IMT 573: Module 7 Lab

Regression

Ali Qazi

Due: August 06, 2021

Collaborators: Akeel Qazi, Anthony Mercado List collaborators here.

Objectives

Regression is one of the fundamental and important data models we will encounter in this course. Our aim in this exercise is to see regression models in practice. Recall, our emphasis in class and here will also be on interpreting the results of these models in context, and always to evaluate the assumptions of the model to ensure valid and accurate statistical inference.

Instructions

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Open the `07_lab_regression.Rmd` and save a copy to your local directory. Supply your solutions to the assignment by editing `07_lab_regression.Rmd`.
2. First, replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and I encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`. When the PDF report is generated rename the knitted PDF file to `lab7_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

In this lab you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Sports Statistics: Predicting Runs Scored in Baseball

Baseball is a played between two teams who take turns batting and fielding. A run is scored when a player advances around the bases and returns to home plate. The data we will use today is from all 30 Major League Baseball teams from the 2011 season. This data set is useful for examining the relationships between wins, runs scored in a season, and a number of other player statistics.

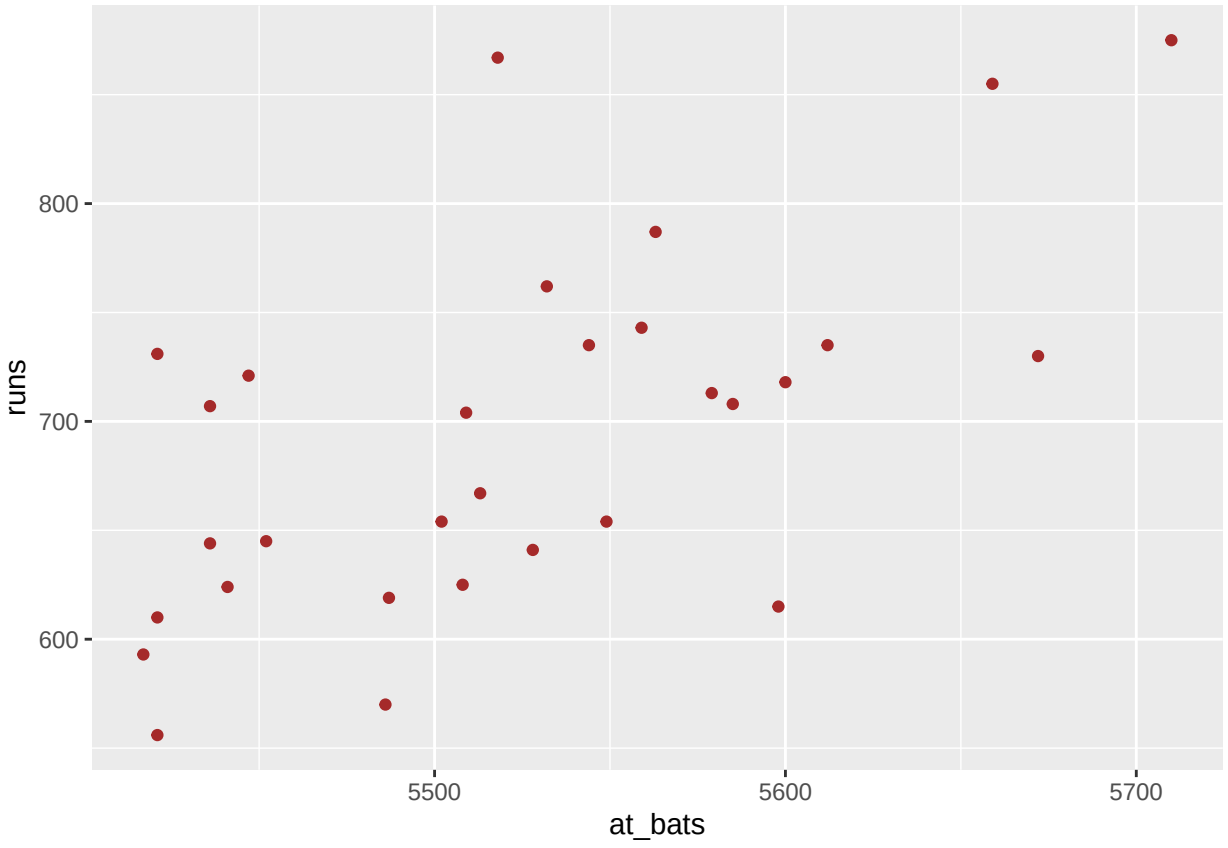
Note: More info on the data can be found here: https://www.openintro.org/stat/data/mlb11.php

```
# Download and load data
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "data/mlb11.RData")
load("data/mlb11.RData")
```

Use the baseball data to answer the following questions:

- Plot the relationship between runs and at bats. Does the relationship look linear? Describe the relationship between these two variables.

```
ggplot(mlb11, aes(x = at_bats, y = runs)) +
geom_point(colour = "brown")
```



- If you knew a team's at bats, would you be comfortable using a linear model to predict the number of runs?

I am comfortable using a linear model

- If the relationship looks linear, quantify the strength of the relationship with the correlation coefficient. Discuss what you find.

```r
# quantify strength of relationship with the correlation coefficient
cor(mlb11$runs, mlb11$at_bats)
```

```
## [1] 0.610627
```

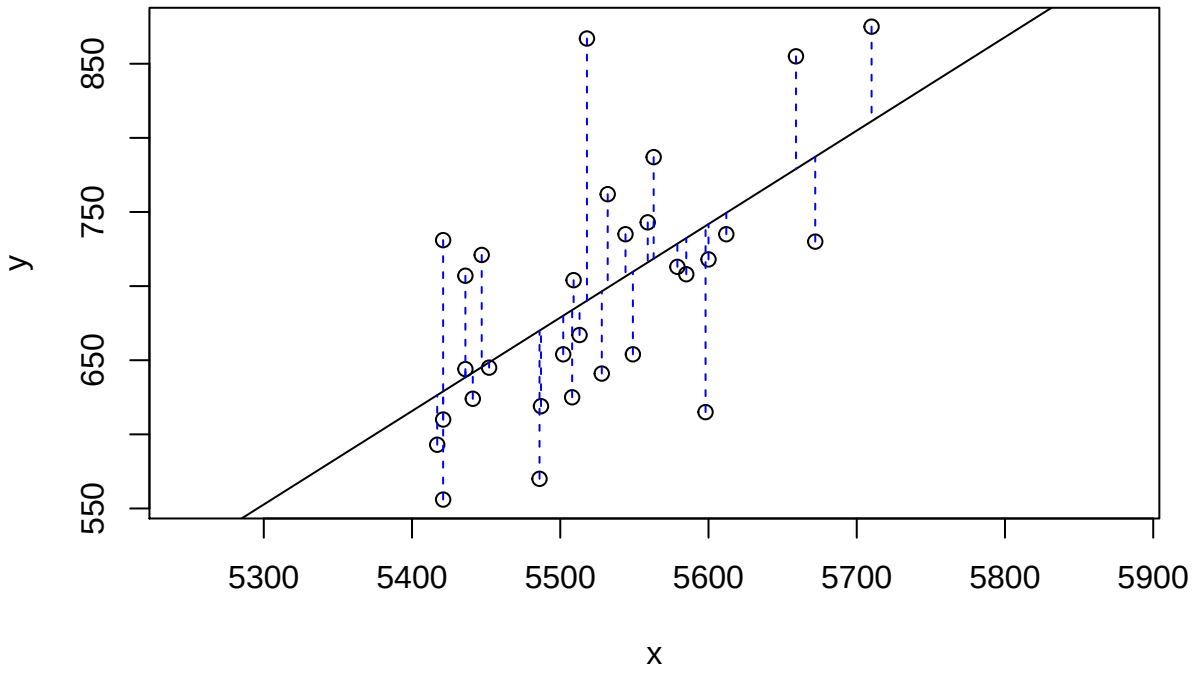
When the r value of two variables is more than 0.7, that link is typically regarded as being strong. The r value mentioned above is 0.6, indicating that it is not particularly powerful. In order to forecast the number of runs, I would not feel particularly confident using a linear model.

- Use the `lm()` function to fit a simple linear model for runs as a function of at bats. Write down the formula for the model, filling in estimated coefficient values.

```r
m1 <- lm(runs ~ at_bats, data = mlb11)
summary(m1)
```

```
##
## Call:
## lm(formula = runs ~ at_bats, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.58  -47.05  -16.59   54.40  176.87
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2789.2429   853.6957  -3.267 0.002871 **
## at_bats         0.6305     0.1545   4.080 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.47 on 28 degrees of freedom
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

```r
plot_ss(x = mlb11$at_bats, y = mlb11$runs)
```



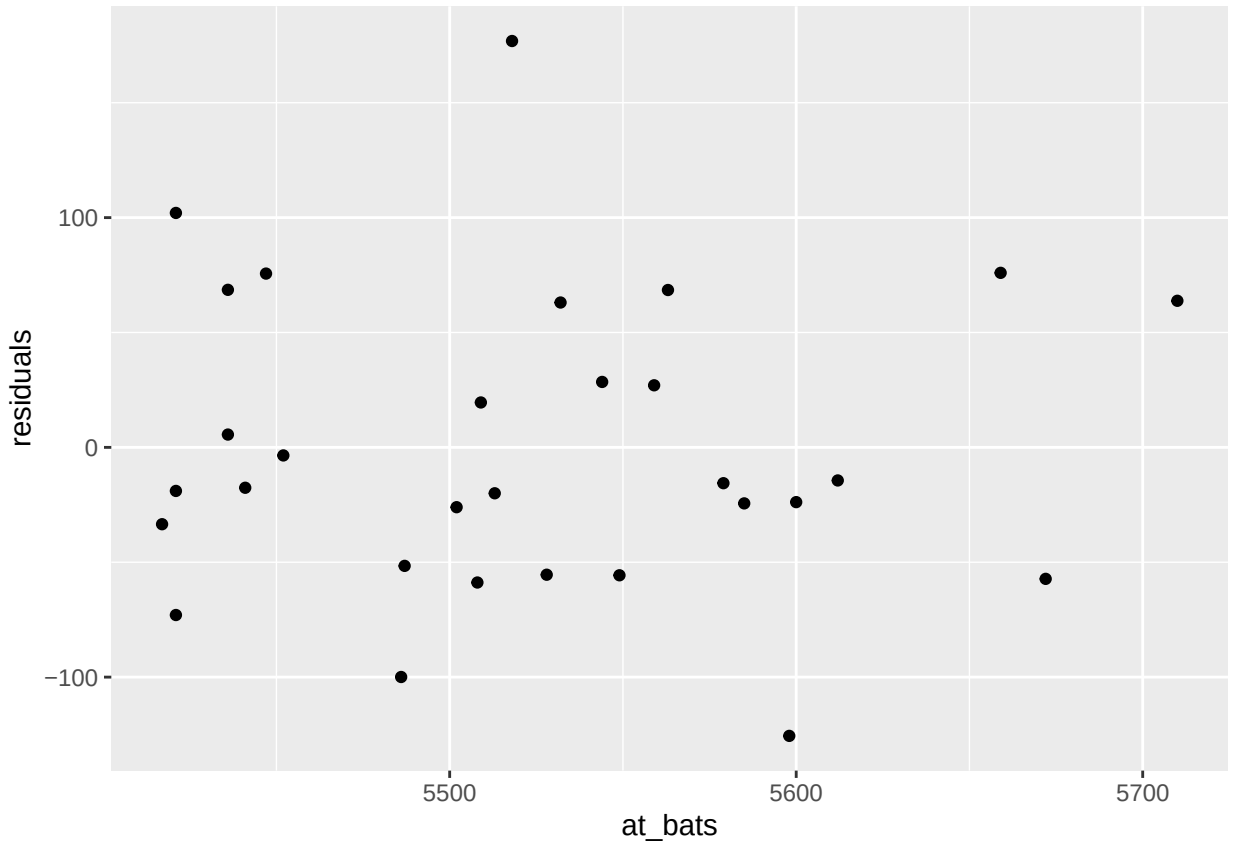

```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##  -2789.2429       0.6305
##
## Sum of Squares:  123721.9
```

- Describe in words the interpretation of β_1.

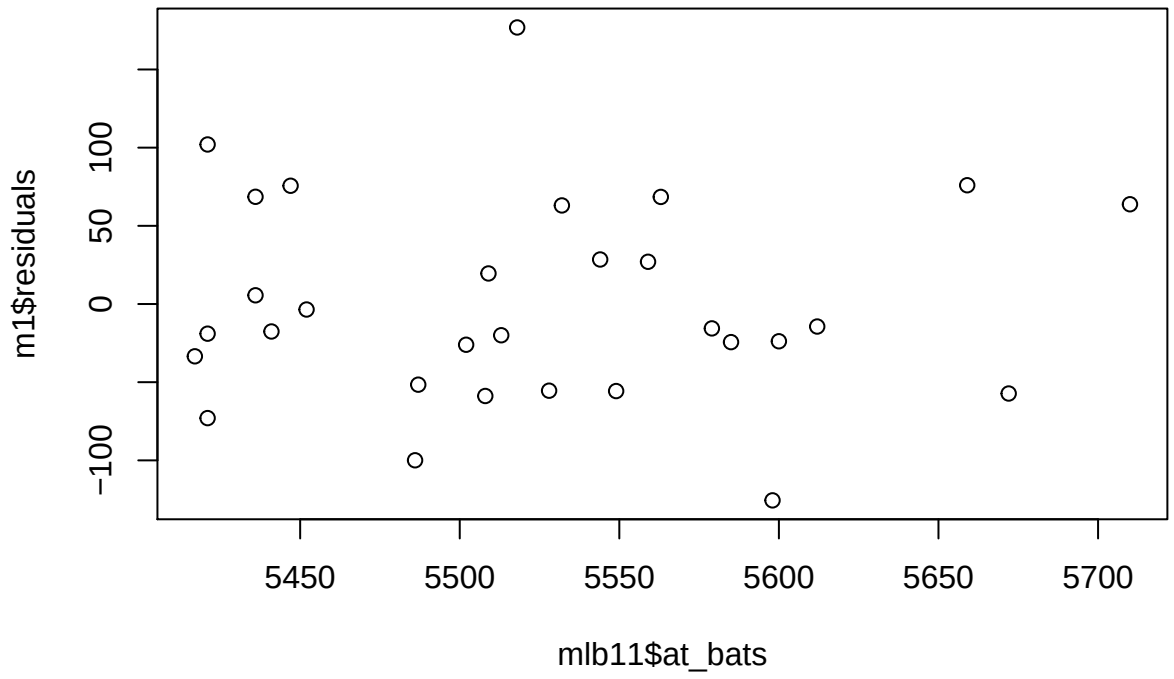
There is a projected gain in runs of 0.6 for each extra at-bat (at-bat). That represents the increase in runs scored for each extra at-bat.

- Make a plot of the residuals versus at bats. Is there any apparent pattern in the residuals plot?

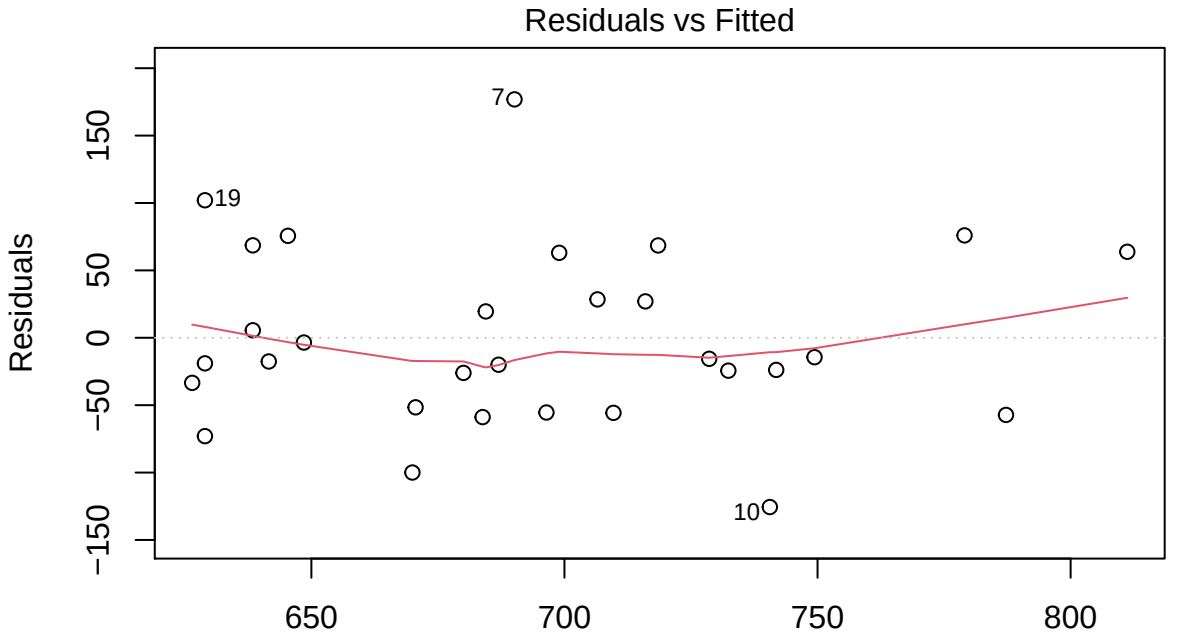
```
mlb11_model <- mlb11 %>%
mutate(residuals = m1$residuals)
ggplot(mlb11_model, aes(x = at_bats, y = residuals)) +
geom_point()
```

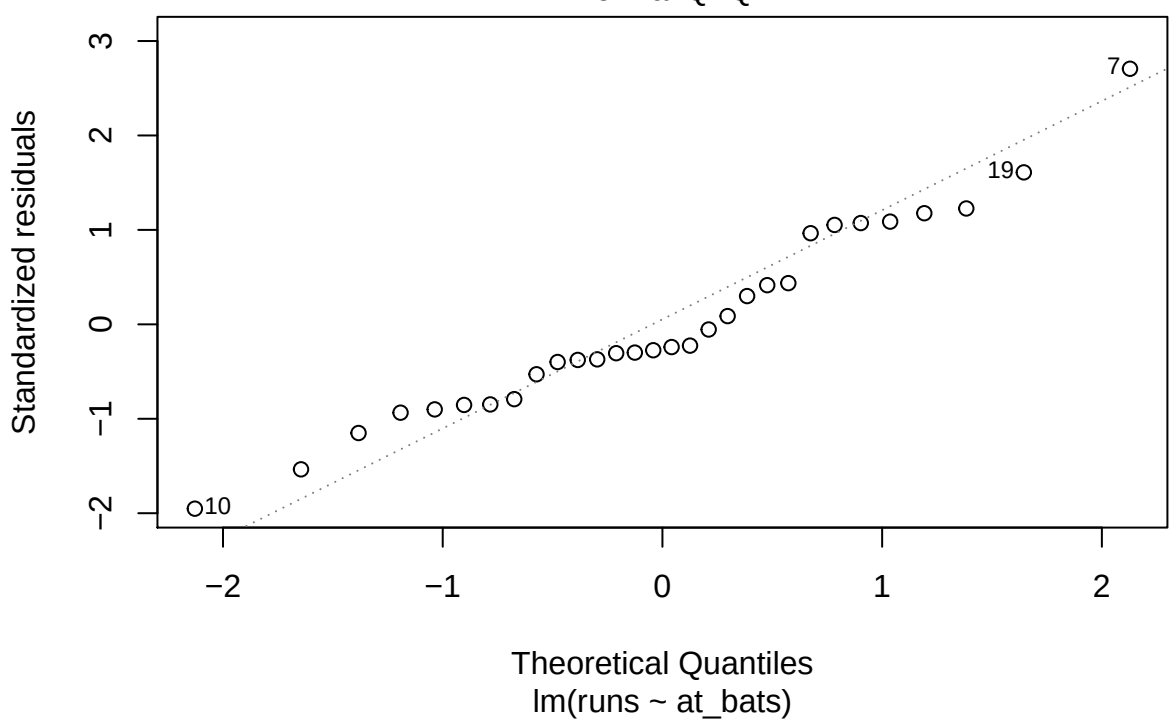


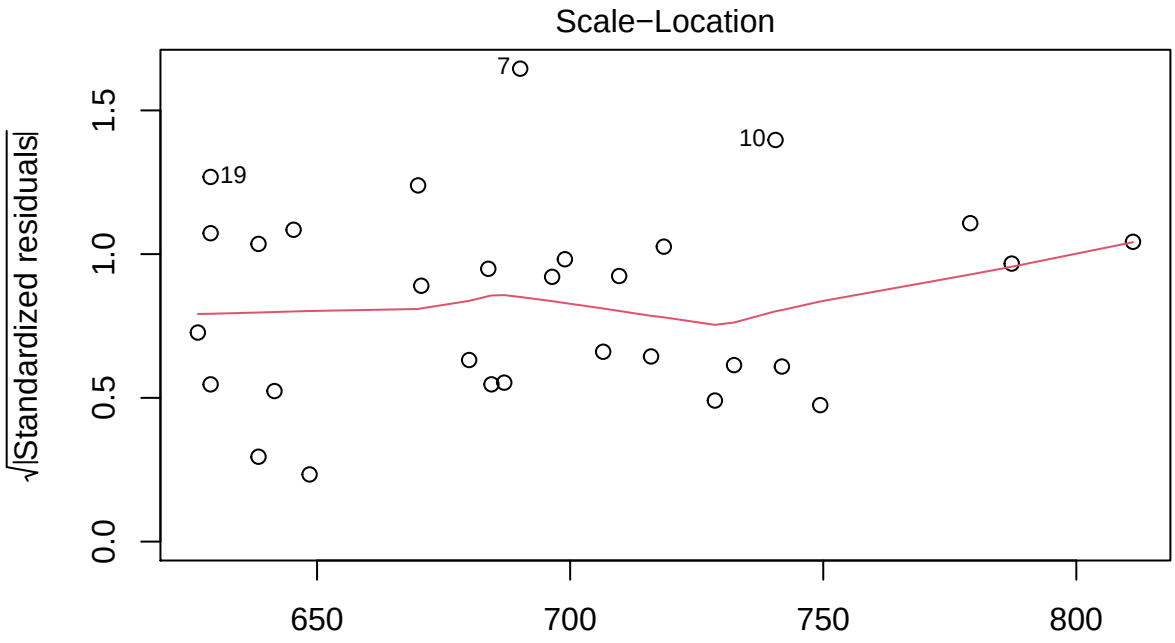

```
plot(m1$residuals ~ mlb11$at_bats)
```

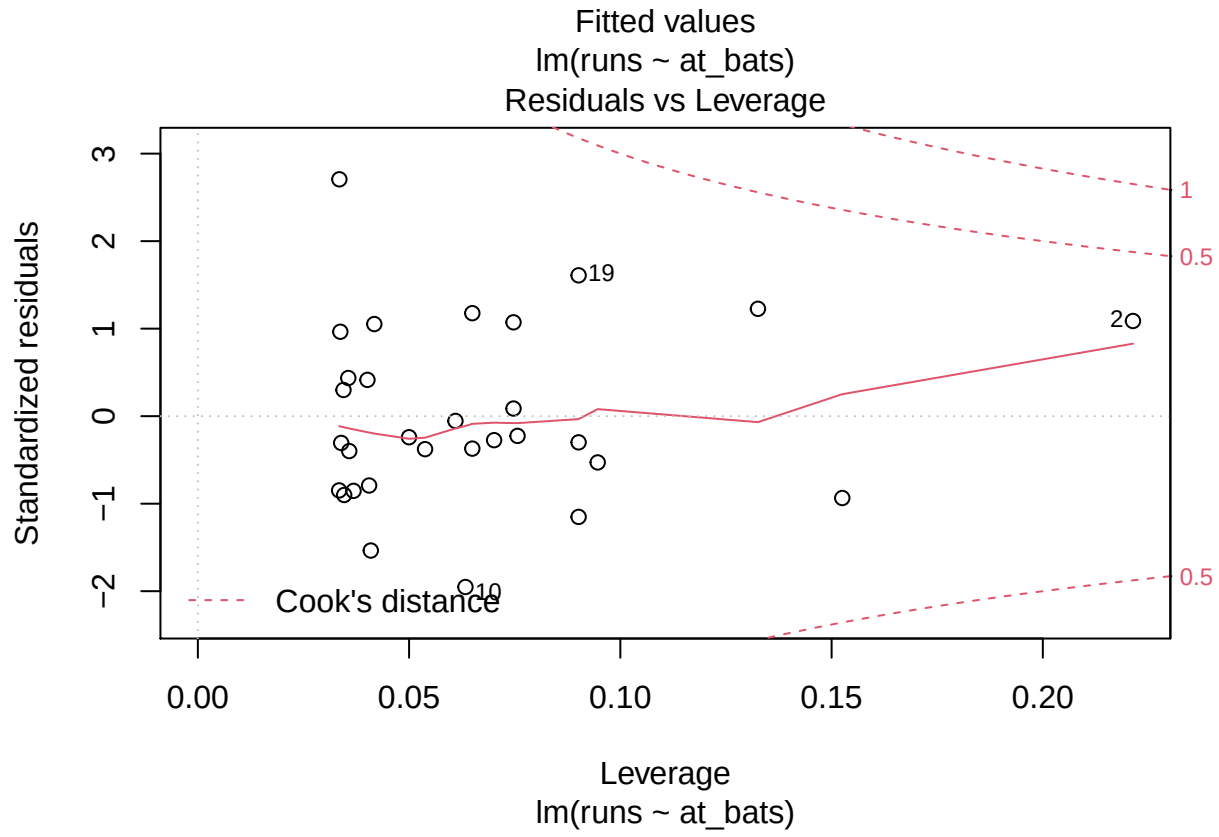


```
plot(m1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(runs ~ at_bats)
7
19
10
150
50
0
−50
−150
650
700
750
800


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(runs ~ at_bats)
7
19
10
3
2
1
0
−1
−2
−2
−1
0
1
2





I analyzed residual graphs and look for trends. There are no obvious patterns, such as funnels or linear trends, in this graph. Random noise is uniformly distributed around 0.

- Comment of the fit of the model.

I gauged how well the model fits by using the model diagnostic. It demonstrates that the residuals have constant variance and zero. No pattern is apparent to me. The r square adjustment was 0.35. Not bad, but neither fantastic.