Koby Manning

Prof. Sanders

STA 4102

4/30/2023

STA 4102 Project Writeup

Growing up in the Tampa Bay area, I have been surrounded by professional sports teams, be it the Bucs, Lightning, or Rays, a proximity which has brought great affinity to the latter and the sport of baseball as a whole. From being a player in my younger years, to becoming an avid spectator through my teens to the present, my love for the game has only grown with time. This love for the game has since been paired alongside my affinity for statistical analysis, and it is this pairing which I aim to take into a future career in MLB front office analytics. Since focusing on this potential career path, I have taken many endeavors into statistical analysis during my education, ranging from analysis of Cy Young seasons to the impact of pitching on a team's success. For this project, my aim is to shift into the other side of the game – offense. The aim of this project is to answer the question "to what extent do team batting statistics impact the winning percentage of an MLB team?"

In order to answer my research question, I found it imperative to find a substantial database to supply my investigation and decided to use stathead.com. Stathead provides a wealth of knowledge from the history of professional baseball, not just modern MLB teams, although the latter group is my focus at the moment. The next pressing matter is choosing a time period, which I decided to focus on the 3 most recent full seasons of MLB play: 2019, 2021 and 2022. As a result of the COVID-19 pandemic, the 2020 season was a short sixty games, rather than the full 162, potentially leading to rate stats that were skewed, and counting stats that cannot realistically be extrapolated to a full season, reasons which explain its exclusion from my dataset. On the Stathead website, I ran three queries, one for each year, with the criteria as follows. For each year, I exported the data to an excel file and created a master data file including the three-year data for each team, which would eventually be used to perform my analysis.

Search Criteria

Click on the red text to pre-fill the form with various values

Sort By

Descending

Wins

Seasons

2022 to 2022

Any • 2023 • 2022 • 2021 • Wild-Card Era • Divisional Era • Expansion Era • Integration Era • Live-Ball Era • Modern Era

Game Type

- (•) Regular Season
- ( ) Postseason

Team

Any Team

League

- [x] American League (1901-present)
- [x] National League (1876-present)
- [x] Negro American League (1937-1948)
- [x] Negro National League I (1920-1931)
- [x] Negro National League II (1933-1948)
- [x] East-West League (1932)
- [x] Negro Southern League (1932)
- [x] American Negro League (1929)
- [x] Eastern Colored League (1923-1928)
- [x] Federal League (1914-1915)
- [x] American Association (1882-1891)
- [x] Players League (1890)
- [x] Union Association (1884)
- [x] National Association (1871-1875)

All/Any • Active • Inactive • Clear

Team Success

- (•) All Teams
- ( ) World Series Champion
- ( ) League Champion
- ( ) Division Champion
- ( ) Made Playoffs
- ( ) Missed Playoffs

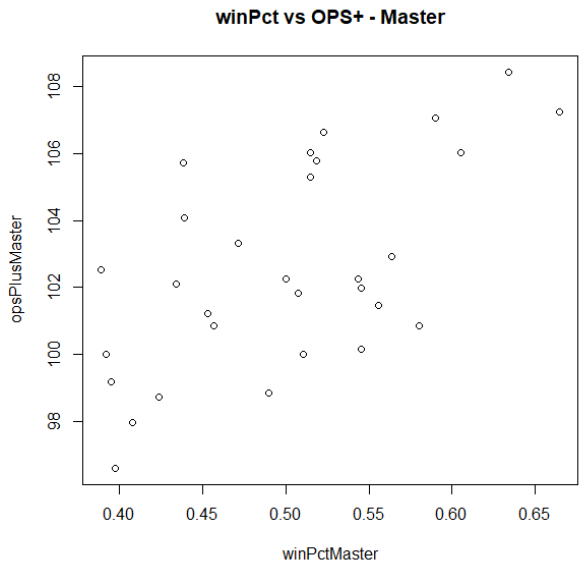
Statistical Filters (AVG, HR, RBI, WAR, etc.)

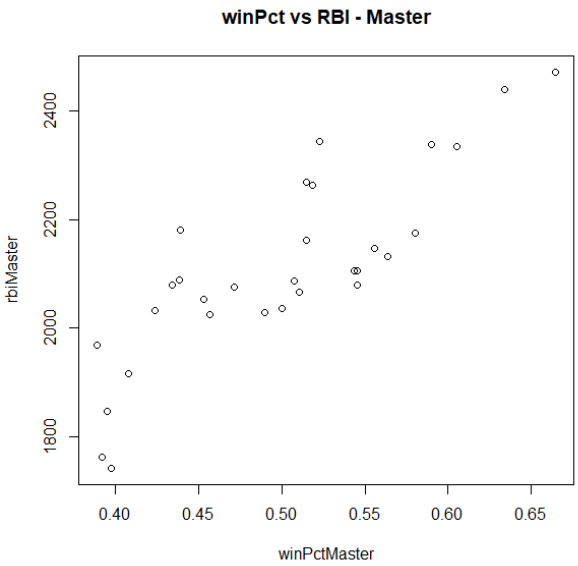
Choose a Statistical Filter

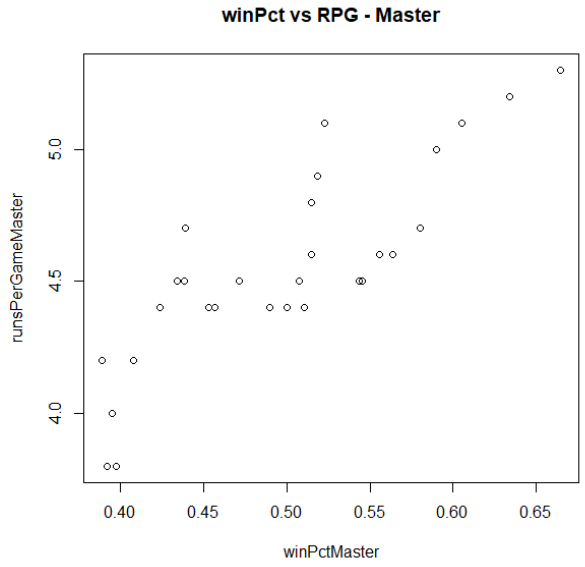
The analysis to answer this research question was conducted in two stages: preliminary and primary analysis. The variables to be analyzed alongside winning percentage include: **Batting Average** (proportion of at-bats during which a batter reached base safely with a hit), **Plate Appearances** (total number of batting opportunities a team had), **Walks** (total number of times a team worked a walk), **OPS+**, (On Base % plus Slugging %, a measure of a team's ability to reach base and hit for power, standardized where 100 is league average), **Runs per Game** (the average number of runs scored per game by a team), and **Runs Batted In** (total number of runs driven in by hits, walks, or sacrifice flies by a team). The preliminary analysis included importing my data into R, creating vectors for each individual statistic year by year, and creating scatterplots to determine which statistics predict winning percentage best.

```
#Import Project Data
batting2019 <- read.csv("C:\\Users\\kobym\\OneDrive\\Documents\\
                        USF Coursework\\STA 4102\\Project&Report\\
                        2019 Batting Data.csv")
batting2021 <- read.csv("C:\\Users\\kobym\\OneDrive\\Documents\\
                        USF Coursework\\STA 4102\\Project&Report\\
                        2021 Batting Data.csv")
batting2022 <- read.csv("C:\\Users\\kobym\\OneDrive\\Documents\\
                        USF Coursework\\STA 4102\\Project&Report\\
                        2022 Batting Data.csv")
battingMaster <- read.csv("C:\\Users\\kobym\\OneDrive\\Documents\\
                          USF Coursework\\STA 4102\\Project&Report\\
                          Master Batting Data.csv")
```

```
#master Vectors
winPctMaster <- battingMaster$WL.
avgMaster <- battingMaster$BA
paMaster <- battingMaster$PA
bbMaster <- battingMaster$BB
opsPlusMaster <- battingMaster$OPS.
runsPerGameMaster <- battingMaster$R.Gm
rbiMaster <- battingMaster$RBI
```







The 3 statistics, OPS+, Runs per Game (RPG), and Runs Batted In (RBI), which predicted winning percentage best were used alongside the master dataset to perform the primary analysis. First, these three statistics from the master dataset were plotted alongside the winning percentage for each team on the same dataset, producing the following scatterplots. Following, I ran linear correlation tests for each statistic, giving the following results.

```
opsPlusModelMaster <- lm(WL. ~ OPS.,
                         data = battingMaster)
summary(opsPlusModelMaster)
```

```
Coefficients:
             Estimate Std. Error
(Intercept) -1.024301   0.372941
OPS.         0.014860   0.003634

t value Pr(>|t|)
 -2.747 0.010408 *
  4.089 0.000331 ***
```

```
rpgModelMaster <- lm(WL. ~ R.Gm,
                     data = battingMaster)
summary(rpgModelMaster)
```

```
Coefficients:
              Estimate Std. Error
(Intercept) -0.2568558  0.0944930
RBI          0.0003584  0.0000446

t value Pr(>|t|)
 -2.718   0.0111 *
  8.036 9.46e-09 ***
```

```
rbiModelMaster <- lm(WL. ~ RBI,
                     data = battingMaster)
summary(rbiModelMaster)
```

```
Coefficients:
             Estimate Std. Error
(Intercept) -0.27159    0.09929
R.Gm         0.16958    0.02176

t value Pr(>|t|)
 -2.735   0.0107 *
  7.795 1.72e-08 ***
```

From the linear models from each pairing of statistics, it is clear that there is a correlation between winning percentage and each of OPS+, RBI, and RPG. Relating back to my research question, which asks "to what extent" these statistics impact winning percentage, it is integral to compare the R^2 value of each pairing to determine the extent to which we can rely on these statistics, respectively, as predictors of a team's ability to win.

```
Adjusted R-squared:  0.3515
```

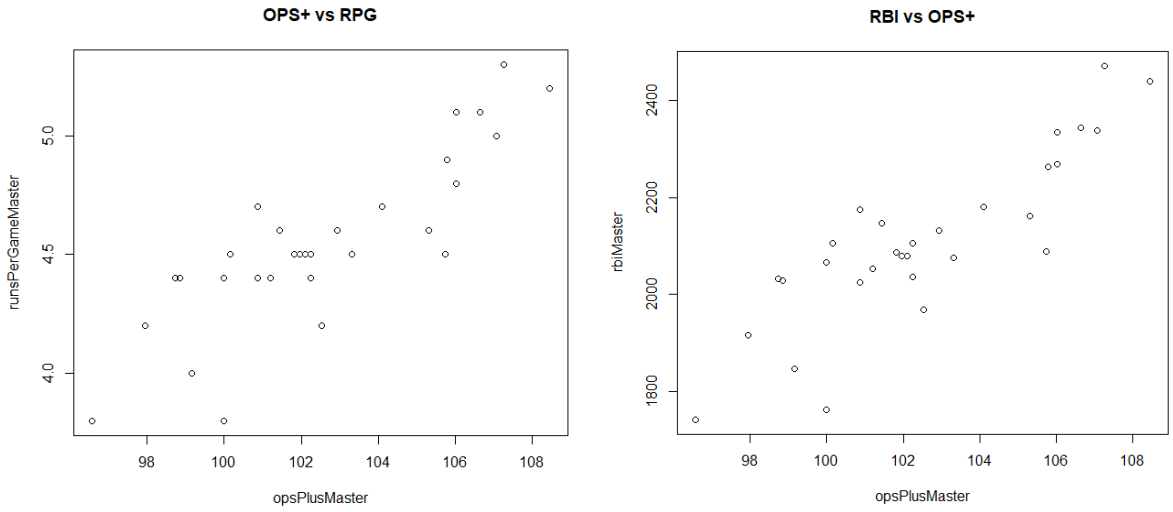
```
Adjusted R-squared:  0.6867
```

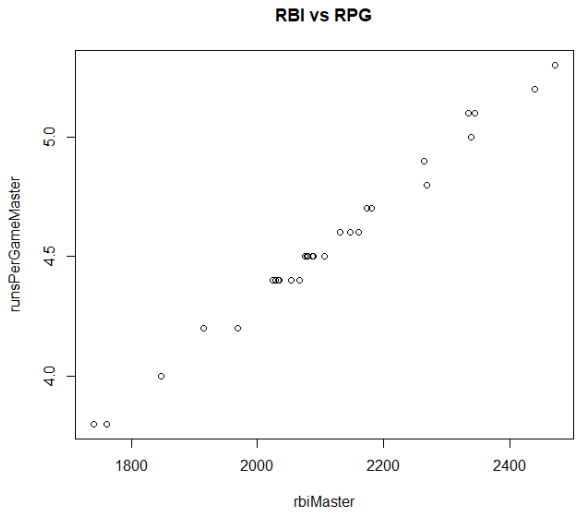
```
Adjusted R-squared:  0.6733
```

As displayed by the R^2values above, the statistic which contributes to the most variation in winning percentage is that of RBI at 68.67%, a strong coefficient of determination. Thus, for one

variable, batting statistics can impact the winning percentage of an MLB team to a fair extent, a not too unreasonable conclusion considering that baseball is a game of two halves, where one would not be realistically able to dominate the other over a substantial period of time.

Following the single-variable analysis, I wanted to explore further and determine whether these three statistics could prove viable in a multi-variable analysis. Upon researching multi-variate regression, I found ridge regression, which is a methodology that exists for data with collinearity between each variable. Upon building the scatterplots between each of the individual variables, and running the linear model tests, it was clear that each of the variable pairs are collinear.





Upon this determination, I was able to undergo the ridge regression, which consists of five steps, including the setup in R[1]. For this methodology, I used the glmnet package, installing and loading the library into R. I then defined my input and output variables, with the former as a matrix of the three successful batting statistics, and the latter the vector of winning percentages.

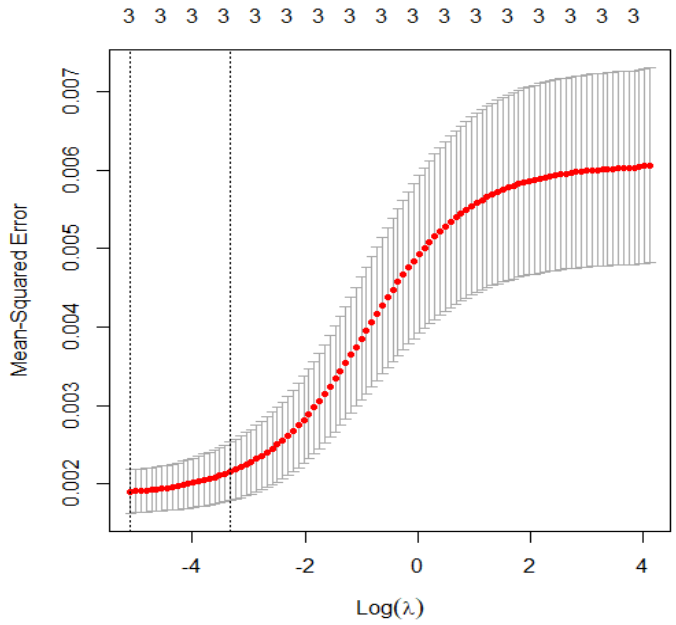
```
#advanced model // ridge regression
#install necessary packages
install.packages("glmnet")
library(glmnet)

#Load Data
#Define Response Variable
y <- winPctMaster
#Define Matrix of Predictors
x <- data.matrix(masterBatting[,c('RBI','OPS.','R.Gm')])
```

The first step within ridge regression is to fit the model, through using the glmnet function and an alpha value of 0.

```
#Fit Regression Model
#Fit model
ridgeModel <- glmnet(x,y, alpha=0)
#View Summary
summary(ridgeModel)
```

Following, the goal is to optimize the lambda value, which minimizes the Mean-Squared Error (MSE) of the function. This step also includes cross validation with k=10 folds.

```
#Choose Optimal Lambda
#k-fold cross validation, k=10
cv_model <- cv.glmnet(x,y,alpha=0)
#find optimal lambda, minimizing MSE
best_lambda <- cv_model$lambda.min;best_lambda
#plot MSE by lambda
plot(cv_model)
```



Once the ideal lambda is determined, I built the ridge model using the glmnet command, specifying the lambda value and running the coef function of the new "best_model."

```
#Analyze model
#find coefficients of best model
best_model <- glmnet(x,y,alpha=0,lambda=best_lambda)
summary(best_model)
coef(best_model) # winpct = 0.0927738331 + 0.0002305
#produce ridge trace plot
plot(ridgeModel, xvar = "lambda")
```

winpct = 0.0927738331 + 0.0002305022(RBI) - 0.0045702695(OPS+) + 0.0855540324(RPG)

Finally, I calculated the R^2 by finding the SST and SSE values of the data, resulting in a value of 0.7205, meaning that 72.05% of the variation in winning percentage of a team can be explained by the change in OPS+, RBI, and RPG from the past three years of MLB play.

While the findings from my research and analysis were successful, there are certain caveats that must be addressed when considering the applications of said findings. First, the period of the data was rather short as I wanted to keep the data manageable, and because baseball is an evolving game, especially in the modern, statcast era, with teams taking every advantage they can get, constantly innovating to win. Secondly, the inclusion of seasonal data rather than single game data may have sacrificed some accuracy for the sake of simplicity in the data. Statcast reports queries 200 results per page on my system, and considering that there are 4,860 games in a season, it was not time efficient to manually create this dataset. Finally, of course, is a caveat mentioned earlier: offense is only half of the game. Pitching and defense also provide a massive boon towards a team's chances of winning and this aspect cannot be understated. In essence, batting statistics impact a team's winning percentage to a fair extent, within the restraints of the nature of the sport, and the run scoring environments in which the games are played.

References

"Team Batting Season Stats Finder." *Stathead.com*, https://stathead.com/baseball/team-batting-season-finder.cgi.

Zach. "Ridge Regression in R (Step-by-Step)." *Statology*, 13 Nov. 2020, https://www.statology.org/ridge-regression-in-r/.