UNITED STATES MILITARY ACADEMY

PROJECT DATA ANALYSIS

MA376: APPLIED STATISTICS

SECTION G25

DR. ANNY-CLAUDE JOSEPH

By

CADET AUDREY HAMILTON '21, CO A1

CADET JOSEPH ZUCCARELLI '21, CO F1

WEST POINT, NEW YORK

05 NOVEMBER 2020

JZ OUR DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

_____ WE DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: Joseph Zuccarelli

Are Sluggers More Likely to Strikeout?

Audrey Hamilton and Joseph Zuccarelli

05 November, 2020

Introduction

In 2019, New York Mets first baseman Pete Alonso led Major League Baseball with 53 home runs. However, Alonso was also atop a less desirable league leaderboard–strikeouts. Alonso struck out a total of 183 times during the 2019 season, just shy of the league leader Eugenio Suarez with 189 strikeouts on the season. For sluggers such as Alonso, is it safe to assume that they are swinging for the fences during most at-bats, and does this mean that they are more likely than other players to strike out? The objective of the following study is to determine if there is a relationship between rate of strikeouts and home runs for Major League Baseball (MLB) players.

Methods

The observational units in this study are MLB players who had over 100 at-bats during the 2019 season. There are 390 complete observational units in the data set. Each unit represents a player's total regular season numbers. Note that the data set was created using the "Sean Lahman Baseball Database" in R (see Appendix A).[1] The variables included in the study are as follows: *bats*, *strike out rate*, *home run rate*, and *batting average*. *Bats* is a categorical variable that represents the player's batting hand. This variable has two levels: L for those who bat left-handed and R for those who bat right-handed. Note that any switch hitters, or players that bat both left-handed and right-handed, were not included in the study. In this study, *bats* is a potential confounding variable. *Strike out rate* is a quantitative variable that is calculated by dividing the number of total strike outs by the total number of at-bats. In this study, *strike out rate* is the response variable. *Home run rate* is a quantitative variable that is calculated by dividing the number of total home runs by the total number of at-bats. In this study, *home run rate* is the explanatory variable. *Batting average* is a quantitative variable that is calculated by dividing the number of total hits by the total number of at-bats. In this study, *batting average* is a potential confounding variable.

The statistical analysis method that will be employed in this study is a linear regression involving both quantitative and categorical variables. The null hypothesis for this analysis is that there is no linear relationship between *strike out rate* and *home run rate*. The alternate hypothesis for this analysis is that there is a linear relationship between *strike out rate* and *home run rate*. There are two potential confounding variables, *bats* and *batting average*. Note that we will not test for an interaction between these variables. Also note that our analysis will include four validity conditions. These conditions will be considered met if: the residuals vs. predicted values graph does not show any strong evidence of curvature or other patterns (linearity), the responses can be considered independent of each other (independence), the histogram of the residuals is approximately symmetric with no large outliers (normality), and the residuals vs. predicted values graph shows a constant width (equal variance). We can confirm that the condition of independence has already been met, as the responses included in the data set are all different players from the 2019 season. The other three validity conditions will be assessed later on in the report.

Results

Figure 1, included below, summarizes the mean, standard deviation, and median for the variables *strike out rate*, *home run rate* and *batting average* based on the player's batting hand. Note that all of the values are rates, as each value is a decimal less than one. Also that there are three rows in the table: L (left-handed batters), R (right-handed batters), and overall (both left-handed and right-handed batters). Left-handed batters represent about 37.18% percent of the data set, while right-handed batters represent about 62.18%. Based on the table, left-handed and right-handed batters share very similar mean, standard deviation, and median values for each variable. This appears to suggest that neither of these three variables are affected by the player's batting hand.

	L (N=145)	R (N=245)	Overall (N=390)
strikeoutRate			
Mean (SD)	0.256 (0.0739)	0.259 (0.0721)	0.258 (0.0727)
Median [Min, Max]	0.251 [0.0890, 0.473]	0.260 [0.0421, 0.476]	0.257 [0.0421, 0.476]
homerunRate			
Mean (SD)	0.0409 (0.0209)	0.0415 (0.0204)	0.0413 (0.0206)
Median [Min, Max]	0.0389 [0, 0.102]	0.0386 [0, 0.0997]	0.0388 [0, 0.102]
battingAverage			
Mean (SD)	0.250 (0.0364)	0.253 (0.0361)	0.252 (0.0362)
Median [Min, Max]	0.251 [0.151, 0.334]	0.255 [0.125, 0.344]	0.254 [0.125, 0.344]

Figure 1: MLB Batting Statistics (2019) by Batting Hand

Figure 2, provided below, displays the distribution of strike out rates for players during the 2019 season. The code used to produce this plot can be found in Appendix A. The plot of the distribution of strike out rates has one main peak (unimodal) at around 0.250 and is relatively symmetric. Strike out rates range between 0.0421 and 0.4760 with a mean of 0.258 and a median of 0.257.

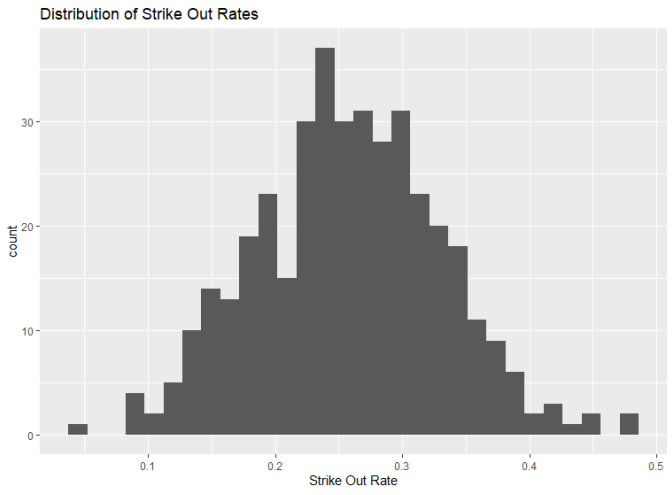

Figure 2: Distribution of Strike Out Rates

Figure 3, included on the next page, displays the relationship between strike outs and home runs for players from the 2019 season. The code used to produce this plot can be found in Appendix A. Note that the plot is colored by batting hand, as left-handed hitters are displayed in red and right-handed players are displayed in blue. Based on the plot, there appears to be a positive linear relationship between strike out rate and home run rate. The general trend displayed by the plot is that as a player's strike out rate increases,

their home run rate increases as well. The relationship appears to be the same for both left-handed and right-handed batters.

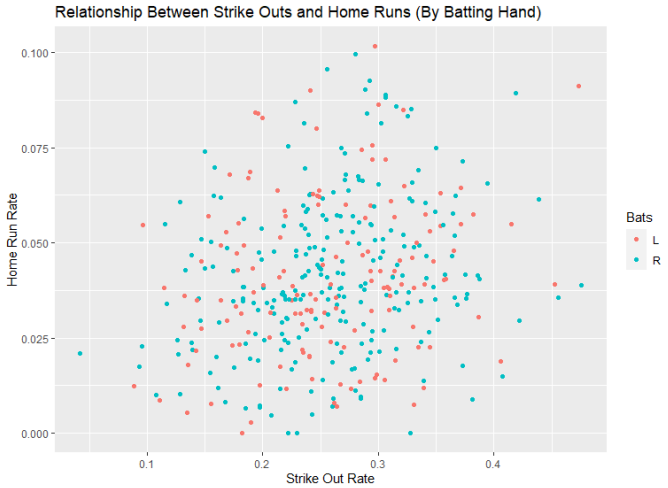

Figure 3: Relationship Between Strike Outs and Home Runs (By Batting Hand)

Figure 4, provided below, displays the relationship between strike outs and hits for players from the 2019 season. The code used to produce this plot can be found in Appendix A. Note that the plot is also colored by batting hand, as left-handed hitters are displayed in red and right-handed players are displayed in blue. Based on the plot, there appears to be a negative linear relationship between batting average and home run rate. This linear relationship appears to be a bit stronger than the linear relationship between strike out rate and home run rate indicated in **Figure 3**. The general trend displayed by the plot is that as a player's strike out rate increases, their batting average decreases. Note that batting average is simply a ratio of hits to at-bats, and thus it appears that the more a player strikes out, the less hits they accumulate. The relationship again appears to be the same for both left-handed and right-handed batters.

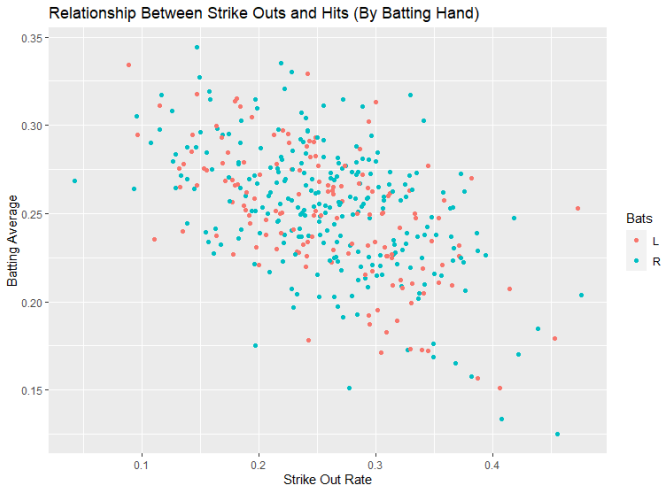

Figure 4: Relationship Between Strike Outs and Hits (By Batting Hand)

Author Contributions

1. CDT Zuccarelli completed all the R code. He wrote the description about Figures 2, 3, and 4. Additionally, CDT Zuccarelli edited and made changes to the sections that CDT Hamilton wrote.

2. CDT Hamilton wrote the initial paragraph outlining the statistical analysis and validity conditions, and described Figure 1.

Works Cited

[1] Lahman, Sean. *Batting*. Lahman's Baseball Database, http://www.seanlahman.com/baseball-archive/statistics/. Accessed 25 September 2020.

Appendix A

The following R Code was used to obtain the data set and produce all figures included in this report.

```
##Libraries Required to Run Code
library(Lahman)
library(tidyverse)
library(table1)
library(knitr)


##Merge Lahman Data Frames to Obtain Player Name and Batting Hand
Master$name <- paste(Master$nameFirst, Master$nameLast, sep=" ")
batting <- merge(Batting,
                 Master[,c("playerID","name","bats")],
                 by="playerID", all.x=TRUE)


##Filter Batting Data Based on Inclusion Criteria
battingData <- batting %>%
  filter(yearID == 2019) %>%
  filter(AB >= 100) %>%
  filter(bats == "R" | bats == "L") %>%
  mutate(strikeoutRate = SO/AB,
         homerunRate = HR/AB,
         battingAverage = H/AB) %>%
  select(name,bats,strikeoutRate,homerunRate,battingAverage)


##Table 1 Output
table1(~strikeoutRate + homerunRate + battingAverage | as.factor(bats), data = battingData)


##Plot Distribution of Strike Outs
battingData %>%
  ggplot(aes(x= strikeoutRate))+
  geom_histogram()+
  labs(x = "Strike Out Rate") +
  ggtitle("Distribution of Strike Outs")

##Plot Relationship Between Strike Outs and Home Runs (By Batting Hand)
battingData %>%
  ggplot(aes(x= strikeoutRate, y= homerunRate, color= bats))+
  geom_point()+
  labs(x = "Strike Out Rate", y = "Home Run Rate", color = "Bats") +
  ggtitle("Relationship Between Strike Outs and Home Runs (By Batting Hand)")


##Plot Relationship Between Strike outs and Hits (By Batting Hand)
battingData %>%
  ggplot(aes(x= strikeoutRate, y= battingAverage, color= bats))+
  geom_point()+
  labs(x = "Strike Out Rate", y = "Batting Average", color = "Bats") +
```

```
ggtitle("Relationship Between Strike Outs and Hits (By Batting Hand)")
```