Major League Baseball

Moneyball: The Statistics of the Game

Final Project – STA 3000
Max Shalom

Introduction

What metrics most closely correlate to a strong baseball player? How can we find the most relevant gauge to predict which players are best?

Using publicly available data from the MLB and Kaggle, players' statistics can be analyzed to understand which statistics are the most important when drafting new players or trading between teams. Such techniques were used by the Oakland A's in 2002, as portrayed in the movie Moneyball, starring Brad Pitt. Since the team was strapped for cash, they turned to alternate ways of understanding how a player can be developed and the potential each player has based on statistics that may go overlooked. In recent years, statistics in the MLB have drawn more attention from researchers and mathematicians. The MLB stats are much more advanced than that of any other professional sports league because of the massive number of variables being tracked. These variables allow for composite scores to be computed, such as Fangraph's, fWAR (wins above replacement). One of baseball's most advanced and comprehensive statistics, fWAR (and other WAR formulas by other statistic agencies), takes into account both the offensive and defensive statistics of a player and compares that to a normal player of that position. The composite score is analyzed to calculate how many more wins the player has given their team compared to that of a normal player in that position.

The goal of my project is to analyze the evolution of the league over time. In seeing the growth of the league and the betterment of players becoming more competitive, I would like to research which parts of the game have become more important in recent years. I can analyze the correlation between certain metrics, and how they translate to a more efficient player. I will use charts to analyze the average statistics by year and I can also calculate that correlation with newer statistical formulas being used in the league to define which is the most efficient and thus, finding the best way to measure a player.

Data Sources

Savant & Statcast

The MLB runs a website called Baseball Savant which includes troves of data on every player and team, across all positions. This comprehensive website includes historical data as far back as 1950, current data, and even short-term forecasts on future data predictions, such as probable pitchers. Savant tracks all data and can compute live leaderboards for daily, weekly, monthly, and pretty much any other timeframe's data to show, for example, top pitch velocity of the day, top exit velocity of the week, or even fastest pitch speed of the month. Much of this data is run by Google Cloud.

In 2015, the MLB entered the "Statcast era". Statcast is an innovation in tracking technology that allows for the real-time collection and analysis of massive amounts of baseball data, in ways that were never possible in the past. After a trial run in 2014, all 30 MLB ballparks were retrofitted with 12 'Hawk-Eye' cameras, installed in an array around the field. With high frames-per-second rates, pitch tracking is meticulously watched, and other cameras track all players' fielding and batting metrics. The new system assists in raising the percentage of tracked batted balls from ~89% to ~99%, complete with additional metrics on these balls we did not have before. Statcast data is hosted on the Savant platform.

> *"These metrics allow front offices, broadcasters and fans alike to quantify the raw skills of players in ways that were previously available only to scouts or not available at all. Even in a 2020 regular season shortened to 60 games by the pandemic, more than 260,000 pitches and 43,000 batted balls were tracked. Meanwhile, terms such as "spin rate" and "launch angle" have become ubiquitous not just on broadcasts, but also on ballpark video boards and even on the field, as players across the league use the data and the thinking behind it to elevate their games."*
>
> *MLB.com Glossary - STATCAST*

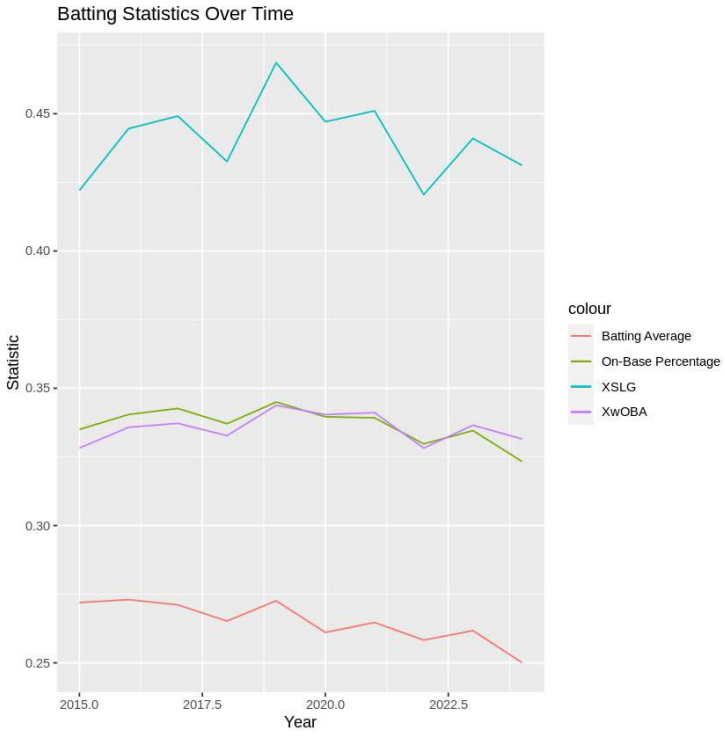
Project Data

As mentioned above, teams, media, and fans are becoming more familiar with advanced technical analysis. This elevates the game by teams understanding the importance of forgotten or previously untraceable metrics, thus giving more importance to advanced statistics in valuing players and acquiring new ones. The data I will be exploring from Statcast includes all individual player statistics from its inception in 2015 to 2024. It includes 30 columns of various advanced Statcast data metrics. I will be delving into the select few below:

- Basic data – player name, year of season, player age
- OBP (On base %) – measures how frequently a batter reaches base per plate appearance.

- SLG (slugging percentage) - is a measure of the power of a hitter. It calculates the total number of bases a player records per at-bat.
- xSLG (expected slugging percentage) - estimates a player's slugging percentage based on the quality of contact (exit velocity, launch angle, and batted ball types)
- OPS+ (On-base Plus Slugging Plus)- takes a player's on-base plus slugging percentage and normalizes the number across the entire league.
- WOBA (weighted on-base average) - A statistic that measures a player's overall offensive contributions per plate appearance.
- xWOBA (expected weighted on-base average) – An estimate/forecast of a player's wOBA based on the quality of their contact (exit velocity, launch angle, etc.).
- Sweet spot % - The percentage of batted balls that are hit in the "sweet spot" for optimal performance. These are hit within a certain optimal range (between 8 and 32 degrees) of launch angle and exit velocity.
- Barrel batted rate - The percentage of batted balls that are "barreled" (well-struck).
- Hard hit % - The percentage of batted balls that are hit hard (typically with an exit velocity of 95 mph or higher).

FIGURE 1



Over this time period, average batting statistics for the league have varied. As an introduction for this data, I have created a chart [fig. 1] portraying the evolution of the game over the studies time period (2015-2024) [Appendix Code Block C]. During this time period, the mean batting average actually dropped closer to 0.250. Historically, this is low and has been dropping steadily possibly due to an influx of stronger pitchers introducing alternate pitching methods unheard of in the past. While AVG did drop, OBP and xWOBA were relatively unchanged. In addition, xSLG, while varying throughout, has also leveled out. xSLG is a predictive metric which gives a more accurate picture of a player's power by focusing on different factors that contribute to the original SLG. As mentioned above, this includes quality of contact and estimates total bases. As such, this can prove the hypothesis correct that batter contact may be steady, yet batting average is decreasing due to a near 6% increase in walks in 2023 vs the base year (2015) (Walks are not counted as hits in a batting average but are counted as a base for OBP purposes).

Statistical Analysis

Question

Which statistics are the most significant in choosing a player based on skill?

Correlation Matrix

Using select important metrics, I created a correlation coefficient matrix [fig. 2] using the `corrplot` package [code block B]. With this, one is able to understand the relevance between these certain metrics. In addition, dependencies become clear. For instance, WOBA, has strong correlations with OBP and SLG (0.85 and 0.92, respectively) because these are inputs into the larger WOBA formula. The formulas that are the most comprehensive and incorporate as many other as possible seem to be the most efficient way of valuing a player.

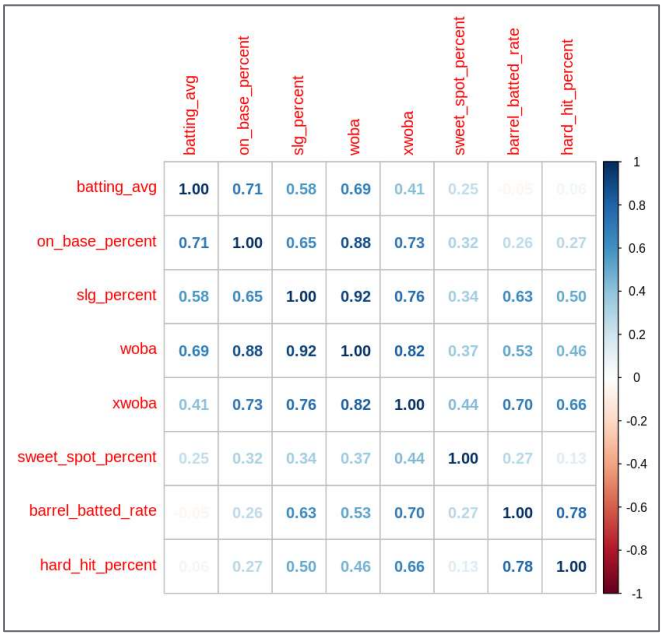


FIGURE 2: CORRELATION MATRIX OF SELECTED VARIABLES

The barrel batted rate (percentage of 'well-struck' balls) is closely correlated with SLG and hard hit %. Well-struck/barreled is defined as a batted ball with similar hit types in exit velocity and a launch angle that has led to a minimum .500 batting average and 1.500 slugging percentage. The batted ball requires an exit velocity of 98 miles per hour to be barreled. On the matrix, it is strongly correlated with SLG and hard hit %. When barreling a ball, it is likely that a strong base hit will come of it, which is one of the most heavily weighted parts of the SLG formula.

As shown in fig. 3 [appendix code block D], there is a significant correlation of 0.59 between barreling a ball and hitting a home run. A batter who can achieve this is more likely to be a strong all-around hitter and a worthy player for your team.

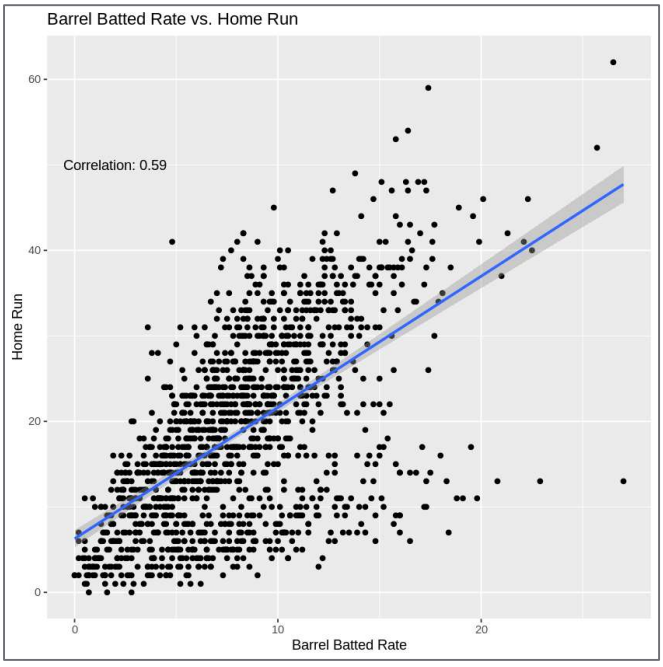


FIGURE 3: CORRELATION BETWEEN BARREL RATE AND HOME RUNS HIT

Linear Regression Model

Expanding on the relationship between barrel rate and home run, I have performed a linear regression analysis with the below results [appendix code block E]:

```
Correlation: 0.59

Call:
lm(formula = home_run ~ barrel_batted_rate, data = stats)

Residuals:
    Min      1Q  Median      3Q     Max
-34.725  -5.260   0.820   6.053  27.357

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.27356    0.51579   12.16   <2e-16 ***
barrel_batted_rate  1.53524    0.05651   27.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.86 on 1404 degrees of freedom
Multiple R-squared:  0.3446,   Adjusted R-squared:  0.3441
F-statistic: 738.2 on 1 and 1404 DF,  p-value: < 2.2e-16
```

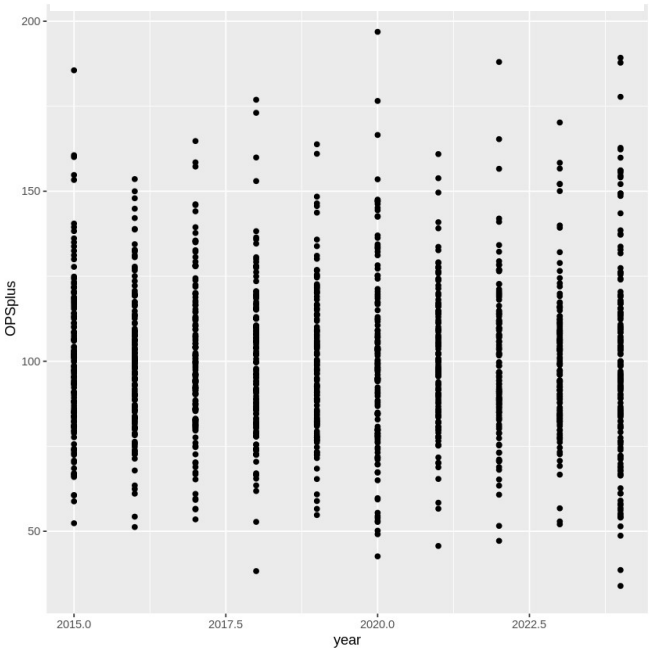
This model was fit with home_run as the dependent variable and barrel_batted_rate as the independent variable using the data in the stats dataframe. With this correlation we are able to tell a lot about its impact. The intercept of 6.27 is the expected number of home runs when the barrel_batted_rate is 0. We can also tell that for every 1% increase in barrel_batted_rate, the number of home runs is expected to increase by approximately 1.54! The low p value indicates that the model as a whole is statistically significant.

OPS+ (On-base Plus Slugging Plus)

OPS+ is an extremely versatile statistic because it involves relative averages between the league, grouped by year. The formula is $OPS+ = 100 * (OBP/(leagueAvg\ OBP) + SLG/(leagueAvg\ SLG)) - 1$. By looking at players comparatively throughout the league, one is able to determine the athletic advantage one player has over another because the league average statistic takes into account variables such as ballpark difference or external factors. The baseline for this statistic is always set at 100 and a player 50% better than the average would have an OPS+ of 150.

In the figure to the right, I have computed OPS+ by grouping all OBP/SLG data by year and determining the new column for each player using a scatterplot [code block A (see appendix)]. I have used the above formula to implement this alongside metrics I already had. By design, all players are strongly correlated near the mean of 100. The upper outliers are the true stars of the game. The greatest OPS+ in this dataset is Juan Soto, who achieved a 196.87 during his 2020 season on the Washington Nationals. This means he was nearly twice as good as the average player (Note: 2020 was a limited season due to the pandemic).

FIGURE 4: OPS+ BY YEAR PER PLAYER



When looking to trade for a new player, it would be logical for a general manager to study a prospect's OPS+ compared to that of the other prospects. This is an extremely significant statistic that comprehensively shows a player's offensive performance and how that compares to the league average, just with a quick glance.

Conclusion

Barrel rate and OPS+ are very strong metrics to look at when deciding on players for your team, perhaps as someone working in a team's front office. These have strong correlations with other data that is inputted into their formulas, therefore they are the most comprehensive analysis you can see at a quick glance. I am excited to see how Statcast evolves and continues to grow the sport of Baseball with an eye towards statistical analysis.

Major League Baseball

Project Appendix

Max Shalom - STA 3000

Setup Code

```
library(readr)
stats <- read_csv("stats.csv")
summary(stats)

library(dplyr)
library(ggplot2)
```

Code Block A

Calculating an additional statistic of OPS+

```
# Calculate league averages for OBP and SLG by year which is needed for the OPS+ formula
league_stats <- stats %>%
  group_by(year) %>%
  summarise(lgOBP = mean(on_base_percent),
            lgSLG = mean(slg_percent))

# Add OPS+ data by using the formula
stats <- stats %>%
  left_join(league_stats, by = "Year") %>%
  mutate(OPSplus = 100*((on_base_percent/lgOBP)+(slg_percent/lgSLG)-1))

# Create scatterplot of OPS+ by year with each player plotted
ggplot(stats, aes(x = year, y = OPSplus)) +
  geom_point()
```

Code Block B

Creating a correlation matrix

```
install.packages("corrplot")
library(corrplot)

correlation_matrix <- cor(stats[, c("batting_avg", "on_base_percent", "slg_percent",
"woba", "xwoba", "sweet_spot_percent", "barrel_batted_rate", "hard_hit_percent")])
corrplot(correlation_matrix, method = "number")
```

Code Block C

Line chart to show different statistics over time

```
stats_by_year <- stats %>%
  filter(year >= 2000) %>%
  group_by(year) %>%
  summarise(
    batting_avg = mean(batting_avg, na.rm = TRUE),
    on_base_percent = mean(on_base_percent, na.rm = TRUE),
    xslg = mean(xslg, na.rm = TRUE),
    xwoba = mean(xwoba, na.rm = TRUE)
  )

ggplot(stats_by_year, aes(x = year)) +
  geom_line(aes(y = batting_avg, color = "Batting Average")) +
  geom_line(aes(y = on_base_percent, color = "On-Base Percentage")) +
  geom_line(aes(y = xslg, color = "XSLG")) +
  geom_line(aes(y = xwoba, color = "XwOBA")) +
  labs(title = "Batting Statistics Over Time",
       x = "Year",
       y = "Statistic")
```

Code Block D

```
# Calculate the correlation between barrel_batted_rate and home_run
correlation <- cor(stats$barrel_batted_rate, stats$home_run)

# Create a scatterplot of barrel_batted_rate vs. home_run
ggplot(stats, aes(x = barrel_batted_rate, y = home_run)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Barrel Batted Rate vs. Home Run",
       x = "Barrel Batted Rate",
       y = "Home Run") +
  annotate("text", x = 2, y = 50, label = paste("Correlation:", round(correlation, 2)))
```

Code Block E

```r
# Calculate the correlation between barrel_batted_rate and home_run
correlation <- cor(stats$barrel_batted_rate, stats$home_run)

# Fit a linear model to the data
model <- lm(home_run ~ barrel_batted_rate, data = stats)

# show the correlation and model summary
print(paste("Correlation:", round(correlation, 2)))
summary(model)
```