MLB Batting Average Predictor

Setup Project on Local Machine

- Download files
- Move downloaded folder to directory you wish to run script in
- Open RStudio or other R software
- In Console: Set working directory to filepath leading to downloaded folder was placed
  - setwd(filePath)
- In Console: Verify working directory points to downloaded folder
  - getwd()
- In Console: Load Script holding functions to perform analysis of player performance data
  - source("average_estimator_functions.R")
- In Console: Run main function to output analysis of player performance data
  - main()

Methodology

Player's season-long batting averages were predicted through two steps:

1. Estimating a player's season-long strikeout and walk percentage based on historical league averages.
2. Utilizing the estimated strikeout and walk rates to predict the batting average a player should have based on a range of conservative BABIP values.
3. Utilizing median estimated batting averages from Step 2 and current March/April batting average to output a final estimate for season-long batting average

Estimating Season-Long Strikeout and Walk Rates

Because a player may be showing inflated or deflated play in his first month of games, we can try to stabilize his performance to better represent his performance for an entire season. Historical averages for strikeout and walk rates and a stabilization points for number of plate appearances can be used to calculate a player's stabilized performance. A stabilization point of 100 PA for strikeout rate and 168 PA for walk rate has been found in previous studies to serve as conservative thresholds.

Estimates for season-long strikeout and walk rates are found using the following formula:

- K%_Est = (Player K% * Player PA + League AVG K% * 100 PA) / (Player PA + 100 PA)
- BB%_Est = (Player BB% * Player PA + League AVG BB% * 168 PA) / (Player PA + 168 PA)

League averages of 8.28% and 18.73% for walk rate and strikeout rates respectively, were found by using season long averages for both statistics on Fangraphs (https://goo.gl/Kmf1po).

A player's season-long batting average can now be calculated using our estimates for that player's season-long strikeout and walk rates.

Predicting Season-Long Batting Average with Historical BABIP Values

Historically, 90% of players in the league have a BABIP in the range of 0.250 - 0.360. We can use this range to gather an estimate for a player's batting average. BA is calculated through the following formula:

- Num. Hits / Num. At Bats

Using our estimates found in Step 1 and our range of BABIP values (0.380, 0.360, 0.340, 0.270, 0.250, 0.230), we can estimate the number of hits a player should have had during the March/Apr games for each BABIP value:

- Est # of Plate Appearances (PA_est) = (PA - (BB_est * PA) - (K_est * PA))
- Est # of Hits (Hits_est) = PA_est * BABIP_i + HR

We can estimate the number of At Bats by subtracting the estimated number of walks the player would have had using our season-long estimate:

- Est # of At Bats (AB_est) = PA - (BB_est * PA)

We then calculate our estimated batting average for each BABIP value to produce a range of season-long batting average estimates:

- BA_est = Hits_est / AB_est

Which produces a data table like so:

Name	MarApr_AVG	FullSeason_AVG	0.380	0.360	0.340	0.270	0.250	0.230
Adam Eaton	0.298	0.284	0.336	0.318	0.302	0.241	0.224	0.207
Adam Jones	0.224	0.265	0.311	0.295	0.280	0.225	0.210	0.194
Addison Russell	0.214	0.238	0.346	0.329	0.313	0.254	0.237	0.221
Adeiny Hechavarria	0.222	0.236	0.333	0.317	0.301	0.244	0.229	0.212
Adonis Garcia	0.282	0.273	0.309	0.293	0.278	0.222	0.207	0.192

Outputting Final Estimate for Season-Long Batting Average

With our range of batting averages for each BABIP value, we can develop a final estimate for a player's season-long estimate by utilizing their March/April batting average and the median batting average estimates in our BABIP range (0.340 & 0.270). The 0.340 and 0.270 were chosen to offset any inflated or deflated March/April batting averages and served as conservative estimates for a player's average estimated performance.

- Season-Long Batting Average (BA_Season_est) = mean(March/April_BA + BA_0.340 + BA_0.270)

Once this formula is applied to all players with more than zero plate appearances, we can take the mean season-long batting average estimate and apply that to all players with zero plate appearances. Our final data table is as follows:

Name	MarApr_AVG	FullSeason_AVG	avgEstimate	percentError
Adam Eaton	0.298	0.284	0.280	1.29 %

Name	MarApr_AVG	FullSeason_AVG	avgEstimate	percentError
Adam Jones	0.224	0.265	0.243	8.26 %
Addison Russell	0.214	0.238	0.260	9.38 &
Adeiny Hechavarria	0.222	0.236	0.256	8.44 %
Adonis Garcia	0.282	0.273	0.261	4.45 %
Adrian Beltre	0.289	0.300	0.291	2.87 %

Estimated Batting Averages for Players with Greater than Zero Plate Appearances

Name	MarApr_AVG	FullSeason_AVG	avgEstimate	percentError
Adam Duvall	NA	0.241	0.276	14.58 %
Brad Miller	NA	0.243	0.276	13.64 %
Charlie Blackmon	NA	0.324	0.276	14.77 &
Cheslor Cuthbert	NA	0.274	0.276	0.78 %
Danny Valencia	NA	0.287	0.276	3.78 %
Eduardo Nunez	NA	0.388	0.276	4.12 %

Estimated Batting Averages for Players with Zero Plate Appearances

Results

Accuracy of this method can be evaluated by considering the median and mean percent error generated by two methods used to estimate a player's season-long batting average: 1. Holding a player's current Mar/Apr batting average fixed, assuming they will continue to play at this level the entire season 2. Using a player's current Mar/Apr batting average, adjusting for their strikeout and walk rates as described in Step 1, and utilizing the estimated batting averages for target BABIP values as described in Step 2.

The resulting median and mean percent error for each method is shown below:

Method	Median Percent Error	Mean Percent Error
Method 1	10.24 %	11.82 %
Method 2	7.55 %	8.81 %

From the error analysis, we see that Method 2 demonstrates a better accuracy for estimating a player's season-long batting average, displaying approximately a 3% decrease in median and mean percent error.

References

1. Methodology and assumptions made regarding historical BABIP range and stabilization points for strikeout and walk rates inspired by Jeff Zimmerman's article, "The Metric System: Predicting Batting Average": https://goo.gl/9Aiw6d