





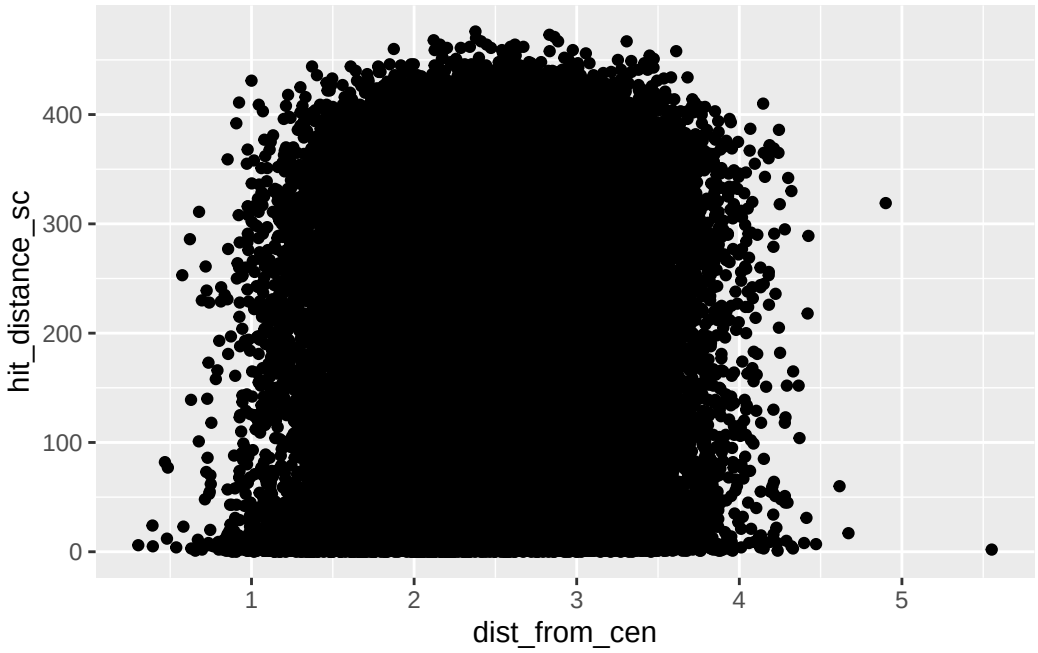

```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



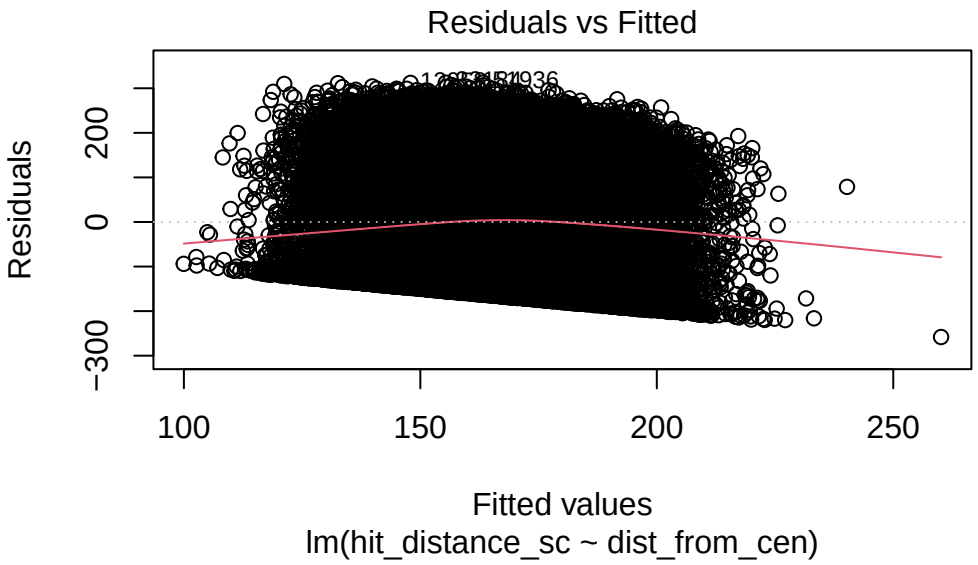

```r
modeldfg <- lm(hit_distance_sc ~ dist_from_cen, data = df_baseball_clean)
summary(modeldfg)
```

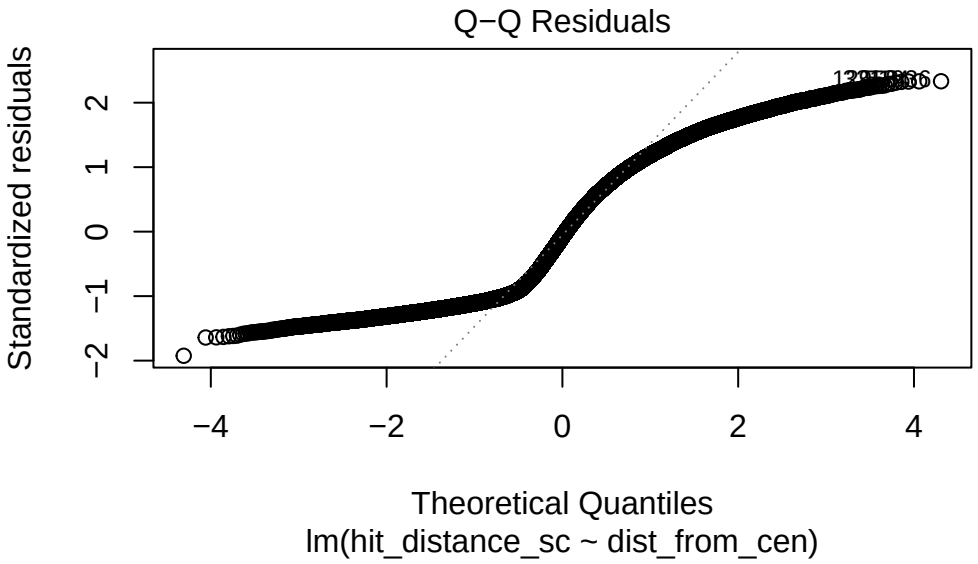
```
Call:
lm(formula = hit_distance_sc ~ dist_from_cen, data = df_baseball_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-258.11 -135.30  -10.34  121.42  312.76

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    90.7003     2.4312   37.31   <2e-16 ***
dist_from_cen  30.5120     0.9742   31.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 134.1 on 60809 degrees of freedom
Multiple R-squared:  0.01588,   Adjusted R-squared:  0.01586
F-statistic: 980.9 on 1 and 60809 DF,  p-value: < 2.2e-16
```

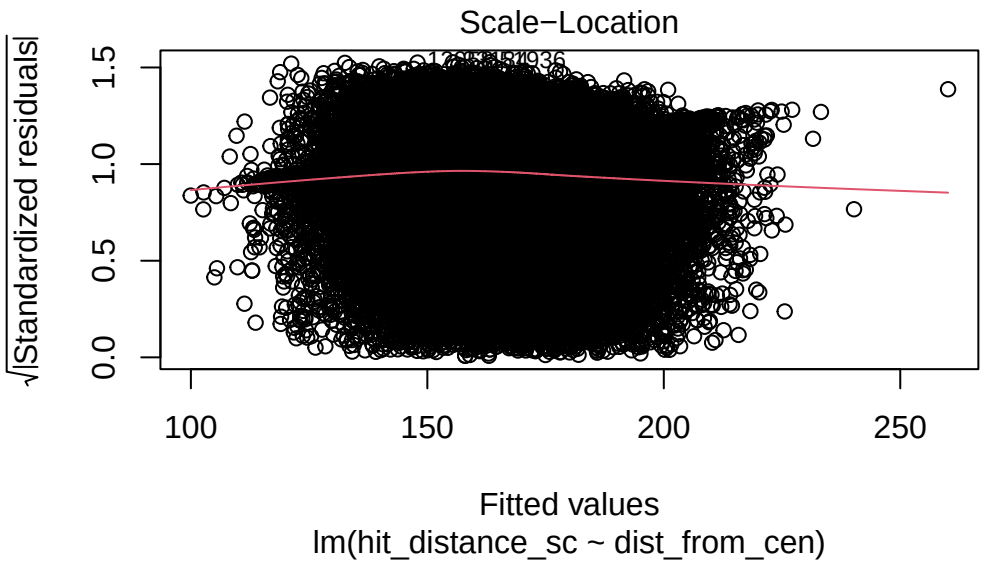
```
modeldfg |> ggplot(aes(x=dist_from_cen, y=hit_distance_sc)) +
  geom_point()
```

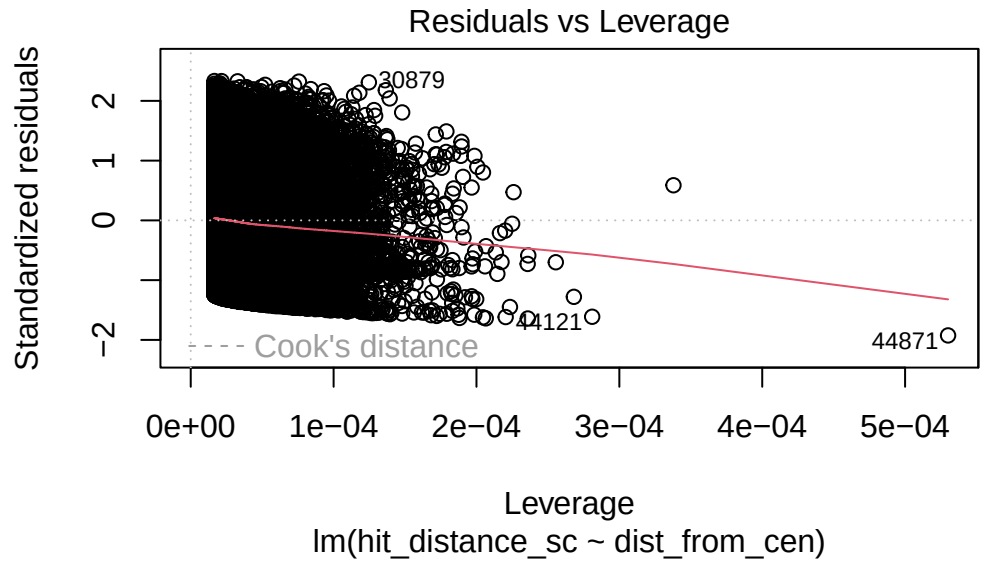



```
plot(modeldfg)
```


Residuals vs Fitted
Residuals
200
0
−300
100
150
200
250
Fitted values
lm(hit_distance_sc ~ dist_from_cen)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−4
−2
0
2
4
Theoretical Quantiles
lm(hit_distance_sc ~ dist_from_cen)

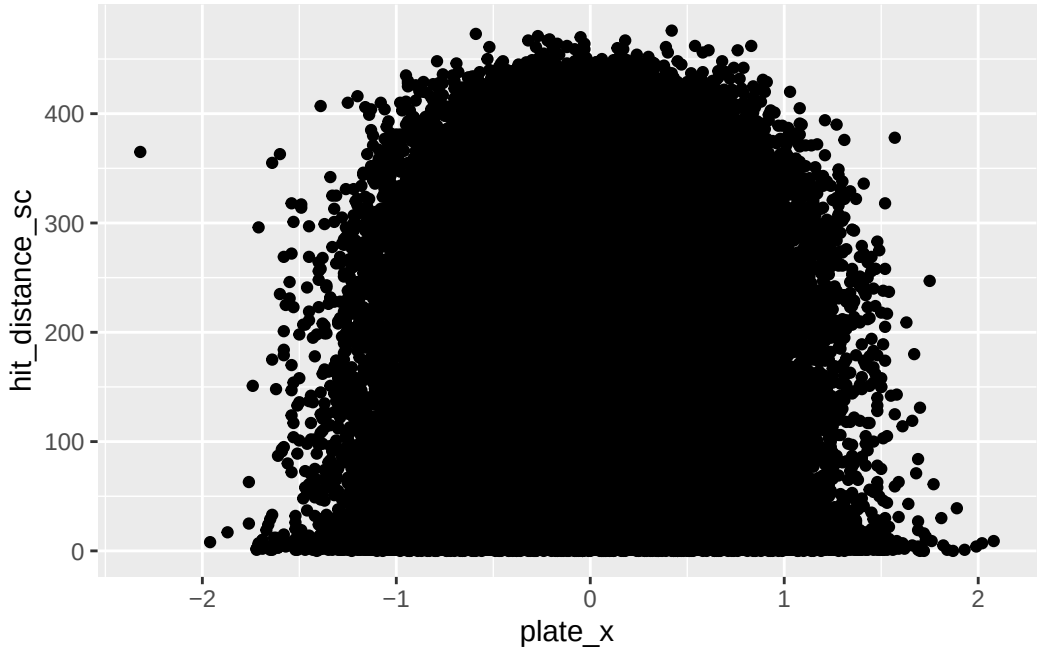

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
100
150
200
250
Fitted values
lm(hit_distance_sc ~ dist_from_cen)


Residuals vs Leverage
Standardized residuals
−2
0
1
2
30879
44121
44871
Cook's distance
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
Leverage
lm(hit_distance_sc ~ dist_from_cen)

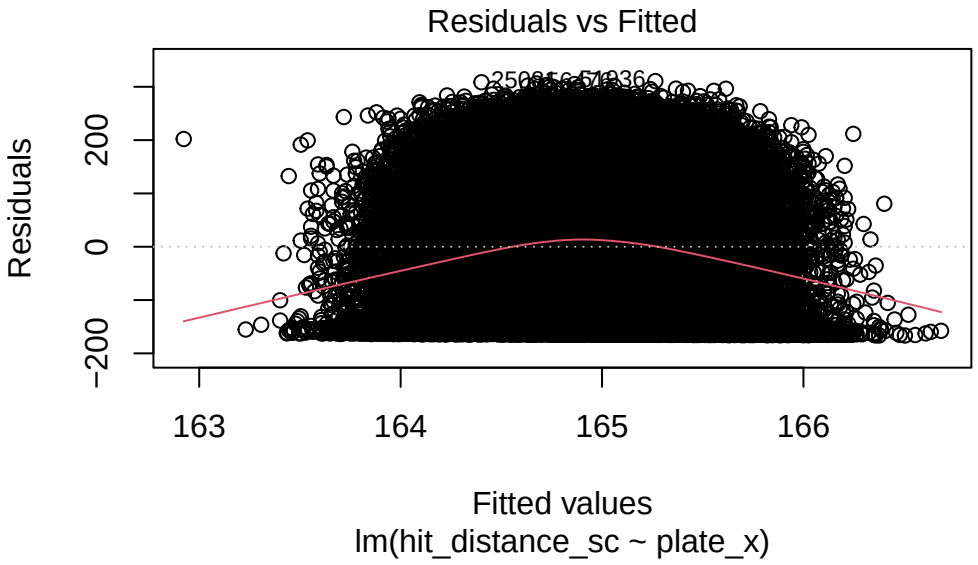
```r
modelxcoord <- lm(hit_distance_sc ~ plate_x, data = df_baseball_clean)
summary(modelxcoord)
```

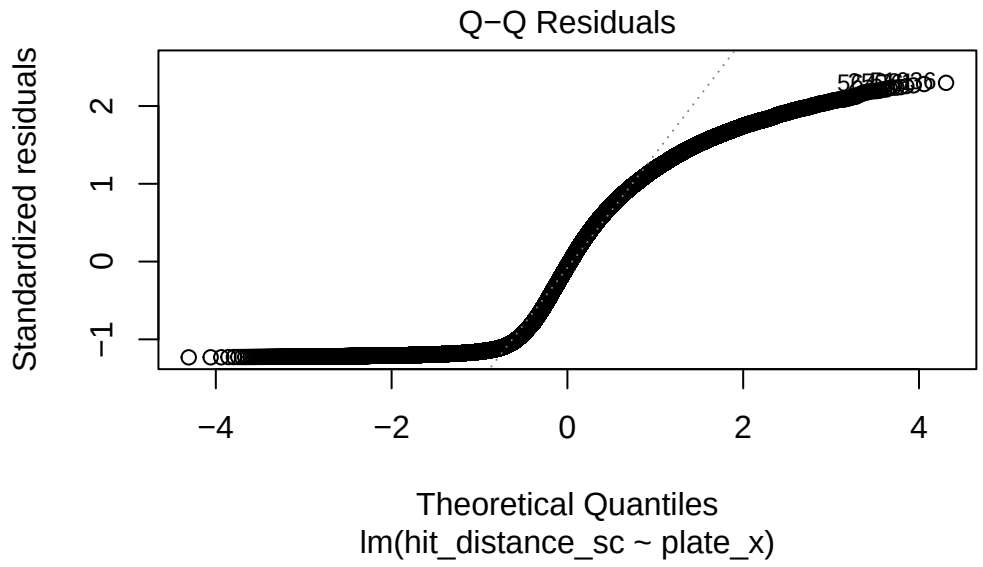
```
Call:
lm(formula = hit_distance_sc ~ plate_x, data = df_baseball_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-166.504 -144.666   -7.974  122.881  310.735

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 164.9058     0.5483 300.748   <2e-16 ***
plate_x       0.8546     1.0290   0.831    0.406
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 135.2 on 60809 degrees of freedom
Multiple R-squared:  1.134e-05, Adjusted R-squared:  -5.101e-06
F-statistic: 0.6898 on 1 and 60809 DF,  p-value: 0.4062
```

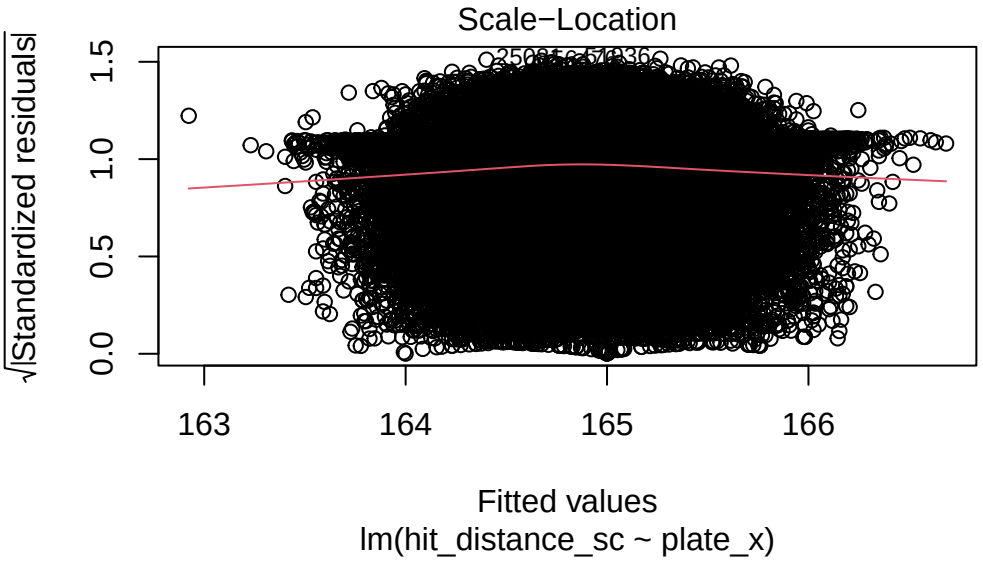
```r
modelxcoord |> ggplot(aes(x=plate_x, y=hit_distance_sc)) +
  geom_point()
```

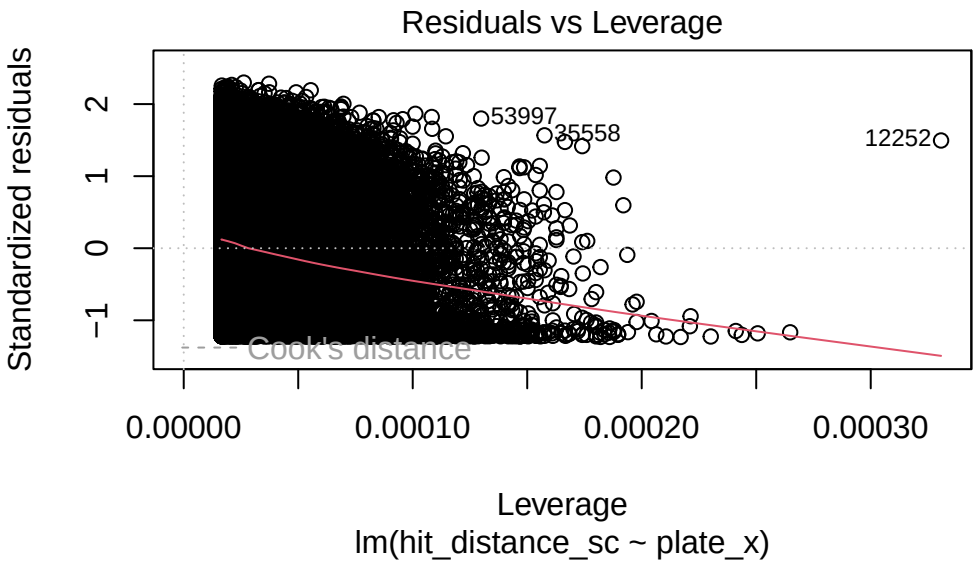



```
plot(modelxcoord)
```




Q−Q Residuals
Standardized residuals
−1
0
1
2
−4
−2
0
2
4
Theoretical Quantiles
lm(hit_distance_sc ~ plate_x)

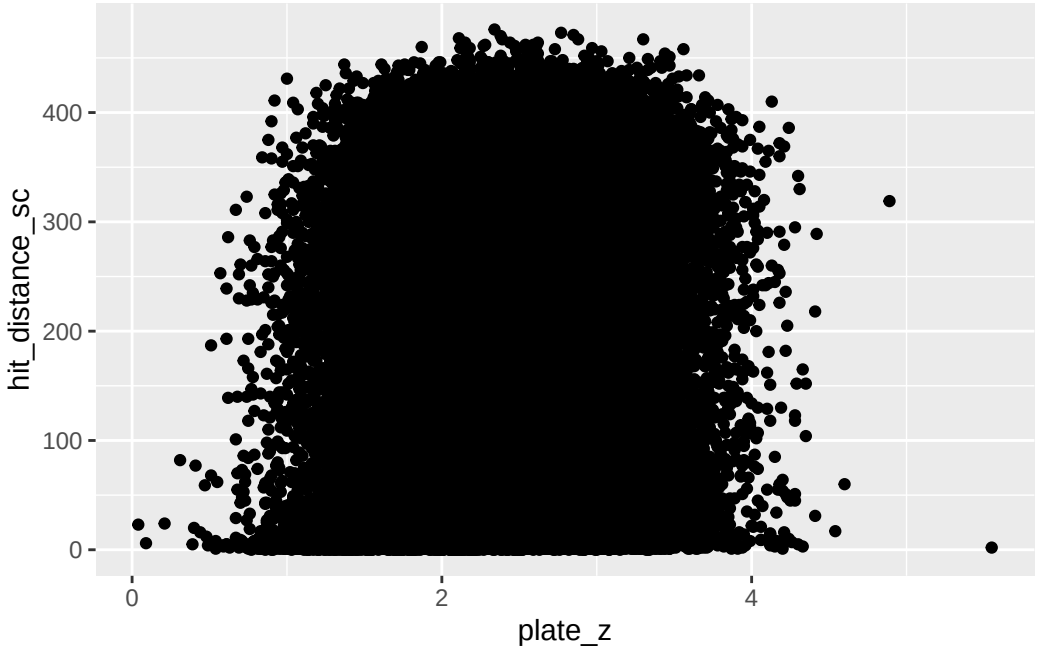

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
163
164
165
166
Fitted values
lm(hit_distance_sc ~ plate_x)



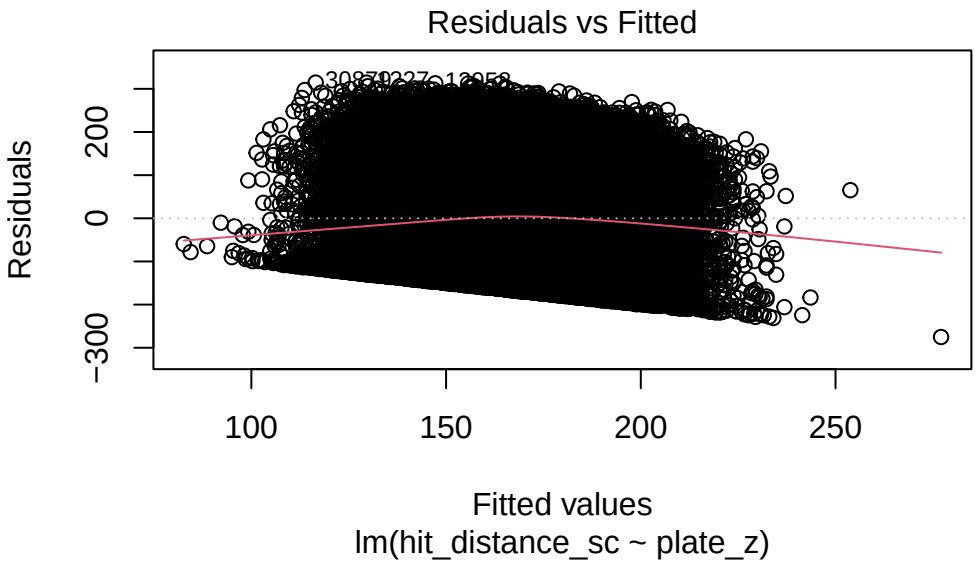

```r
modelzcoord <- lm(hit_distance_sc ~ plate_z, data = df_baseball_clean)
summary(modelzcoord)
```

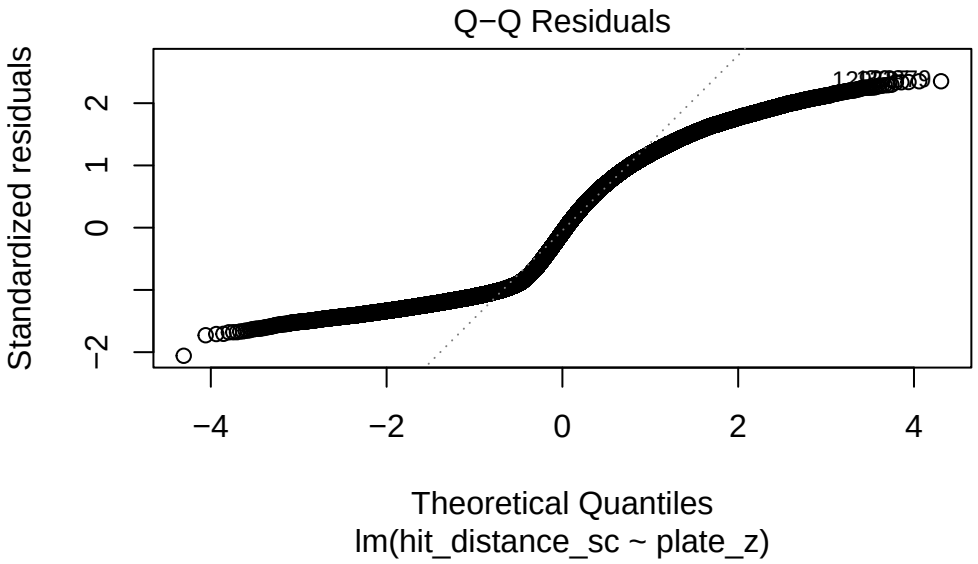
```
Call:
lm(formula = hit_distance_sc ~ plate_z, data = df_baseball_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-275.10 -132.98  -10.69  120.84  314.49

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.2214     2.3353   34.78   <2e-16 ***
plate_z      35.2934     0.9579   36.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.7 on 60809 degrees of freedom
Multiple R-squared:  0.02184,   Adjusted R-squared:  0.02182
F-statistic:  1358 on 1 and 60809 DF,  p-value: < 2.2e-16
```

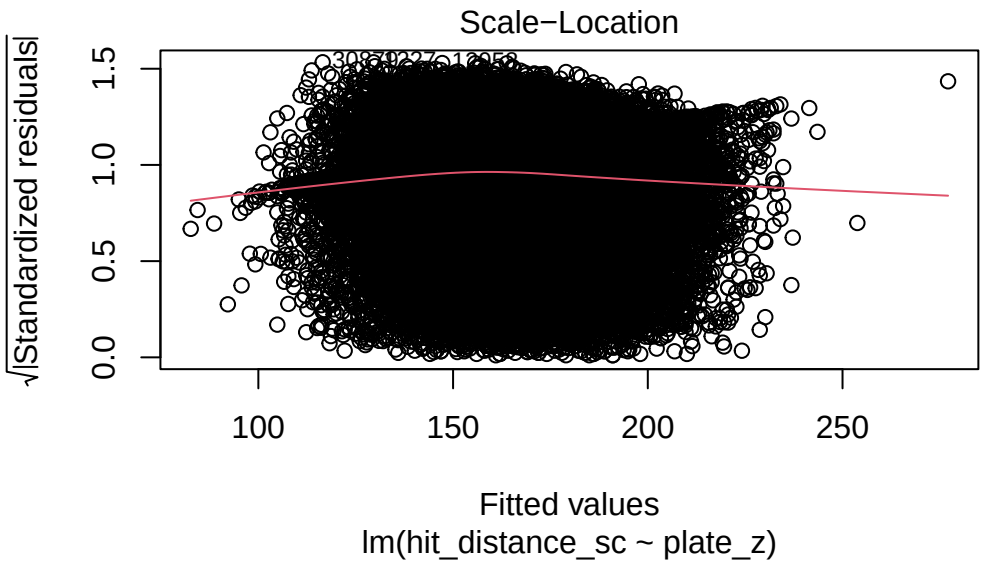
```
modelzcoord |> ggplot(aes(x=plate_z, y=hit_distance_sc)) +
  geom_point()
```

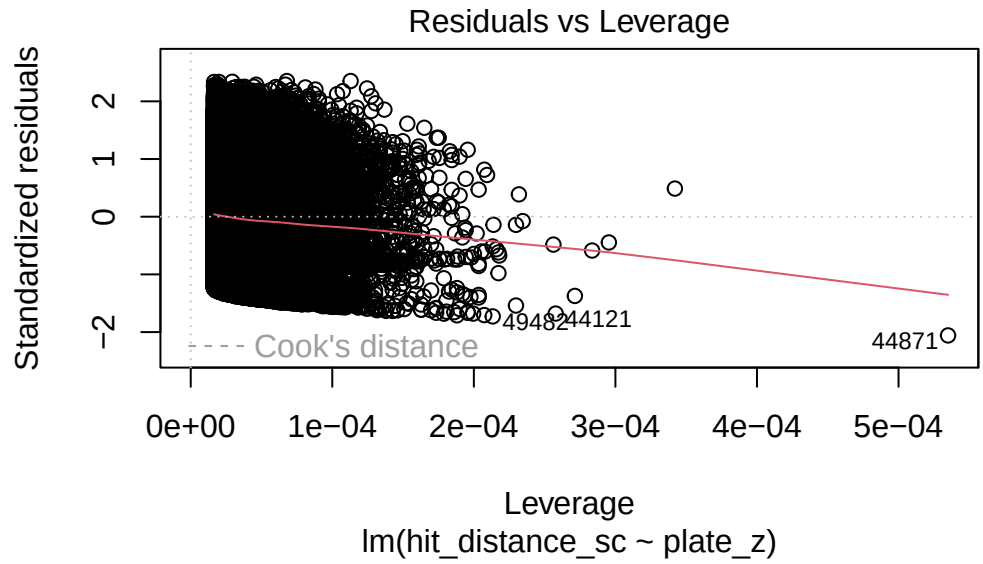



```
plot(modelzcoord)
```


Residuals vs Fitted
Residuals
200
0
−300
100
150
200
250
Fitted values
lm(hit_distance_sc ~ plate_z)


Q−Q Residuals
Standardized residuals
2
1
0
−2
−4
−2
0
2
4
Theoretical Quantiles
lm(hit_distance_sc ~ plate_z)

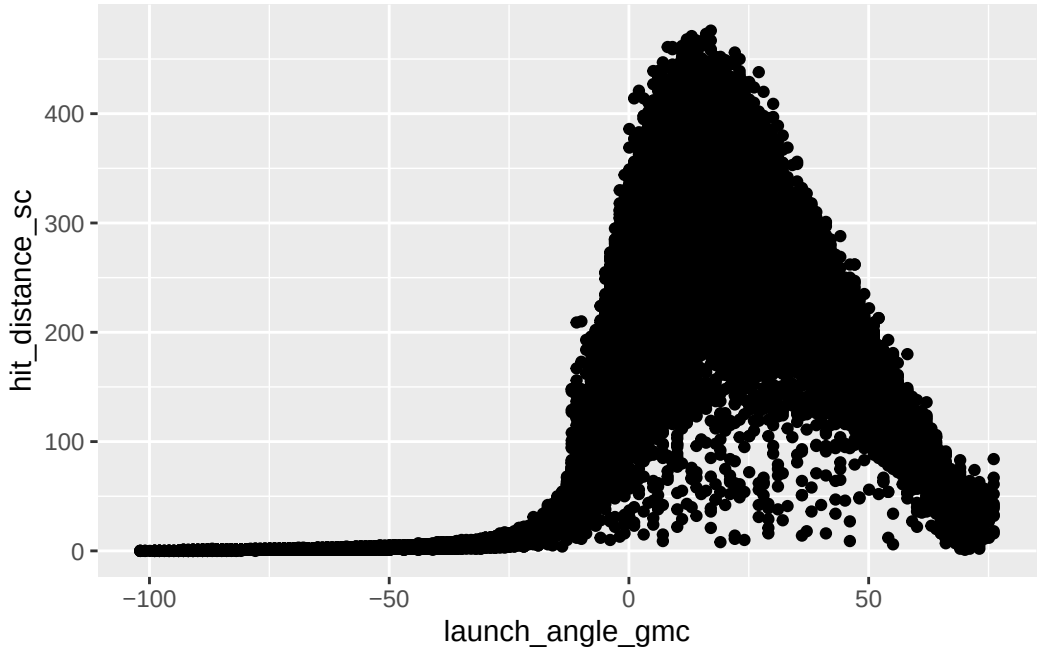

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
100
150
200
250
Fitted values
lm(hit_distance_sc ~ plate_z)


Residuals vs Leverage
Standardized residuals
−2
0
1
2
Cook's distance
44871
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
Leverage
lm(hit_distance_sc ~ plate_z)

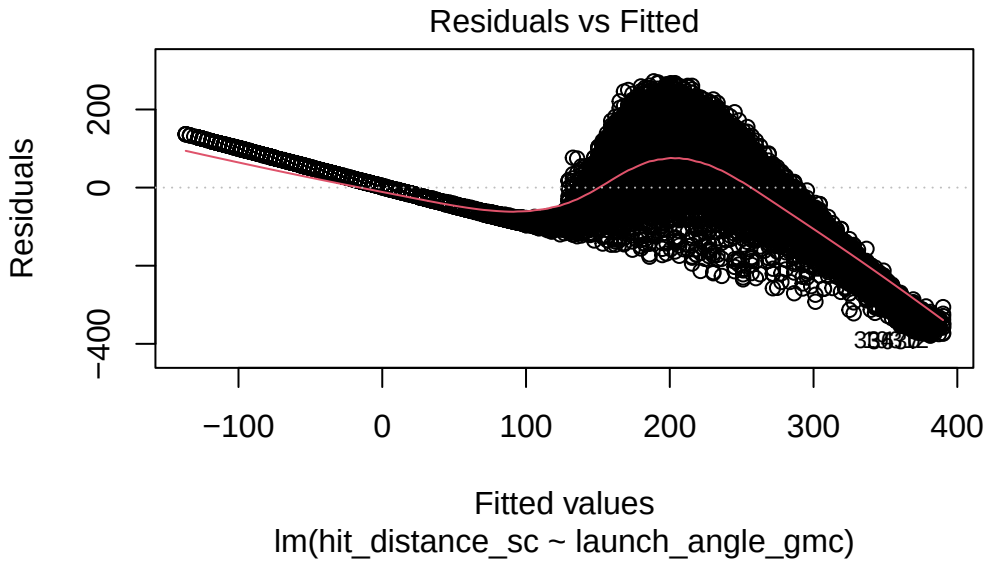
```r
modellag <- lm(hit_distance_sc ~ launch_angle_gmc, data = df_baseball_clean)
summary(modellag)
```

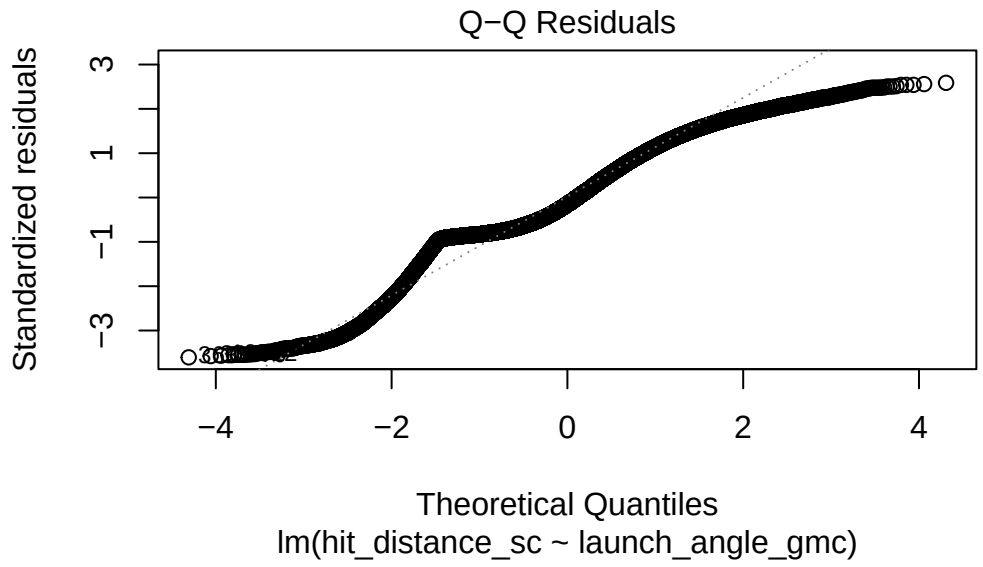
```
Call:
lm(formula = hit_distance_sc ~ launch_angle_gmc, data = df_baseball_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-379.21  -76.69  -15.78   81.48  272.16

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      164.91370    0.42656   386.6   <2e-16 ***
launch_angle_gmc   2.95957    0.01486   199.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 105.2 on 60809 degrees of freedom
Multiple R-squared:  0.3946,    Adjusted R-squared:  0.3946
F-statistic: 3.964e+04 on 1 and 60809 DF,  p-value: < 2.2e-16
```

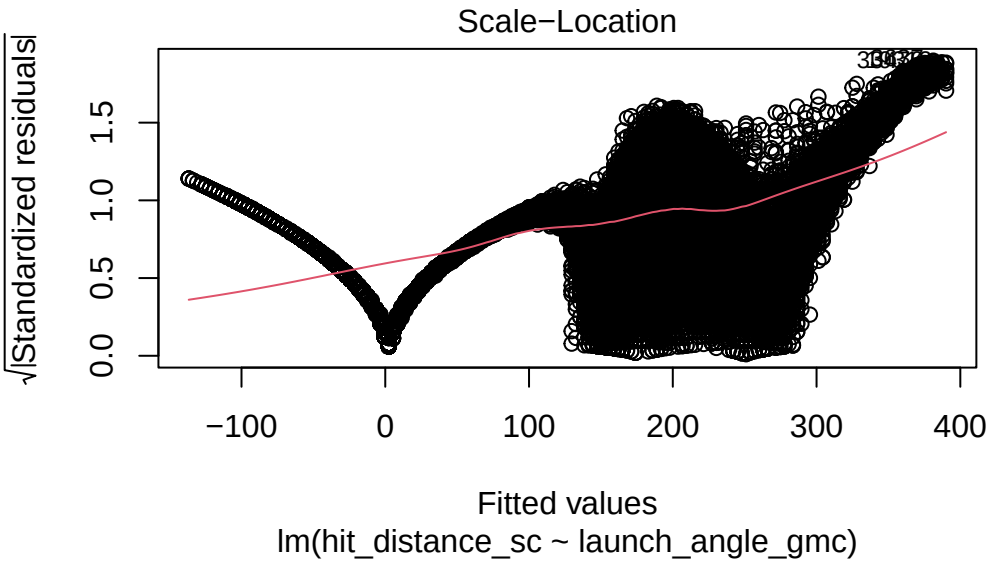
```r
modellag |> ggplot(aes(x=launch_angle_gmc, y=hit_distance_sc)) +
  geom_point()
```

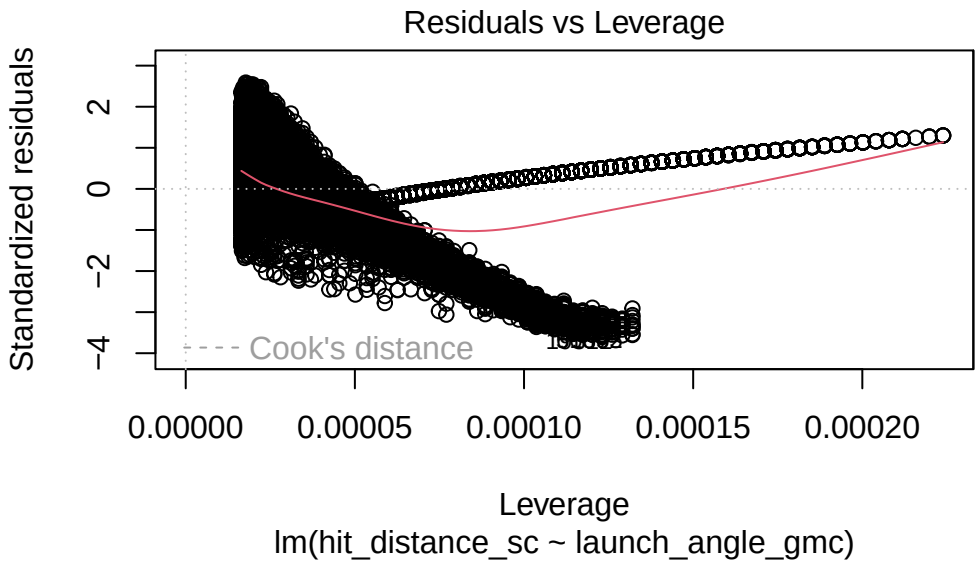



```
plot(modellag)
```




Q−Q Residuals
Standardized residuals
3
1
−1
−3
−4
−2
0
2
4
Theoretical Quantiles
lm(hit_distance_sc ~ launch_angle_gmc)

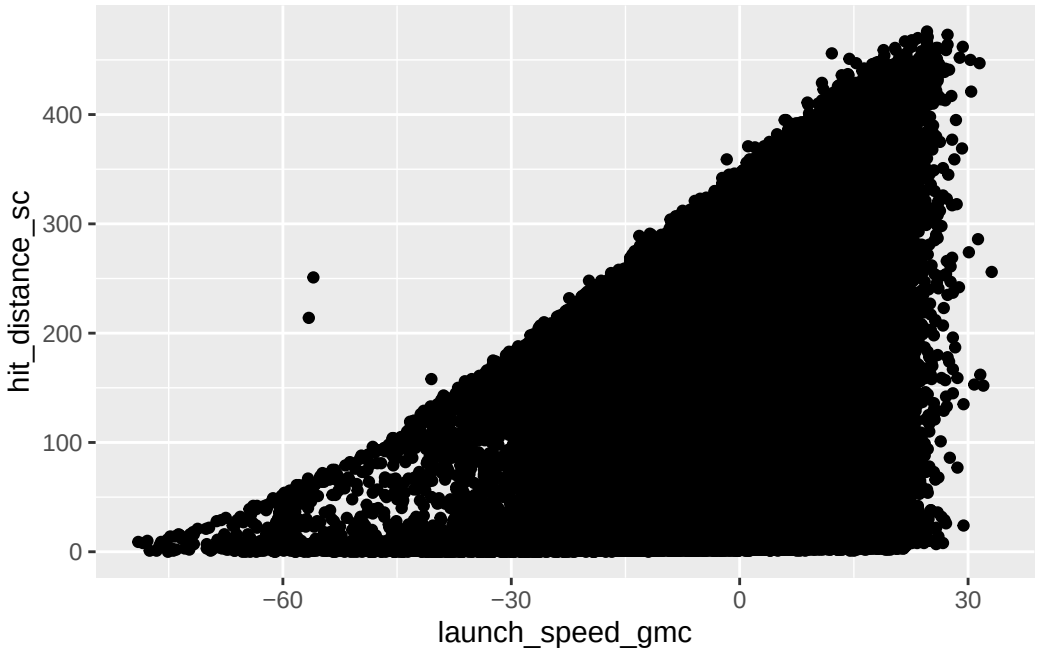

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−100
0
100
200
300
400
Fitted values
lm(hit_distance_sc ~ launch_angle_gmc)



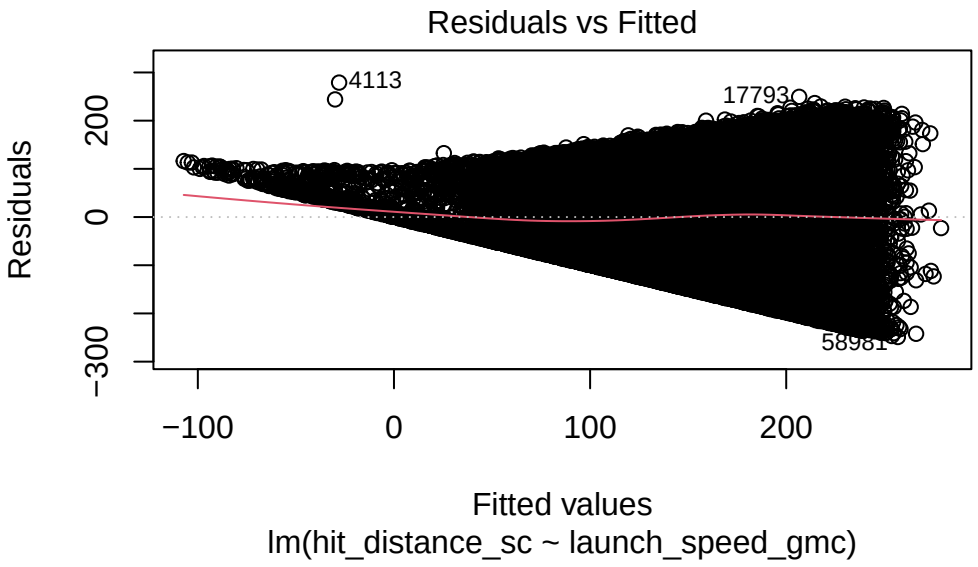

```r
modellsg <- lm(hit_distance_sc ~ launch_speed_gmc, data = df_baseball_clean)
summary(modellsg)
```

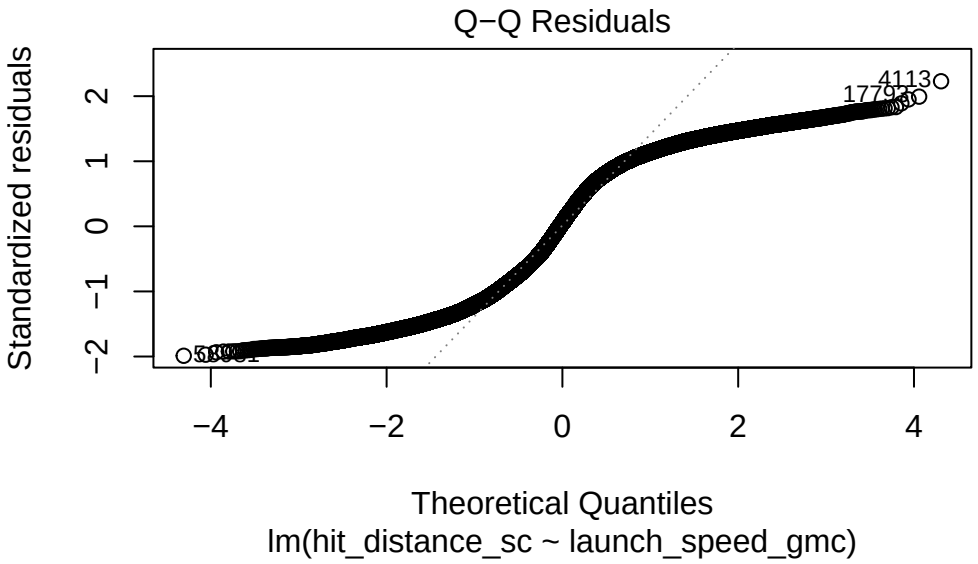
```
Call:
lm(formula = hit_distance_sc ~ launch_speed_gmc, data = df_baseball_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-248.904 -116.205    7.088  119.691  278.965

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      164.91370    0.50752   324.9   <2e-16 ***
launch_speed_gmc   3.44461    0.03419   100.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 125.2 on 60809 degrees of freedom
Multiple R-squared:  0.143, Adjusted R-squared:  0.143
F-statistic: 1.015e+04 on 1 and 60809 DF,  p-value: < 2.2e-16
```

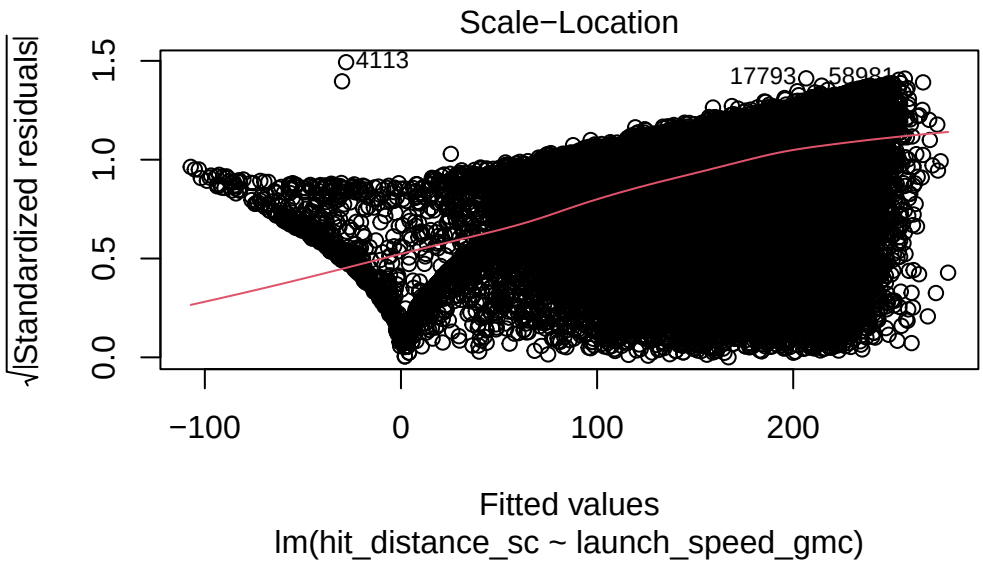
```
modellsg |> ggplot(aes(x=launch_speed_gmc, y=hit_distance_sc)) +
  geom_point()
```

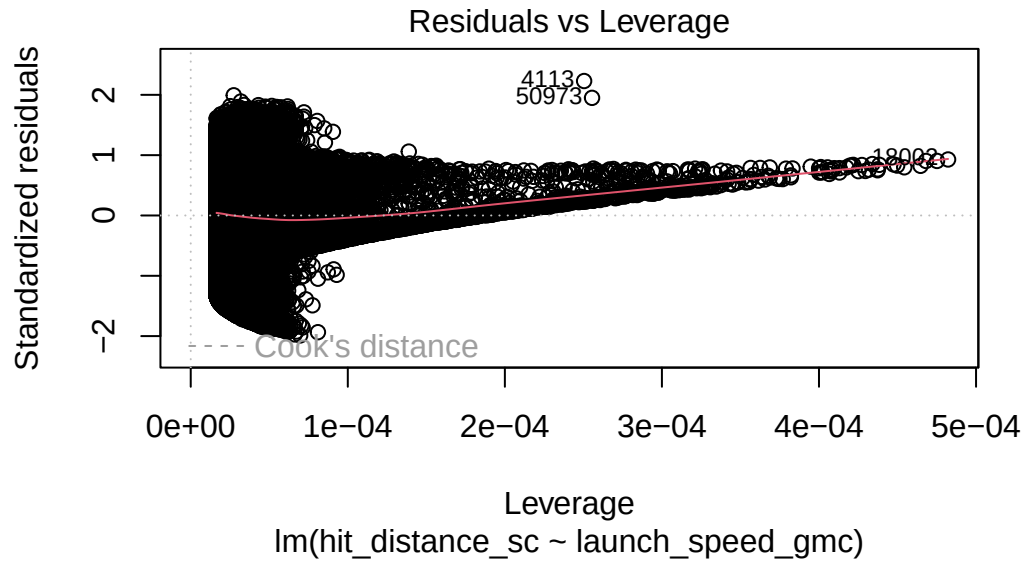



```
plot(modellsg)
```


Residuals vs Fitted
Residuals
Fitted values
lm(hit_distance_sc ~ launch_speed_gmc)
4113
17793
58981


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(hit_distance_sc ~ launch_speed_gmc)
4113
17793

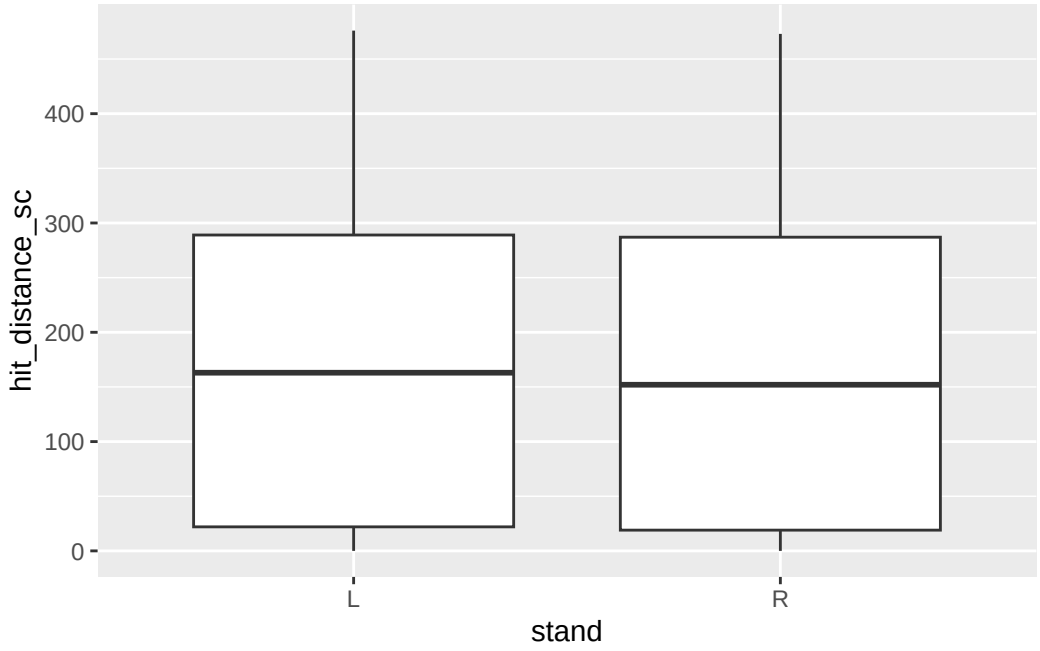

Scale−Location
√|Standardized residuals|
4113
17793
58981
0.0
0.5
1.0
1.5
−100
0
100
200
Fitted values
lm(hit_distance_sc ~ launch_speed_gmc)


Residuals vs Leverage
Standardized residuals
4113
50973
18003
Cook's distance
−2
0
1
2
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
Leverage
lm(hit_distance_sc ~ launch_speed_gmc)

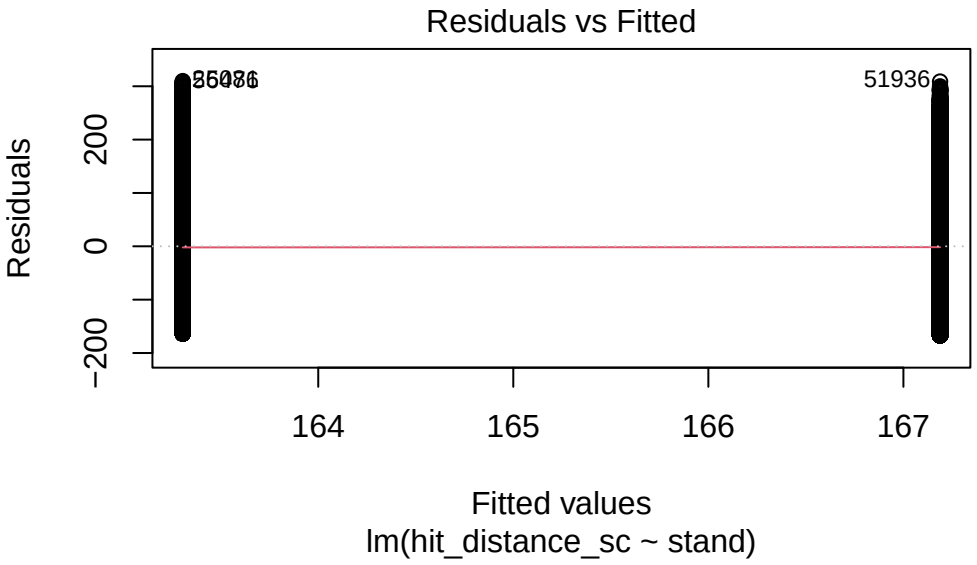
```r
modelside <- lm(hit_distance_sc ~ stand, data=df_baseball_clean)
summary(modelside)
```

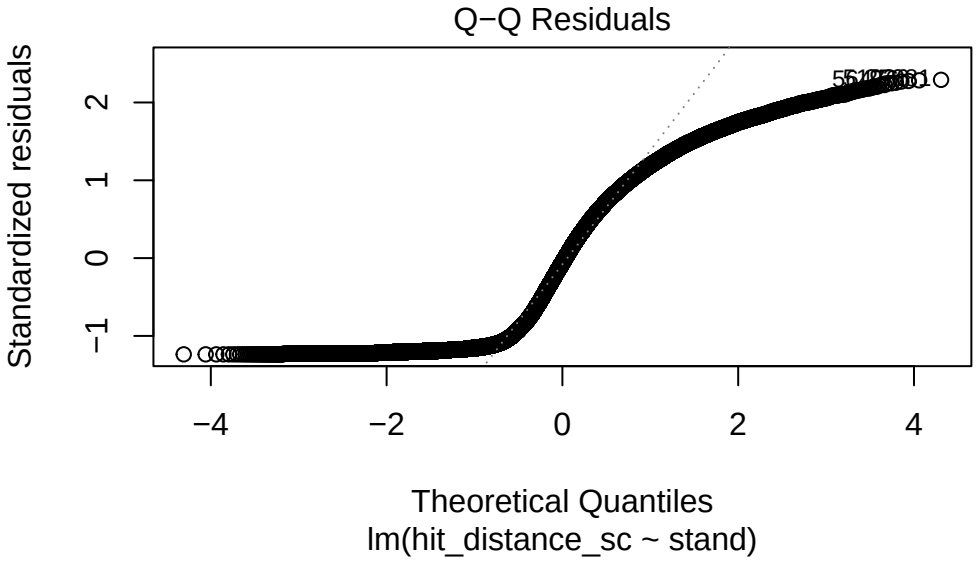
```
Call:
lm(formula = hit_distance_sc ~ stand, data = df_baseball_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-167.188 -144.305   -8.188  122.695  309.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 167.1880     0.8517 196.300  < 2e-16 ***
standR       -3.8828     1.1128  -3.489 0.000485 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 135.2 on 60809 degrees of freedom
Multiple R-squared:  0.0002002, Adjusted R-squared:  0.0001837
F-statistic: 12.17 on 1 and 60809 DF,  p-value: 0.000485
```

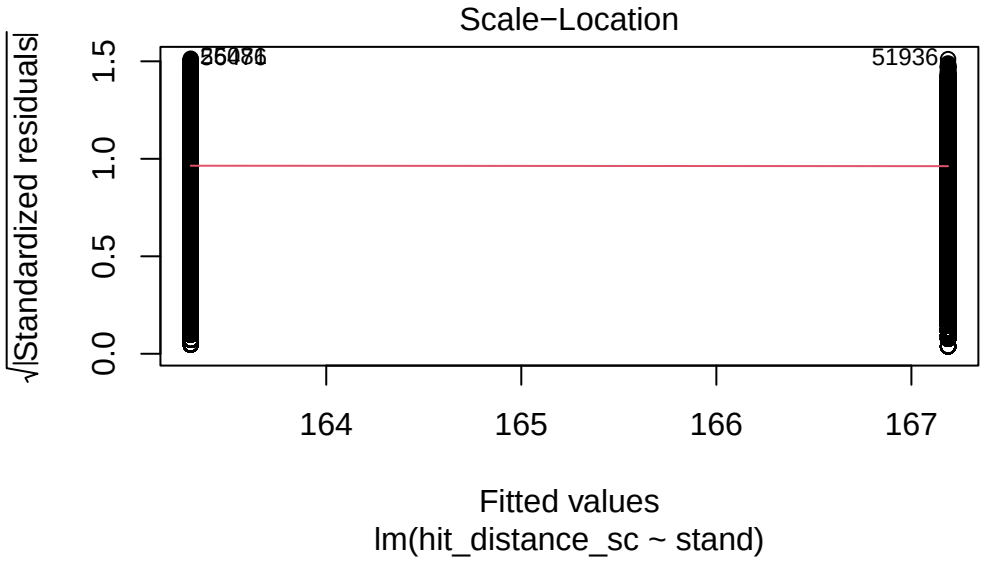
```r
modelside |> ggplot(aes(x=stand, y=hit_distance_sc)) +
  geom_boxplot()
```

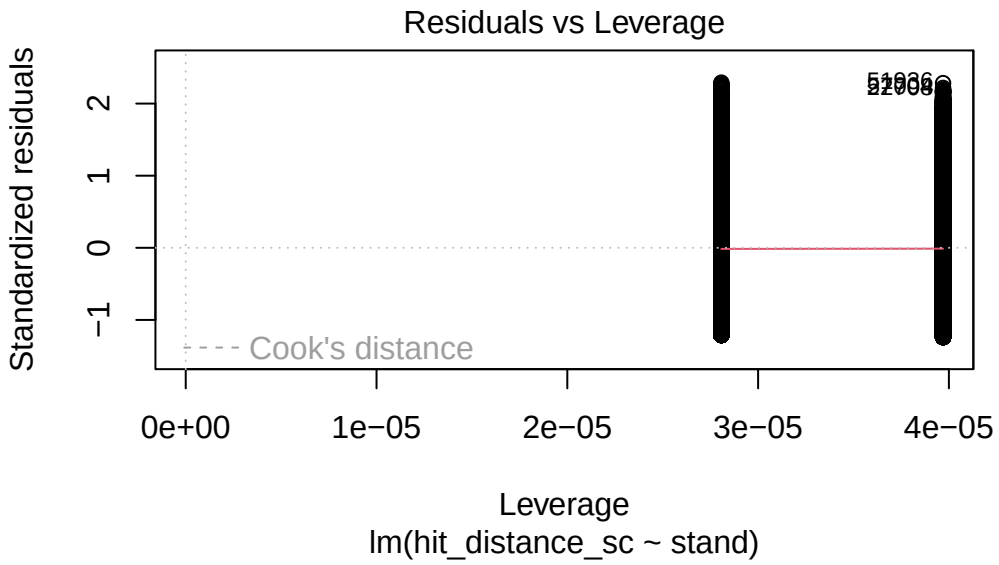



```
plot(modelside)
```




Q−Q Residuals
Standardized residuals
−1
0
1
2
−4
−2
0
2
4
Theoretical Quantiles
lm(hit_distance_sc ~ stand)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
51936
164
165
166
167
Fitted values
lm(hit_distance_sc ~ stand)




```
## Finding the number of Batters and Hits
n_distinct(df_baseball_clean$batter)
```

```
[1] 554
```

```
## Total Number of hits per player
df_baseball_clean |>
  group_by(batter) |>
  summarize(total_hits = sum(hit_distance_sc)) |>
  summarize(average = mean(total_hits))
```

```
# A tibble: 1 x 1
  average
    <dbl>
1  18102.
```

```
## Total Number of Data Points
df_baseball_clean |>
  summarize(total_hits = sum(hit_distance_sc))
```

```
# A tibble: 1 x 1
  total_hits
       <dbl>
1   10028567
```