

# Implicitly Defined Baseball Statistics

December 9, 2012

Joe Scott

## 1 Introduction

Major League Baseball uses statistics to determine awards every season. The batting champion is given to the player with the highest batting average. The Cy Young Award is given to the top pitcher which is determined by many different statistics including earned run average (ERA). Batting average and ERA have been used for many years and are major statistics in baseball. Neither batting average or ERA consider the skill of the opposing pitcher or batter. Thus, every pitcher and batter is considered to have the same skill level. We develop an implicitly defined statistic that determines the skill or value of a player. The value of a batter and the value of a pitcher is based on the skill of the opposing pitcher and batter respectively. We use linear algebra to find eigenvector solutions to the eigenvalue problem,  $A\lambda = \lambda x$ , which generates each player's statistical value.

## 2 Idea

Consider a baseball league in which there are  $N_b$  players who bat, represented by  $b_i$  for  $1 \leq i \leq N_b$ . We represent the number of pitchers in the league as  $p_j$ ,  $1 \leq j \leq N_p$  where  $N_p$  is the number of pitchers.  $N_b$  is defined as the number players who record an at bat during a specific season and  $N_p$  is the number of players who record a pitching appearance during a season. The total number of players in the league,  $N_{tp}$ , is represented by the inequality  $N_{tb} \leq N_b + N_p$ . This inequality considers players who both hit and pitch. Since in the National League pitchers hit as well as pitch we need to add the pitchers to the total number of batters and in interleague play (which is when American League teams face National League teams in the regular season) American League pitchers bat when the National League team is home. For each batter, a batting average,  $ba_i$ , is produced by

$$ba_i = \frac{h_i}{ab_i} = \frac{1}{ab_i} \sum_{j=1}^{N_p} h_{i,j}$$

where  $h_i$  is the number of hits recorded for batter  $i$  and  $ab_i$  is the number of at bats recorded by batter  $i$ . A similar statistic for pitchers is called opponents batting average,  $oba$ . Opponents batting average is determined by number of hits against a pitcher  $j$  divided by the number of at bats against a pitcher  $j$ . So,

$$oba_j = \frac{oh_j}{oab_j},$$

where  $oh_j$  is number of hits against pitcher  $j$  and  $oab_j$  is number of at bats against pitcher  $j$ . Opponents batting average is not sufficient because it places emphasis on the success of the hitter so we define pitcher effectiveness of pitcher  $j$  as

$$pe_j = \frac{oab_j - oh_j}{oab_j}.$$

Thus, we consider how successful the pitcher is at recording outs. However, pitching is not only defined by at bats and hits. A pitcher could walk or hit a batter. Thus, we consider plate appearance,  $pa_j$  and the number of times a player gets on base,  $ob_j$ . We redefine pitcher effectiveness as

$$pe_j = \frac{pa_j - ob_j}{pa_j} = \frac{1}{pa_j} \sum_{i=1}^{N_b} (pa_{i,j} - ob_{i,j}).$$

That is,  $pe_j$  takes plate appearance minus the number of men that reach base and divide the difference by number of plate appearances.

There exist many situations where batting average and pitcher effectiveness is not a good indication of the player's skill. In a division where the average pitcher effectiveness is low, a batter could obtain more hits, thus raising his batting average. This creates a problem when the batter is compared to a different batter who plays in a division where the pitcher effectiveness is high. The same argument could be made for pitcher effectiveness. We seek a metric that "levels the playing field".

### 3 The Statistic

We assign weights to each batter and pitcher,  $wba$  and  $wpe$  respectively. For batters, we define the weighted batting average,

$$wba_i = \frac{1}{ab_i} \sum_{j=1}^{N_p} wpe_j h_{i,j}$$

where  $wpe_j$  is the weighted pitching effectiveness of pitcher  $j$ . Similarly,

$$wpe_j = \frac{1}{pa_j} \sum_{i=1}^{N_b} wba_i (pa_{i,j} - ob_{i,j}).$$

That is, we define  $N_b$  weights,  $wba_i$  for  $1 \leq i \leq N_b$ , and  $N_p$  weights,  $wpe_j$  for  $1 \leq j \leq N_p$  as

$$\begin{aligned}
wba_1 &= \frac{1}{ab_1} \sum_{j=1}^{N_p} wpe_j h_{1,j} \\
wba_2 &= \frac{1}{ab_2} \sum_{j=1}^{N_p} wpe_j h_{2,j} \\
&\vdots \\
wba_{N_b} &= \frac{1}{ab_{N_b}} \sum_{j=1}^{N_p} wpe_j h_{N_b,j} \\
wpe_1 &= \frac{1}{pa_1} \sum_{i=1}^{N_b} wba_i (pa_{i,1} - ob_{i,1}) \\
wpe_2 &= \frac{1}{pa_2} \sum_{i=1}^{N_b} wba_i (pa_{i,2} - ob_{i,2}) \\
&\vdots \\
wpe_{N_p} &= \frac{1}{pa_{N_p}} \sum_{i=1}^{N_b} wba_i (pa_{i,N_p} - ob_{i,N_p}).
\end{aligned} \tag{1}$$

We define the weighted batting average vector  $wba$ , and the weighted pitching effectiveness,  $wpe$  as

$$wba = \begin{bmatrix} wba_1 \\ wba_2 \\ \vdots \\ wba_{N_b} \end{bmatrix}, \quad wpe = \begin{bmatrix} wpe_1 \\ wpe_2 \\ \vdots \\ wpe_{N_p} \end{bmatrix}.$$

Combining these vectors, we achieve the weight vector

$$w = \begin{bmatrix} wba \\ wpe \end{bmatrix}.$$

Let

$$(\mathbf{AB})_{i,j} = ab_{i,j},$$

$$(\mathbf{H})_{i,j} = h_{i,j}$$

$$(\mathbf{PA})_{i,j} = pa_{i,j},$$

$$(\mathbf{OB})_{i,j} = ob_{i,j}$$

be  $N_b \times N_p$  matrices for at bats, hits, plate appearances, and on-base respectively. So, system (1) can be written as

$$\begin{bmatrix} wba \\ wpe \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{M} \cdot \mathbf{H} \\ \mathbf{N} \cdot (\mathbf{PA} - \mathbf{OB})^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} wba \\ wpe \end{bmatrix}, \quad (2)$$

and

$$M = \begin{bmatrix} \frac{1}{ab_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{ab_2} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \cdots & \frac{1}{ab_{N_b}} \end{bmatrix}$$

and

$$N = \begin{bmatrix} \frac{1}{pa_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{pa_2} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \cdots & \frac{1}{pa_{N_p}} \end{bmatrix}.$$

If we define

$$\mathbf{C} = \begin{bmatrix} \mathbf{0} & \mathbf{M} \cdot \mathbf{H} \\ \mathbf{N} \cdot (\mathbf{PA} - \mathbf{OB})^T & \mathbf{0} \end{bmatrix}, \quad (3)$$

then system (2) can be expressed as the following system

$$w = \mathbf{C}w.$$

However, this system may not have a solution except when  $w=0$ . So, we consider the problem

$$\lambda w = \mathbf{C}w.$$

In general, there could be up to  $N_b + N_p$  number of options for  $\lambda$ . To we find a unique  $\lambda$ , we enforce the following properties. We want  $\lambda$  to be positive and real whose corresponding eigenvector is either non-negative or non-positive. In the past ten years of baseball statistics, only one  $\lambda$  that fits the criteria each year.

The following theorem gives conditions under which  $\mathbf{C}$  meets all requirements.

**Theorem 1. Perron-Frobenius Theorem** *Let  $A$  be an irreducible non-negative  $n \times n$  matrix. Then  $\mathbf{A}$  has a real eigenvalue  $\lambda_1$  with the following properties:*

- a)  $\lambda_1 > 0$
- b)  $\lambda_1$  has a corresponding positive eigenvector.

All of our matrices are non-negative due to the fact that in baseball, at bats, hits, plate appearances and ability to be on base are only considered in the natural numbers. Thus,  $\mathbf{C}$  is an  $(N_p + N_b) \times (N_p + N_b)$  matrix that is non-negative. We must now consider the irreducibility. An irreducible  $n \times n$  matrix  $\mathbf{A}$  exist if and only if there does not exist a permutation matrix  $\mathbf{P}$ , such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix}$$

where  $A_1 \neq 0$  and  $A_3 \neq 0$ . In the past ten years, baseball has satisfied this condition but conditions under which this is guaranteed are unknown.

## 4 Example

Consider Joe's Baseball League,

hits/ at bats	Jim	Greg	Rich	Mike	Evan
Brian	1/4	1/3	2/8	0/1	1/2
Tommy	2/5	2/8	5/8	1/5	0/0
Cody	5/6	2/5	0/2	2/4	0/1
Derek	3/7	3/6	0/9	5/7	4/9
James	1/4	0/0	1/8	3/10	1/3
Bob	2/7	0/6	0/3	1/5	3/4

on base/plate appearances	Jim	Greg	Rich	Mike	Evan
Brian	2/5	1/7	2/8	1/3	1/5
Tommy	2/5	2/8	5/8	1/5	0/4
Cody	6/8	5/9	0/2	2/5	0/6
Derek	5/9	3/8	0/9	5/7	4/12
James	1/8	2/2	1/8	6/13	2/4
Bob	3/9	0/8	0/3	2/6	5/7

$$\mathbf{AB} = \begin{bmatrix} 4 & 3 & 8 & 1 & 2 \\ 5 & 8 & 8 & 5 & 0 \\ 6 & 5 & 2 & 4 & 1 \\ 7 & 6 & 9 & 7 & 9 \\ 4 & 0 & 8 & 10 & 3 \\ 7 & 6 & 3 & 5 & 4 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 & 1 & 2 & 0 & 1 \\ 2 & 2 & 5 & 1 & 0 \\ 5 & 2 & 0 & 2 & 0 \\ 3 & 3 & 0 & 5 & 4 \\ 1 & 0 & 1 & 3 & 1 \\ 2 & 0 & 0 & 1 & 3 \end{bmatrix}$$

$$\mathbf{PA} = \begin{bmatrix} 5 & 7 & 8 & 3 & 5 \\ 5 & 8 & 8 & 5 & 4 \\ 8 & 9 & 2 & 5 & 4 \\ 9 & 8 & 9 & 7 & 12 \\ 8 & 2 & 8 & 13 & 4 \\ 9 & 8 & 3 & 6 & 7 \end{bmatrix}, \quad \mathbf{OB} = \begin{bmatrix} 2 & 1 & 2 & 1 & 1 \\ 2 & 2 & 5 & 1 & 0 \\ 6 & 5 & 0 & 2 & 0 \\ 5 & 3 & 0 & 5 & 4 \\ 1 & 2 & 1 & 6 & 2 \\ 3 & 0 & 0 & 2 & 5 \end{bmatrix}$$

Using these matrices and (3), we solve the eigenvalue/eigenvector problem  $\lambda \mathbf{w} = \mathbf{Cw}$ . This yields the following solutions.

$$\lambda \approx 0.4592$$

Batter	BA (Rank)	WBA (Rank)	Pitcher	PE (Rank)	WPE (Rank)
Brian	0.278(4)	0.224 (4)	Jim	0.432 (2)	0.276(4)
Tommy	0.385 (3)	0.304 (2)	Greg	0.310 (4)	0.377 (3)
Cody	0.500(1)	0.327(1)	Rich	0.211 (5)	0.409(1)
Derek	0.395 (2)	0.281 (3)	Mike	0.436 (1)	0.283 (5)
James	0.240 (5)	0.167 (6)	Evan	0.333 (3)	0.384 (2)
Bob	0.240 (5)	0.173 (5)			

## 5 2012 results

Using play-by-play files, we were able to calculate all weights for the 692 batters and 715 pitchers. The results for 2012 are as follows:

Batter	WBA (Rank)	BA (Rank)	Pitcher	OBA (Rank)	WPE (Rank)
Miguel Cabrera	0.59785(1)	0.330 (2)	Felix Hernandez	0.241 (26)	1.5708(1)
Adam Jones	0.56685 (2)	0.287(44)	James Shields	0.239 (23)	1.5472(2)
Albert Pujols	0.56255 (3)	0.285 (50))	Hiroki Kuroda	0.249 (39)	1.5062(3)
Alex Gordon	0.55757 (2)	0.294 (30)	Clayton Kershaw	0.210(2)	1.3641 (4)
Starlin Castro	0.55172 (5)	0.283(53)	Mat Latos	0.230 (9)	1.2680(5)
Mike Trout	0.48024 (N/A)	0.326(4)	Justin Verlander	0.237(4)	1.2301(6)

According to this statistic, Miguel Cabrera would be the MLB batting champion in this set of data according to our metric. Interestingly enough, Cabrera won the American League triple crown which is given to a player who leads their league in batting average, runs batted in, and home runs. Also, Mike Trout is included because he was a front runner for the MVP but finished runner-up to Miguel Cabrera. Trout was not in the top five in *wba* but we should note that this metric determines the top hitter and pitcher, not the best all around

player. This is because we do not consider the game as a whole, which is the reason Trout was considered for the MVP. It is interesting that in *wpe*, the top 5 pitchers are all starting pitchers because they face the most batter. This gives confirmation that the metric, *wpe*, could be used to determine the best pitcher since normally the Cy Young award is given to a starter.

## 6 Future Studies

We need to prove that yearly baseball data always produces an irreducible matrix  $\mathbf{C}$  by showing that the adjacency matrix corresponding to batter-pitcher interactions is strongly connected. Next, consider making a small adjustment to create a weighted on base percentage by changing  $\mathbf{C}$  to be

$$\mathbf{C} = \begin{bmatrix} \mathbf{0} & \mathbf{J} \cdot \mathbf{OB} \\ \mathbf{N} \cdot (\mathbf{PA} - \mathbf{OB})^T & \mathbf{0} \end{bmatrix}$$

where,  $\mathbf{J} = \begin{bmatrix} \frac{1}{pa_1} & 0 & \dots & 0 \\ 0 & \frac{1}{pa_2} & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \dots & \frac{1}{pa_{N_b}} \end{bmatrix}$ .

We also may consider slugging percentage and relate *wpe* to the slugging percentage for a batter versus a pitcher. This would require looking at the amount of extra base hits a pitcher gives up and more specifically, the amount of doubles, triples, and homeruns. The data would determine how difficult it is to achieve a certain extra base hit off a pitcher.

The metric might also be used to predict the *wba* or *wpe* for a player who changes teams. An example is Mike Napoli who batted 0.238 in 2010 with the Texas Rangers. However, he was in the top 30 players in *wba*. He was then traded to the Texas Rangers where his batting average was 0.330 in the following year. Was this a result of facing worse pitching with Texas' schedule or was this just by chance? Finally, are there other sports where this implicitly define statistic can be used?

## References

- [1] Retrosheet. (n.d.). Retrieved from <http://www.retrosheet.org/>
- [2] Statistics. (2012, June 6). Retrieved from <http://mlb.mlb.com/stats/>.
- [3] JE and DM. 2011. *An Implicitly Defined Baseball Statistic*. (unpublished notes).