**Base Accumulation by Player Ability**

CSCI 4831 - Sabermetrics Final Project

*Jackson Mediavilla*

*Professor Hoenigman*

*7 May 2018*

**Explanation of Statistic**

For the final project in this course, I am proposing a new statistic for evaluating hitters. The statistic accounts for hitting and baserunning factors over which the batter has control. The purpose of this statistic is to evaluate how well players can accumulate bases on the basis of their own skill. I include the following factors in my equation: singles, doubles, triples, home runs, steals, caught stealing, and at bats. I have not included factors over which the batter does not have full control (such as base on balls and hit by pitch).

I have named my new statistic Base Accumulation by Player Ability (or BAPA for short). The equation I have developed is as follows:

where *SB* is stolen bases and *CS* is caught stealing. The numerator of this equation first sums all bases accumulated by hits alone (, then adds successfully stolen bases, and finally subtracts caught stealing. This total should give an idea of how well a player can accumulate bases based on their batting and baserunning abilities. There exists a denominator to this equation to place all players on a more even playing field. The first part of the denominator calculates at-bats where the batter did not get a hit, and then caught stealing is added to lower the BAPA for getting out on a steal attempt. Ideally, the denominator should adjust the numerator to a more normalized scale so that players can more easily be compared.

The equation above is similar to a couple of other statistics. The first statistic, total bases, is included in the above equation. Total bases is a total number representing all accumulated bases from hits. *Total bases =* . The second similar statistic is called Total Average. The equation for Total Average calculates a very similar statistic to BAPA, except that it includes factors not totally under the batter/runner’s control (base on balls, hit by pitch, and ground into double plays). *Total Average = [(TB + BB + HBP + SB – CS)/(AB – H + CS + GIDP)].*

**Evaluation of Statistic**

I have evaluated my statistic by comparing it to other common baseball statistics. I created a Jupyter Notebook to perform the statistical analyses. The notebook can be found under the *code* folder of this project’s GitHub repository.An HTML version of the file can also be found in the repository under the *data* folder. For all of the following regression analyses, I split the data 80% for the training set and 20% for the test set.

The first statistic to which I made comparisons is on-base percentage (OBP). Linear regression analysis yielded an equation of

*On-Base Percentage = 0.367 \* BAPA + 0.0946*

with a mean squared error of 0.01 and a variance score of 0.72. The low mean squared error of 0.01 shows that BAPA is very highly correlated to on-base percentage.

FIGURE 1

Fitted Linear Regression Comparing BAPA and OBP

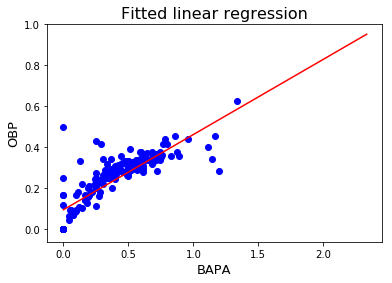


Figure 1 depicts the relationship between BAPA and on-base percentage. The visual representation shows a clear, positive relationship between the two variables.

I also performed linear regression analysis between BAPA and batting average (BA). This yielded an equation of

*Batting Average = 0.313 \* BAPA + 0.0784*

with a mean squared error of 0.00 and a variance score of 0.74. The extremely low mean squared error shows a very high correlation between BAPA and batting average.

FIGURE 2

Fitted Linear Regression Comparing BAPA and BA

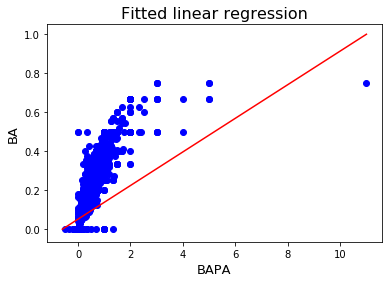


Figure 2 depicts the relationship between BAPA and batting average. While the mean squared error rounds to 0.00, the graph shows the existence of outliers in the relationship. These outliers become more apparent in my later analyses.

Another relationship I examined with linear regression analysis was BAPA vs slugging percentage (SLG). The resulting equation is as follows

*Slugging Percentage = 0.594 \* BAPA + 0.0556*

with a mean squared error of 0.00 and a variance score of 0.95. The extremely low mean squared error shows a very high correlation between BAPA and slugging percentage.

FIGURE 3

Fitted Linear Regression Comparing BAPA and SLG

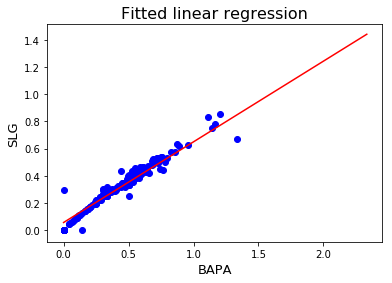


Figure 3 shows the relationship between BAPA and SLG. Of the linear regression graphs generated, this one shows the strongest relationship. This is probably because the very high outlier BAPA value was dropped. The value was dropped because in all of these regression calculations, all rows were dropped that did not contain both the BAPA statistic and the dependent variable.

The last regression analysis I performed was between BAPA and the Linear Weights (LWTS) formula developed by Pete Palmer.

*LWTS = .46\*1B + .80\*2B + 1.02\*3B + 1.40\*HR + .33\*(BB + HBP) + .30\*SB – .60\*CS – .25\*(AB – H)*

The regression equation is as follows

*LWTS = 19.2 \* BAPA + -6.01*

with a mean squared error of 141.97 and a variance score of 0.20.

FIGURE 4

Fitted Linear Regression Comparing BAPA and LWTS

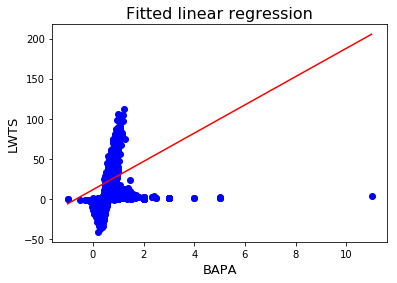


Figure 4 and the regression equation show a very weak relationship between BAPA and the linear weights statistic.

**Conclusions**

Based on the analysis above, I believe BAPA has potential for evaluating baseball players’ offensive skills, but it does not provide a great overall picture of players. This is evident in the fact that Base Accumulation by Player Ability is highly correlated to on-base percentage, batting average, and slugging percentage, but is not highly correlated to LWTS. The linear weights formula is intended to give an overall score to players to analyze their ability to contribute runs. This is very similar to what I intended to evaluate with my statistic. Perhaps BAPA could be a much more effective statistic by weighting events in the same manner as the linear weights formula.

More problems with the statistic become evident when looking at the top ranked players under BAPA’s evaluation.

FIGURE 5

Top Ranked Players by BAPA

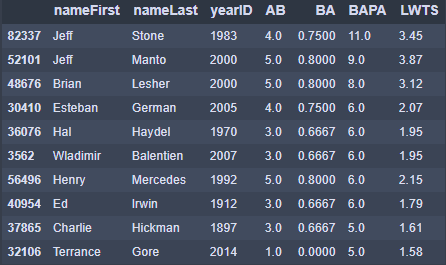


Figure 5 shows the top ranked players in the Lahman database by their BAPA score. It is evident that BAPA needs modification because all of these players have very few at-bats and very low LWTS.

If the query is changed to only include players with more than 300 at-bats, BAPA produces a much more accurate list of top players.

FIGURE 6

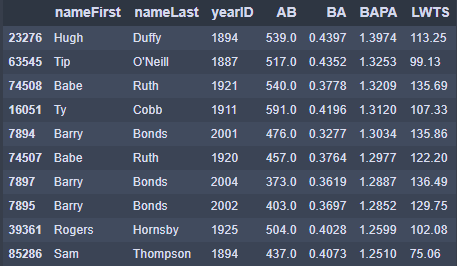
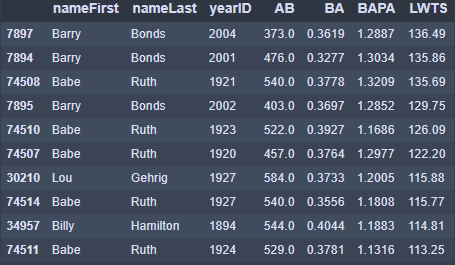
Top Ranked Players by BAPA (AB-Adjusted) 

Figure 6 shows a more realistic list of the top 10 players from the Lahman Database.

The higher accuracy of Figure 6 can be seen in comparison to the top players by LWTS.

FIGURE 7

Top Ranked Players by LWTS

Comparing the list of players in Figure 7 to the list of players in Figure 6 shows a much higher correlation than comparing Figures 5 and 7.

In conclusion, I believe BAPA is an okay evaluation of baseball offensive skill. With some modifications (such as weights), BAPA could be developed into a more accurate representation of players’ ability to accumulate bases on their own skill.