Journal of Quantitative Analysis in Sports

Volume 2, *Issue* 3 2006 *Article* 4

Who Controls the Plate? Isolating the Pitcher/Batter Subgame

Benjamin Alamar Ph.D.[*] Jeff Ma[†]
Gabriel M. Desjardins[‡] Lucas Ruprecht[**]

[*]University of California, San Francisco, quantsports@gmail.com
[†]Protrade Sports Inc, jma@protrade.com
[‡]Averna Corp Inc, gabriel.desjardins@gmail.com
[**]Protrade Sports Inc, lruprecht@protrade.com

Who Controls the Plate? Isolating the Pitcher/Batter Subgame

Benjamin Alamar Ph.D., Jeff Ma, Gabriel M. Desjardins, and Lucas Ruprecht

Abstract

This paper combines an estimated expected run value equation with a probability model on the outcome of batted balls to isolate the game within a game. Using linear regression we were then able to determine the percentage of the outcome of an at bat that is controlled by a pitcher and the percentage that is controlled by the batter.

KEYWORDS: expected runs, baseball, player valuation

1. Introduction

Baseball researchers have long used the concept of Net Expected Run Value (NERV) to analyze a player's performance and manager's strategic decisions (Cook 1966, Lindsey 1963). NERV, typically represented in a 3X8 table, is the run expectation for the remainder of a half inning given the number of outs and the configuration of men on base. The value of a given plate appearance, for example, can be calculated as the change in NERV caused by the plate appearance plus any runs scored on the play. In previous work, the total value of the change in NERV has been assigned to the batter or the pitcher (Lindsey 1963) and then the total value of all of the batters plate appearances or pitcher's batters faced was equal to their total value. For example, with no runners on base and no outs, the expected run value is approximately 0.51. If the batter in this situation hits a homerun, he scores one run and the expected run value, when the next batter comes to the plate is still 0.51. In this case, the NERV is simply 1 because there has been no change in the expected run value and the batter has scored 1 run. In past analysis, researchers have focused either on batters or pitchers and would have credited the focus of their analysis with 1 point of NERV.

Many plate appearances involve only the batter and the pitcher. Strikeouts, homeruns and walks all are principally affected only by the batter and pitcher. Other plays such as any ball put into play, involves multiple defensive players and runners if there are men on base. The change in the run expectancy then is not due entirely to the outcome of the duel between the pitcher and the batter, but rather, that duel is a subgame within the game. For example, a batter may hit a line drive that would ordinarily result in a single; in this case, it would be fair to say that the batter got the better of the pitcher. If however, a fielder makes a great defensive play on the batted ball and creates an out instead of a many on first, the fielder deserves credit for the change in NERV the results, not the pitcher who lost the subgame against the batter. The quantitative problem then becomes splitting the change in NERV in such a way that batters who put a ball in play that is hard to field gets a positive credit, even if a defensive player makes an unlikely play on the ball and the change in NERV for the entire play is negative.

Additionally, as past analysis has used NERV to focus simply on pitchers or batters, the results of the analysis are not comparative across pitchers and batters. There is no way, for example, to determine using a NERV analysis whether a great pitcher or a great batter has more total effect on a team over the course of a season. Splitting NERV between batters and pitchers objectively, allows for the direct comparison of total NERV between pitchers and batters to determine which will have a greater impact on a team.

We have developed a system that estimates NERV while controlling for many more game states than previously considered. For example, different parks have different dimensions that change the run expectancy. Coors Field in Colorado, for example, is known as a hitter's park. It is therefore important to adjust NERV for the park where the game being analyzed takes place.

The system then isolates the pitcher/batter subgame by estimating the expected outcome of any ball in play given the hit type, trajectory, distance and speed of the hit. Using the expected outcome of the ball in play, the effect that the hitter and pitcher have on that expected outcome is estimated using measures of skill for the players involved, independent of other defensive players. Finally, the expected change NERV change is calculated and split between the batter and pitcher according to our estimated split (batter 62% and pitcher 38%). The total for every plate appearance in 2004 was calculated to determine the best pitchers and hitters in MLB in 2004.

2. Estimating NERV

Previous work that has utilized NERV has used tables that calculate the expectancy based upon only the number of outs and the configuration of men on base. There are, however, other variables that can affect the run expectancy. In order to control for more game states, we estimated an equation that would predict NERV based upon not only outs and men on base, but also a host of other game states. These include League the game is played in, whether the batter is at home or away, inning the game is in (only the 1st, 9th and extra inning binary variables were found to be statistically significant), batter's position in the lineup, runs scored by the team at bat, whether the pitcher and batter are of opposite handedness, whether the batter is a pinch hitter, the pitch count and a park variable.[1]

We used play-by-play data for the Major Leagues from 2001-2003[2] from Stats Inc. and calculated the variable ForwardRuns for each play. ForwardRuns is defined as the total runs scored in the rest of the inning. ForwardRuns was used as the dependent variable in a series of regressions to determine the best fit. The residuals for the regressions exhibited heteroskedasticity[3] so weighted least

[1] To construct the park variable, we regressed runs scored on a constant term and a binary for every park except one, which was used as the reference park. The park variable was then calculated as the constant term plus the estimated coefficient from this regression, if the estimated coefficient was statistically significant.

[2] Intentional walks are a decision made by the manager, and not the pitcher on the mound, therefore all intentional walks were removed from the data set for all of the analysis in the paper.

[3] The magnitude of the residuals appeared to increase with outs. Using the reciprocal of outs + 1 corrected for this problem.

DOI: 10.2202/1559-0410.1012

squares was used. The reciprocal of the number of outs + 1 was used as the weight and this corrected for the heteroskedasticity.

Using the weighted R^2 as a measure of fitness, we experimented with many functional forms of the response variable. The use of a logarithmic transformation significantly improved the fit of the model to the data. The functional form that provided the best fit of the functional forms tested was the natural log of (ForwardRuns +9) as the dependent variable with the independent variables described above. The results are consistent with intuition (i.e. the signs of coefficients are reasonable) and exhibit a good fit for the data (Table 1). These estimated parameters can then be used to calculate the expected run value from the beginning of any at bat.

Table 1. Weighted Least Squares Regression of Log(ForwardRuns+9)[4] (Weighted by 1/(outs+1)

Variable	Estimate	Prob.
Outs	-0.0198	<0.001
Runneron 1st/(Outs+1)	0.0407	<0.001
Runner on 2nd/(Outs+1)	0.0619	<0.001
Runner on 3rd/(Outs+1)	0.0878	<0.001
American League	0.0003	<0.001
Ninth Inning	-0.0104	<0.001
Extra Innings	-0.0209	<0.001
Lineup Position	-0.0018	<0.001
Park	0.1126	<0.001
Constant	2.2597	<0.001
N	593,201	
R^2	0.987	

3. Isolating Pitcher/Batter SubGame

In order to isolate the Pitcher/Batter sub-game, we divided plate appearances into two groups: those where the ball was batted into the field of play, and those where the ball was not handled by a fielder, that is, walks, strikeouts, home runs and batters hit by pitch.

[4] The control variables Home Team, 1st Inning, Pitch Count, Pinch Hitter, Runs Scored and Pitcher/Batter handedness were found to have statistically significant effects, but the magnitude of these effects was very small and had no effect on any of the proceeding analysis. The finding of statistical significance was likely due the large data set employed. We have therefore removed these control variables from the analysis.

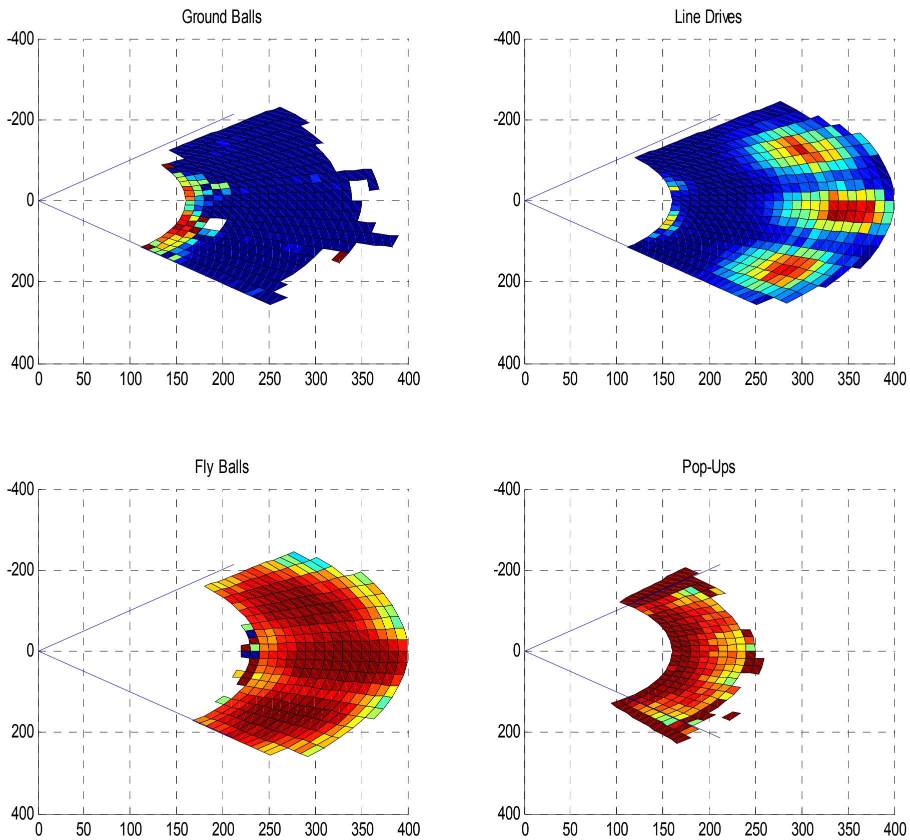

Figure 1. The expected outcome of a batted ball is dependent upon the type of hit (Ground balls, Line Drives, Fly Balls and Pop-ups), the direction the ball was hit and how hard the ball was hit. Balls hit to blue areas tend to result in base hits while balls hit to red areas tend to result in outs. The source data for Figure 1 is available as an additional file.

In the case of balls batted onto the field of play, consider the simplest case of a batter hitting with the bases empty. His plate appearance can result in five possible outcomes: he can reach first, second or third base, score a run, or he can make an out. In order to determine their expected outcome, batted balls were divided into groups based on the type of hit (fly ball, line drive, ground ball, pop up, bunt). For each hit type, they were further subdivided by the location they were hit to on the field. The expected outcome of each sub-group was then estimated empirically from past data. The probability of the batter recording an

DOI: 10.2202/1559-0410.1012

out based on hit location and type for balls hit to the outfield is shown in Figure 1 (red is highest probability, blue is lowest).

Qualitatively, the results are intuitive: fly balls and pop-ups are very likely to be caught, unless they are hit into a region in the outfield that is distant from the fielders, such as to the outfield fences or into shallow center field. Line drives, on the other hand, are likely to result in an out only if they are batted directly at an outfielder. Ground balls do not result in outs because they have already been batted cleanly by the time they reach a fielder.

Plate appearance expectation charts were generated for all baserunner configurations and out situations (e.g. bases loaded, 2 outs). In addition, similar expectation charts were created for the expected base advancement by runners in each situation. Then for a given hit type and location, we may determine the expected game state after the completion of the play, and the change in NERV that results. This calculation provides the NERV change associated with the outcome of a batter/pitcher sub-game.

We then compute the expected change in NERV based upon the expected outcome of the batted ball. This calculation provides the total NERV change associated with the batter/pitcher subgame.

For plays that involved only the batter and the pitcher, the total NERV change from the result of the play was calculated.

4. Control of the Plate

The next step was to determine the proper split of the NERV change between the pitcher and the batter that resulted from their subgame. In order to properly determine the effect a pitcher has on the NERV change vs. the batter's effect on the NERV change, we regressed the NERV change from all of the plays on relevant statistics for the batter (strikeouts per plate appearance and homeruns per plate appearance) and pitcher (strikeouts per homerun allowed and outs per base allowed) involved. These rate statistics were normalized to have a mean of one and standard deviation of one. The variables were normalized so that the analysis of the estimated coefficients is more straightforward. Also, those statistics that were expected to have a negative sign where entered into the regression in the negative so that the estimated coefficient would have a positive sign. The variables used in the regression are all normalized so that the sum of the pitching statistics for the average pitcher is equal to the sum of the statistics for the average hitter. This calculation will provide the importance of the hitter's statistics relative to the pitcher's statistics. The estimated coefficients determine the likely weight that each of these different statistics have on the likely outcome of a plate appearance, therefore the percentages that are calculated below represent the percentage effect that the pitcher (or batter) has on the outcome relative to the batter (or pitcher).

The resulting coefficients from the rate statistics (Table 2) can be used to calculate the total percentage of the expected NERV change from a given at bat is attributable an average pitcher by summing the pitcher's rate coefficients and dividing by the sum of the pitcher's and batter's rate coefficients. The average batter's percentage is then equal to the sum of the batter's estimated rate coefficients divided by the sum of the pitcher and batter's rate coefficients.

Table 2. Weighted Least Squares Regression of Expected NERV Change (Weighted by Expected NERV change)

Variable	Coefficient	Prob.
Strikeouts per Plate Appearance (-)	0.3684	<0.001
Homeruns Per Plate Appearance	0.5069	<0.001
Strikeouts per Homerun Allowed (-)	0.3189	<0.001
Outs per Base Allowed (-)	0.2153	<0.001
Constant	1.8242	<0.001
N	593,201	
R-squared	0.366	

The results of this calculation show that in an at bat between an average batter and an average pitcher, the batter should accrue 62% of the resulting expected NERV change and the pitcher should accrue 38% of the expected NERV change.[5] Intuitively this may be true because the pitcher has significantly more time to plan his action than the batter, who must assess a pitch in a split second. As the batter seems to have a more difficult task, the skill of the batter has a

Table 3a. Top 10 Batters in 2004

Rank	First Name	Last Name	Score
1	Barry	Bonds	62.089
2	Scott	Rolen	29.105
3	Manny	Ramirez	27.363
4	Todd	Helton	25.418
5	Bobby	Abreu	25.366
6	Albert	Pujols	24.052
7	Jim	Edmonds	23.254
8	Aramis	Ramirez	21.181
9	Adrian	Beltre	20.843
10	Mark	Teixeira	18.191

Table 3b. Top 10 Pitchers in 2004

Rank	First Name	Last Name	Score
1	Randy	Johnson	30.659
2	Johan	Santana	27.257
3	Curt	Schilling	27.208
4	Brad	Radke	25.749
5	Ben	Sheets	23.017
6	Jason	Schmidt	22.444
7	Carlos	Zambrano	21.185
8	Doug	Davis	20.285
9	Freddy	Garcia	20.280
10	Roy	Oswalt	20.019

DOI: 10.2202/1559-0410.1012

greater effect on the outcome and thus receives more credit for the outcome of the plate appearance.

We then used data from every at bat for the 2004 Major League Baseball Season to calculate the expected NERV change, and assign the resulting points to the batter and pitcher associated with the at bat. We then summed these play by play values for each batter and pitcher across the entire season to determine the best batters and pitchers in MLB in 2004. Table 3a shows the top ten batters in the season and table 3b shows the top ten pitchers.

Tables 3a and 3b show that, except for Barry Bonds, the best pitchers have approximately the same impact on their team over the course of a season as the best hitters. If NERV had not been split according to the 62%/38% estimated in this paper, but rather using the arbitrary 50%/50% split used in standard analysis, the scores of the best pitchers would have been relatively higher and the scores of the best batters would have been relatively lower. The conclusion from that analysis would be that the best pitchers are relatively more important than the best batters. It is therefore important to make the estimated split in order to more accurately compare batters with pitchers.

5. Conclusion

Careful statistical analysis in an expected points framework shows that, given an average pitcher and an average hitter, the hitter has more control over the outcome than the pitcher. While this result may be unsurprising, this is the first rigorous attempt at quantifying the amount of control that each player in the subgame has. We have also presented a significant advancement in the concept of evaluating players and game situations based on NERV. Our NERV equation controls for more variables than previous NERV tables allow which increases the comparability of the results. For example, if two hitters have the same NERV total based on the classic NERV table, the results are not comparable due to park or league effects. If one of these hitters has a home park that is friendlier to hitters than the other, then his NERV total will be inflated. Our NERV equation accounts for park effects and therefore makes the results more comparable across players from different teams. Our NERV equation is a framework for more rigorous study on such topics such as the effects of pitch count, when to attempt to steal bases and many other strategy and personnel issues.

---

[5] Additional rate statistics were used to test the results. Using statistics such as strikeout to walk ratio, walks per plate appearance and outs per plate appearance the batter share of the resulting change in NERV was found to be as low as 59% and as high as 65%.

References

Cook, E., Percentage Baseball (2nd ed.), (1966) Cambridge, MA: MIT Press.

Lindsey, G. "An Investigation of Strategies in Baseball," Operations Research, (1963) 11, 447-501.

DOI: 10.2202/1559-0410.1012