

THE GREAT RBI PROPHECY

RILEY GOUGH

ABSTRACT. This paper will include a discussion of the concepts of formulating models for baseball statistics. Ideas on a new “Runs Batted In” model will be explored along with an analysis of Bill James’ “Runs Created” formula. The necessity of the new statistic will be emphasized.

(1) Introduction

Throughout the entire world of sports, the most difficult skill to master is, arguably, to hit a baseball. Ted Williams, a Hall of Fame left-fielder for the Boston Red Sox (1939-1942, 1946-1960), once said, “Baseball is the only field of endeavor where a man can succeed three times out of ten and be considered a good performer.” [?] Since the dawn of baseball over a century ago, players of the game have pushed themselves to the limit so as to master this most daunting of tasks. Along with the birth of baseball came the birth of fans and the fans’ unceasing passion to compare and criticize players. Over time, statistics have become the means of choice for baseball coaches and fans alike to evaluate player abilities. The first true baseball statistician, Bill James, in his series of books entitled *The Baseball Abstract*, revolutionized the way baseball statistics were formulated with his Runs Created statistic [4]. This Runs Created formula is 95% accurate in predicting runs scored by a player in a given season. Inspired by James’ formula, this paper will try to create another formula to predict, with 95% accuracy, how many Runs Batted In (RBIs) a player will hit in a given season.

(2) RBI Synopsis

In order to predict RBIs, one must know where RBIs come from. By the rules of baseball, an RBI is given to a batter for each run scored as the result of a batter’s plate appearance. The batter does not receive an RBI if he hits into a double play, if a run is scored on a wild pitch or passed ball or as the result

of an error, or if the pitcher balks. The batter does receive an RBI if he is walked or hit by a pitch with the bases loaded. He will also receive an RBI if a runner scores from his sacrifice fly or sacrifice bunt [6]. So, a list of occurrences that are possible on any plate appearance is necessary. These occurrences are: walks (BB), hit-by-pitch (HBP), sacrifice flies (SF), sacrifice hits (SH), singles (1B), doubles (2B), triples (3B), and home runs (HR). To know how many RBIs the batter will receive from any given occurrence, one must know how many runners are on base. The different situations the batter will come to the plate for are as follows: nobody on base, runner on first base, runner on second base, runner on third base, runners on first and second, runners on first and third, runners on second and third, and bases loaded. This can more easily be described in the table below where “1” means, by the rules of baseball, the possibility of hitting an RBI exists, and “0” means the possibility of hitting an RBI does not exist.

<i>RBI Possibilities</i>								
Runner On	BB	HBP	SF	SH	1B	2B	3B	HR
Nobody	0	0	0	0	0	0	0	1
1st	0	0	0	0	1	1	1	1
2nd	0	0	0	0	1	1	1	1
3rd	0	0	1	1	1	1	1	1
1st , 2nd	0	0	0	0	1	1	1	1
1st , 3rd	0	0	1	1	1	1	1	1
2nd , 3rd	0	0	1	1	1	1	1	1
Bases Loaded	1	1	1	1	1	1	1	1

(3) Linear Model

Now that the possibilities of hitting an RBI in any situation are known, we need to find out how many RBIs will be hit given the certain plate appearance occurrence and situation. In order to do this, we express RBI totals in any given season as a linear combination of the number of occurrences given the situation. In the following equation, let the superscripts represent the different situations and let β represent the number of RBIs scored given the particular occurrence during the specific situation.

$$\begin{aligned}
RBI = & \beta_1 \cdot (HR)^0 + [\beta_2 \cdot (1B) + \beta_3 \cdot (2B) + \beta_4 \cdot (3B) + \beta_5 \cdot (HR)]^1 + \\
& [\beta_6 \cdot (1B) + \beta_7 \cdot (2B) + \beta_8 \cdot (3B) + \beta_9 \cdot (HR)]^2 + \\
& [\beta_{10} \cdot (1B) + \beta_{11} \cdot (2B) + \beta_{12} \cdot (3B) + \beta_{13} \cdot (SF) + \beta_{14} \cdot (SH) + \beta_{15} \cdot (HR)]^3 + \\
& [\beta_{16} \cdot (1B) + \beta_{17} \cdot (2B) + \beta_{18} \cdot (3B) + \beta_{19} \cdot (HR)]^{1,2} + \\
& [\beta_{20} \cdot (SF) + \beta_{21} \cdot (SH) + \beta_{22} \cdot (1B) + \beta_{23} \cdot (2B) + \beta_{24} \cdot (3B) + \beta_{25} \cdot (HR)]^{1,3} + \\
& [\beta_{26} \cdot (SF) + \beta_{27} \cdot (SH) + \beta_{28} \cdot (1B) + \beta_{29} \cdot (2B) + \beta_{30} \cdot (3B) + \beta_{31} \cdot (HR)]^{2,3} + \\
& [\beta_{32} \cdot (SF) + \beta_{33} \cdot (SH) + \beta_{34} \cdot (BB) + \beta_{35} \cdot (HBP) + \beta_{36} \cdot (1B) + \beta_{37} \cdot (2B) + \\
& \beta_{38} \cdot (3B) + \beta_{39} \cdot (HR)]^{1,2,3} + \varepsilon
\end{aligned}$$

In order to find each β we must again look at the rules of baseball. For example, β_{18} must be equal to three because a home run with two runners on base will score exactly three runs which corresponds to three RBIs for the batter. Applying the rules of baseball several more times, every β can be found except for the ones that correspond to the single and double columns in the “RBI Possibilities” table above. To find the values of these numbers, we simplify the linear equation above such that the only remaining coefficients to be determined are $\beta_2, \beta_3, \beta_6, \beta_7, \beta_{10}, \beta_{11}, \beta_{16}, \beta_{17}, \beta_{22}, \beta_{23}, \beta_{28}, \beta_{29}, \beta_{36}$, and β_{37} . Now, the RBI Possibilities table above can be replaced by the table below that includes the β 's that are determined by the rules of baseball and the β 's yet to be determined.

<i>Incomplete Table of Coefficients</i>								
Runner On	BB	HBP	SF	SH	1B	2B	3B	HR
Nobody	0	0	0	0	0	0	0	1
1st	0	0	0	0	β_2	β_3	1	2
2nd	0	0	0	0	β_6	β_7	1	2
3rd	0	0	1	1	β_{10}	β_{11}	1	2
1st , 2nd	0	0	0	0	β_{16}	β_{17}	2	3
1st , 3rd	0	0	1	1	β_{22}	β_{23}	2	3
2nd , 3rd	0	0	1	1	β_{28}	β_{29}	2	3
Bases Loaded	1	1	1	1	β_{36}	β_{37}	3	4

Notice that, under the nobody-on-base situation, the only occurrence that produces RBIs is a home run, and for each home run in this situation, the batter is credited with exactly one RBI. So, the total number of RBIs a player will hit with nobody on base is exactly the number of home runs he hits with nobody on base. For the remainder of this section, we will only

concentrate on the other seven situations. As stated earlier, RBIs can be expressed as a linear combination of the different occurrences in the given situations. In order to evaluate the unknown coefficients with accuracy, we must create seven different equations such that each equation is a linear combination of the occurrences under the same situation. For example, the total number of RBIs with a runner on second base would be

$$RBI^2 = [\beta_6 \cdot (1B) + \beta_7 \cdot (2B) + (3B) + 2 \cdot (HR)]^2$$

Now that some of the coefficients are known, we can simplify the equation algebraically such that all occurrences with known coefficients are on one side of the equation and all occurrences with unknown coefficients are on the other side. Continuing the example above,

$$\begin{aligned} RBI^2 &= [\beta_6 \cdot (1B) + \beta_7 \cdot (2B) + (3B) + 2 \cdot (HR)]^2 \\ \implies [RBI - (3B) + 2 \cdot (HR)]^2 &= [\beta_6 \cdot (1B) + \beta_7 \cdot (2B)]^2 \end{aligned}$$

We do the same simplification procedure for all seven of our situations. Now, each remaining β can be easily found using linear regression on each of the seven situational formulas. All that is needed to complete the regression analysis is to find a random sample that is representative of the entire population of major league ballplayers. Since a comprehensive database was not readily available at the time of this project, the process of entering player data into a spreadsheet proved to be extremely time-consuming. As a result, the sample size only consists of the six ballplayers listed below. These players were chosen according to their position in the batting order, the winning percentage of the teams they played for, the league (American or National) in which their team resides, and the players rank in the top 200 RBI hitters list of currently active players.

1. *Carlos Delgado*

2. *Manny Ramirez*

3. *Todd Helton*

4. *Gary Sheffield*

5. *Alex Rodriguez*

6. *Craig Biggio*

Another restriction to entering player data, besides the slow process at which player data is transferred into spreadsheet format, is that situational statistics were not officially recorded before 1999. Therefore, the statistics of past baseball greats such as Babe Ruth, Ted Williams, and Willie Mays cannot be used. So, only the 1999 - 2005 statistics of the six players listed above are used in this model. In spite of all the restrictions to the data, six players statistics over seven seasons (1999 - 2005) still provides forty-two observations per situation. This is enough data such that, by the Central Limit Theorem [2], this model can still assume an approximately normal distribution. All of the data for these players were acquired from the Major League Baseball website [6]. This is a free resource that is open to the public. After entering the data into a spreadsheet, simplifying each individual situational equation like in the example above, converting the numbers over to SAS, running separate regression analysis procedures for the seven simplified situational equations, checking for outliers, and insuring variable independence, each unknown β is found. Solving the simplified situational equations for RBI^i and summing the seven equations together, the following generalized, seasonal RBI equation is deduced.

$$\begin{aligned}
\text{RBI} = & HR_0 + [.6729 \cdot (2B) + 3B + 2 \cdot HR]_1 + \\
& [.5741 \cdot (1B) + 2B + 3B + 2 \cdot HR]_2 + \\
& [1B + 2B + 3B + SF + SH + 2 \cdot HR]_3 + \\
& [.7724 \cdot (1B) + 1.562 \cdot (2B) + 2 \cdot (3B) + 3 \cdot HR]_{1,2} + \\
& [1B + SF + SH + 1.631 \cdot (2B) + 2 \cdot (3B) + 3 \cdot HR]_{1,3} + \\
& [SF + SH + 1.655 \cdot (1B) + 2 \cdot (2B + 3B) + 3 \cdot HR]_{2,3} + \\
& [SF + SH + BB + HBP + 1.727 \cdot (1B) + \\
& 2.603 \cdot (2B) + 3 \cdot (3B) + 4 \cdot HR]_{1,2,3} + \varepsilon
\end{aligned}$$

Now that the coefficients (β_i) have been found, we must understand what they mean. For example, by the above equation, one double with a runner on first base should produce exactly 0.6729 RBIs. However, this is not the case since runs are strictly integers. So, what this coefficient really stands for is the probability of batting in that runner provided the appropriate occurrence and situation. Hitting a double with a runner on first,

the batter has an approximate 67.29% chance of batting in the runner and therefore producing an RBI. A single with the bases loaded has the coefficient equal to 1.727. This means that the batter is guaranteed one RBI while he has a 72.7% chance of scoring the runner from second base and thereby producing an RBI. The coefficients of the above equation can be seen more clearly in the following table.

<i>Gough's RBI Coefficients</i>								
Runner On	BB	HBP	SF	SH	1B	2B	3B	HR
Nobody	0	0	0	0	0	0	0	1
1st	0	0	0	0	0	.6729	1	2
2nd	0	0	0	0	.5741	1	1	2
3rd	0	0	1	1	1	1	1	2
1st , 2nd	0	0	0	0	.7724	1.562	2	3
1st , 3rd	0	0	1	1	1	1.631	2	3
2nd , 3rd	0	0	1	1	1.655	2	2	3
Bases Loaded	1	1	1	1	1.727	2.603	3	4

(4) RBI Opportunities

Since there is no way to predict how many home runs, triples, doubles, etc. a player will hit in any given season, then it is impossible to predict, using this method, how many RBIs a player will hit in a given season. So, the Great RBI Prophecy has not been solved. Sometimes, however, failure breeds success. It turns out that even though the linear model does not predict RBIs, we can still use it to objectively evaluate and compare players. To have the ability to compare different players' abilities of hitting RBIs, we need to find the proportion of RBIs the player hit to the number of opportunities the player had to hit those RBIs. The importance of an RBI Opportunities statistic when comparing players is given in an example in section 5. Using the following formula, RBI opportunities can be calculated.

$$RBIOpportunities =$$

$$(\text{Runners on Base} + 1)_i \cdot (TPA - HBP - IBB)^i$$

$$i = (0), (1), (2), (3), (1, 2), (1, 3), (2, 3), (1, 2, 3)$$

Note that the abbreviation TPA stands for Total Plate Appearances. This statistic was used to define the term RBI earlier. Total Plate Appearances are counted every time the batter comes to the plate. Once the batter has an occurrence at the plate, then the plate appearance is over until after the other eight players on the team bat at which point the batter appears at the plate again. Hit-by-pitch (HBP) was mentioned earlier. It is the occurrence where the pitcher hits the batter with the ball and the batter is granted first base for free. Intentional Walks (IBB) occur when the pitcher clearly throws four straight balls out of the strike zone on purpose. Intentional walks happen for various reasons which will not be discussed here. For help with any questions about IBBs, contact the baseball fan nearest you. The question lingers on why exactly HBPs and IBBs are subtracted from TPAs. The reason is that the batter has no control over whether or not he is hit by a pitch or if he is intentionally walked. Therefore, by subtracting HBP and IBB from TPA, we do not punish the batter for occurrences in which there was an opportunity to hit RBIs, but because of circumstances out of his control, i.e. being intentionally walked, the RBIs could not be produced. The RBI Opportunities statistic is useful in that it normalizes players to the same level. For example, a player on a good team, batting in the middle of the lineup, is going to have many more opportunities to hit RBIs than, say, the leadoff hitter of a not so good team. This happens because good teams tend to have the ability to get runners on base, while not-so-good teams tend to not possess this ability. Since the number of runners on base greatly affects how many RBIs a batter can receive on a given hit, then it makes sense that good teams get more RBIs per hit than poor teams. Punishing good batters for playing on bad teams is not the goal of this statistic.

(5) **Gough's RBI Equation (GRBI)**

Now that the linear combination has been found and the RBI Opportunities statistic is in place, all that is needed is to somehow merge the two equations into one equation so that players on every team can be evaluated and compared to each other. Letting A be equal to the linear combination and B be

equal to the RBI Opportunities statistic, a new RBI statistic called, for lack of imagination, Goughs RBI (GRBI) equation is formulated by simply dividing the linear combination by the RBI Opportunities.

$$GRBI = \frac{A}{B}$$

$$\begin{aligned} A = & HR_0 + [.6729 \cdot (2B) + 3B + 2 \cdot HR]_1 + \\ & [.5741 \cdot (1B) + 2B + 3B + 2 \cdot HR]_2 + \\ & [1B + 2B + 3B + SF + SH + 2 \cdot HR]_3 + \\ & [.7724 \cdot (1B) + 1.562 \cdot (2B) + 2 \cdot (3B) + 3 \cdot HR]_{1,2} + \\ & [1B + SF + SH + 1.631 \cdot (2B) + 2 \cdot (3B) + 3 \cdot HR]_{1,3} + \\ & [SF + SH + 1.655 \cdot (1B) + 2 \cdot (2B + 3B) + 3 \cdot HR]_{2,3} + \\ & [SF + SH + BB + HBP + 1.727 \cdot (1B) + \\ & 2.603 \cdot (2B) + 3 \cdot (3B) + 4 \cdot HR]_{1,2,3} + \varepsilon \end{aligned}$$

$$\begin{aligned} B = & (\text{Runners on Base} + 1)_i \cdot (TPA - HBP - IBB)_i \\ & i = (0), (1), (2), (3), (1, 2), (1, 3), (2, 3), (1, 2, 3) \end{aligned}$$

As with James Runs Created formula of 2002 [5], a common complaint about the GRBI is its complexity, in that, under all eight different situations, one must find the particular batters HR average, 3B average, 2B average, etc. in order to objectively evaluate the player even though all of these individual averages are free and readily available on the internet. The only complexity left is the time consuming process of entering the data into a spreadsheet. One can only imagine the extensive databases that individual teams keep, on not only their own players statistics, but also players from around the league. With the right database, this statistic is as easy to calculate as Batting Average ($BA = H/AB$). Not only that, but GRBI is, arguably, the most informative statistic in all of baseball. Batting Average is by far the most widely known statistic. The only problem with Batting Average is that grand slams (home runs with bases loaded) are exactly equal to cheap, bloop singles that have no consequence in scoring any runs. Let these non-consequential hits be known as Empty Hits. James Runs Created statistic does a better job at evaluating hits, but in his statistic, potential runs are also counted. The basis of his

statistic is that a double, for example, is worth a little more than half of a run since the batter got half way back to home and if there happens to be any runners on base, those runners would either move ahead on the base paths or score. So, Runs Created is concerned with the batter not only hitting the runs home, but also the potential for the batter to score on another one of his teammates hits. GRBI is different, in that, it is only concerned with the plate appearances that matter the most, the ones that actually score runs. Empty hits and potential runs do not win ballgames. Games are won on each batters ability to actually produce the hits that are needed such that runs are batted in. Empty hits and potential runs are merely peripherals.

As stated earlier, GRBI can be used to objectively evaluate and compare players. When evaluating Craig Biggio, it can clearly be seen that Biggio, who is a leadoff hitter for the Houston Astros, has at most half as many RBI Opportunities as any of the other five players listed above. Since none of the other five players are leadoff hitters, this is a strong indication that leadoff hitters do not get nearly the number of opportunities to hit RBIs as does the rest of the batting order. The importance of RBI Opportunities is clearly seen when comparing Craig Biggio to Alex Rodriguez who is the clean-up hitter for the New York Yankees and 2005 American League Most Valuable Player (MVP). Biggio, who in 2005 had half as many RBI Opportunities as Rodriguez, actually had a better GRBI than Rodriguez under the following situations: runner on third, runners on second and third, and bases loaded. This means that the leadoff hitter for a team that is not known at all for its offense is actually more efficient at hitting RBIs than the American League MVP. But, why does Rodriguez, who won the American League MVP largely due to the phenomenal number of RBIs he hit, get so much more credit than Biggio, who does not hit a large amount of RBIs, when Biggio is actually more efficient at hitting RBIs than Rodriguez in almost half of the situations? No one should dare assume that Rodriguez is merely an average player since Biggios GRBI is higher. Even if Rodriguez is the clean-up hitter for the Yankees, hitting 100+ RBIs in a season is still a fantastic feat. But, using GRBI, the world of baseball comes ever closer to figuring out the questions that all General Managers of every ball club ask: is this player a good hitter that happens to play on an average team? and is this player

just an average hitter who happens to play for a good team?

One problem with GRBI is, however, that the linear combination used in the numerator is only 90% accurate. Of course, since there is no other statistic to relate it to, the RBI Opportunities statistic has no way of judging its accuracy. So, the RBI Opportunities statistic is left for the reader to analyze. But, when entering another players statistics, besides the players used in the model, the linear combination is at least 90% accurate. The goal of this project was to be 95% accurate. The cause for error is largely due to the fact that the sample size was so small. Even though there were forty-two observations for each situation, this does not completely guarantee a high level of accuracy. Not only was the sample size small, but the players used were also hand-picked. They were not selected at random. So, even though, when picking players to use in the model, the goal was to pick players that best represented the entire population of professional ballplayers, this sample might not have been the best combination of players to use. Besides, there are nine players on a baseball team, and only six players were used in the model. This fact alone leaves a sick feeling in the stomach of any baseball statistician. The reasons for the small sample size were stated earlier. In hindsight, a baseball simulation program such as Strat-O-Matic or APBA might have made a good resource for creating a personalized database in which the situational statistics could easily be recorded and used in the model. Not only would this have increased the sample size, but also the diversity and randomness of the sample population.

(6) Further Discussion

A question that is undoubtedly lingering in the readers mind is why the β s are even necessary. Shouldnt taking the total RBIs under the eight situations and dividing by RBI Opportunities result in a more accurate and objective statistic? It absolutely would. However, the β s have more than one use. Not only do they provide a number for runs scored given a certain occurrence under a specific situation, but they can also be used as a team speed index. If the sample size was increased and randomized and thereby creating an accurate coefficient for the average team, then the procedure can be run again using only those players on one particular team. Running the procedure again using the statistics for every player on one individual

team would produce different β s than that used by the larger random sample of players from all teams across Major League Baseball. The individual teams coefficients could then be compared to that of an average team so that the General Manager could evaluate whether or not his team could, per say, score easily from second base, or if his team needs more than just a single to score a runner from second. The higher the coefficients are for the team, the faster the team is as a whole. The faster the team is as a whole, the fewer home run hitters the team needs in order to score runs. In all the research leading up to this project, no objective team speed index was found anywhere. The only way for anybody to know which teams are faster than other teams is simply by gut-feeling. Any Coach or General Manager can see the importance of objectively knowing if his team is faster than the opposing team. So, GRBI can be used to objectively evaluate and compare hitters throughout Major League Baseball with 90% accuracy. This statistic is valuable because it does not include any empty hits or potential runs all while putting all hitters across baseball on a level playing field with the RBI Opportunities statistic. Although it does not solve the Great RBI Prophecy, GRBI can still be a useful tool for General Managers, Coaches, players, and fans throughout baseball nation.

(7) REFERENCES

- [1] Jim Albert, *Curveball*, 2002.
- [2] Hogg and Tanis, *Probability and Statistical Inference*, ver. 7, 2006, Pearson Education, London, p. 265.
- [3] *Brainy Quote*, http://www.brainyquote.com/quotes/authors/t/ted_williams.html
- [4] Bill James, *The Baseball Abstract*, 1977.
- [5] Bill James, *The Baseball Abstract*, 2002.
- [6] Major League Baseball, <http://www.mlb.com>.
- [7] Statistical Association of Baseball Research, <http://www.sabr.org>.