By the Numbers

Volume 24, Number 2 The Newsletter of the SABR Statistical Analysis Committee May, 2014

Review

Review: "The Sabermetric Revolution"

Charlie Pavitt

This new book about analytics in baseball, authored by a sabermetrics practitioner and an academic economist, is an intelligent look at how major league baseball teams have adapted to the sabermetric era. Although the book reveals nothing about what "secret" advances teams may have adopted, readers will still find it worthwhile.

Ben Baumer and Andrew Zimbalist have a new book, "The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball".

I wish to begin by examining the title. Its first half title may imply to some of us that the book is a discussion of the growth of our knowledge about baseball as achieved through statistical means. The second half of the title, however, is a more accurate reflection of the book's actual content.

In particular, the sense of the term "analytics" intended here is consistent with the way in which it is used in the business community: the use of data and statistics to inform and guide organizational decision making. In this case, the organizations in question are major league baseball teams.

As such, the book draws comparison with Ben Alamar's book reviewed by Phil Birnbaum in the previous issue of *By the Numbers.* Like Alamar, Baumer is an academic statistician who has worked for major league teams. Zimbalist is one of the premier academic sports economists, and the approach taken here is fundamentally different than one that an amateur such as I would take if I were writing a book on the "sabermetric revolution."

Not that this is in any way bad. In fact, I found the book to be intelligent throughout and (for me) quite informative about the application of statistical baseball analysis in the baseball industry.

The central thesis of the book is that the popularity of Michael Lewis's book *Moneyball* was instrumental in motivating the growth of that application. Chapter 2, entitled "The Growth and Application of Baseball Analytics Today," illustrates that growth, and will be useful to any interested person (such as myself) happy to be on the outside but curious about looking in.

That chapter, however, comes after one called "Revisiting *Moneyball*" which argues in detail about how wrong Lewis was in attributing the A's 2002 success to the fruits of Paul DePodesta's statistical analysis. I agree with Baumer and Zimbalist 100 percent here, and have long felt that this success had far less to do with Lewis's darlings Scott Hatteberg and Chad Bradford, reasonably productive support players as they were, as much as to having Tim Hudson, Mark Mulder, and Barry Zito all at their peaks anchoring the starting rotation. The authors also draw attention to the then star-level position players Eric Chavez and Miguel Tejada, the latter in particular anything but Lewis's prototypical *Moneyball* player with his refusal take a walk.

Further, if Lewis was truly accurate in reporting that Billy Beane ignored scouting information during the 2002 amateur draft, a decade's hindsight demonstrates this to have been a monumentally incompetent style of decision making that Beane has since outgrown. The authors point out that the only two draftees who have had truly thriving careers, Nick Swisher and Joe Blanton, were obvious early-round choices through conventional scouting information.

In this issue

The previous issue of this publication was February, 2014 (Volume 24, Number 1).

The take-home message to me is how ironic it is that Lewis's fundamentally flawed treatment so influenced what I trust readers of this review would consider a significant improvement in how most ball teams operate.

Chapter 3 is a thoughtful if very thumbnail survey of currently-applied measures of offense, with its strongest part a nice demonstration of why batting average is so inconsistent from year to year. Chapter 4 does the same for pitching (emphasizing DIPS and its implications), fielding (the very disappointing failure of play-by-play data to lead to a trustworthy measure of individual player performance), overall player value (the reasons why various wins above replacement measures cannot be trusted) and situational/strategy issues (Baumer/Zimbalist seem to think that Trent McCotter's work has demonstrated that hitting streakiness exists, but I remain unconvinced given the wealth of very good and contradictory work).

It is not, however, at all clear from these chapters the extent to which "conventional" sabermetric measures (if one can use that term for exemplars such as OPS, DIPS-inspired measures of pitching effectiveness, the various zone rating-based fielding indices) are actually being used by major league organizations. This is no fault of the authors, as teams are unwilling to divulge their methods for fear of losing competitive advantage.

The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball

By Benjamin Baumer and Andrew Zimbalist

University of Pennsylvania Press, 240 pages, $26.50 (US), ISBN 0812245725

amazon.com page: http://tinyurl.com/lzphnvb

They respond to this unwillingness in Chapter 7, where they attempt to estimate the extent to which teams appear to be applying sabermetric principles, no matter where they got them from, in their on-field play. First, they computed the following indices for every team over the years 1985 through 2012:

1. The ratio of OBA to BA, to reveal the extent to which walks contribute to number of baserunners;

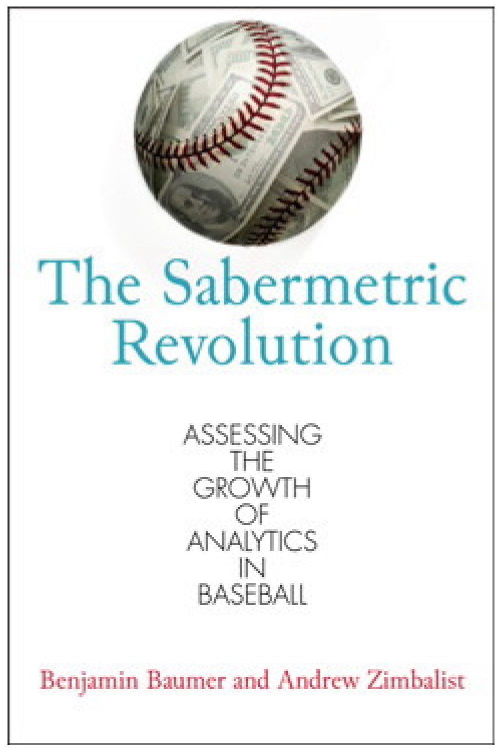
2. The proportion of sacrifice bunts relative to league average;

3. Run value of stolen base attempts, computed as (.18 (successful steals) - .32 (caught stealing));

4. The ratio of Defensive Efficiency Rating to fielding average, to gauge the extent to which range is valued over surehandedness;

5. The inverse of the Fielding Independent Pitching index, to estimate reliance on DIPS factors rather than ERA;

6. Isolated power divided by slugging average, to measure understanding of the importance of extra bases.



Regressing team winning percentage on these indices yielded estimates of the relative importance of each of the principles underlying each. The fielding range index ended up the strongest predictor (regression coefficient of .433), followed by the DIPS (.290) and walk emphasis (.237) measures. Extra base emphasis (.037) and the knowledge to limit sacrifices (.002) and poorly-chosen stolen base attempts (.001) brought up the rear.

Weighing team performance by these coefficients allowed an estimate of team "sabermetric intensity," and with this estimate Baumer and Zimbalist looked for teams that played as if they were guided by the underlying principles. Some were known to take these principles seriously: The *Moneyball* era A's, Epstein's Red Sox, Yankee teams since the late 1990s, recent Rays seasons. Others were not; the Braves are one of the few teams that have refused to formally adopt analytics in their front-office, but team design in the 1990s revealed an intuitive feel for the importance of DIPS (Maddux/Glavine/Smoltz) and fielding range.

In a final analysis, after the impact of payroll had been removed, sabermetric intensity scores accounted for 36.7% of the residual variance in team winning percentage. The take-home message: your team will win more if it takes sabermetric principles seriously.

I have not said anything about Chapters 5 and 6, because they exist outside the narrative flow of the rest of the book and read more like appendices. Chapter 5 is a short but sweet summary of the extent to which analytics has invaded NFL and NBA front offices. Chapter 6 is titled "Analytics and the Business of Baseball," but concentrates on just one of the many aspects of this topic, competitive balance. This is one of Zimbalist's central research areas, and the chapter summarizes the gist of his work on this topic. However, the research literature on competitive balance in baseball is both huge and nuanced, covering important topics left unmentioned here. For example, there is significant work on the impact of the 2000 Blue Ribbon Panel report, which decried the loss of competitive balance without including any good analysis in support of its existence or deleterious effects. The failure to consider any research other than Zimbalist's own makes this the weakest chapter in the book.

I end by repeating the descriptor for this book I used above: intelligent. The secrecy shrouding specific team application of analytics limits what the authors can accomplish, but what they have done is consistently insightful and interesting. I recommend a read to those among us curious about what can be inferred about the adoption of sabermetric principles in the game.

Charlie Pavitt, chazzq@udel.edu ♦

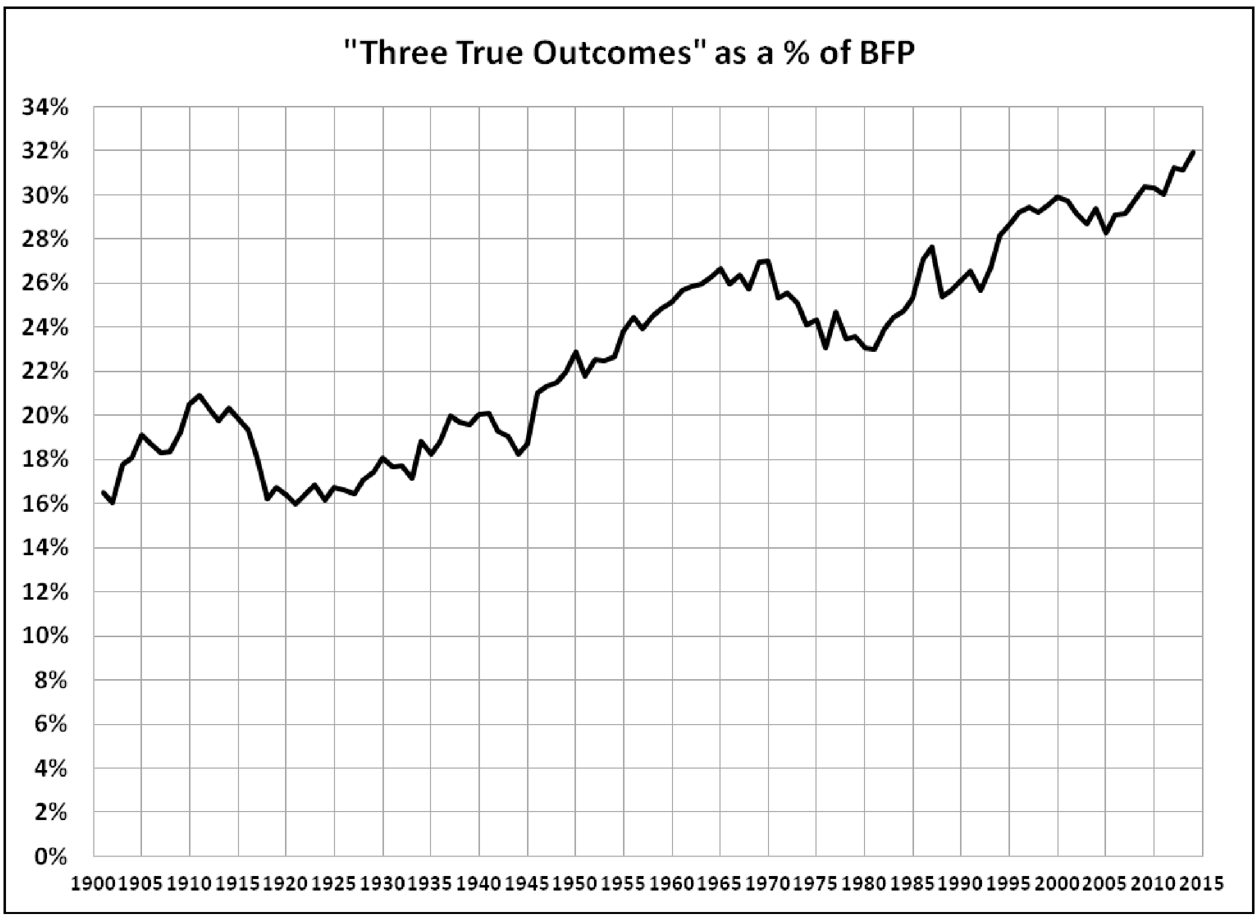
The Single Most Important Change in Baseball?

Donald A. Coffin

What's the most important difference in baseball these days, as seen in its statistics? Many would guess the increase in home runs, or power in general. But it's actually something else, the author argues.

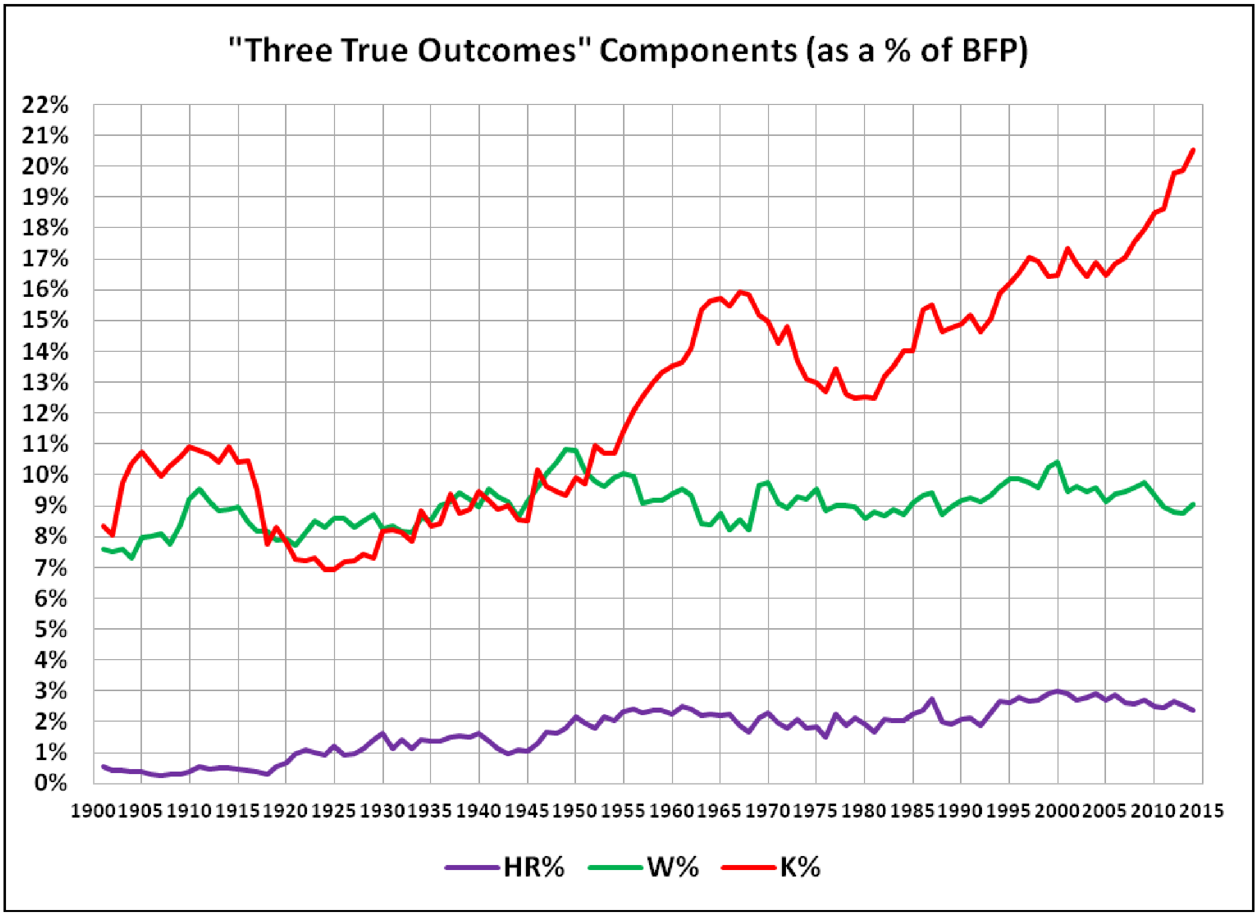
I will argue here for the single, most important change in Major League Baseball from historical times to today. And it's not the increase in hitting home runs (or, more generally, hitting for power).

One way of looking at events in MLB that has become more common over the past few years has been to look at the "Three True Outcomes"-- home runs, walks (here, including hit-by-pitch), and strikeouts -- most commonly measured as a percentage of total plate appearances (or batters-faced-pitchers, BFP). So let's start with the total TTO, as shown in the following chart.



In general, what has happened is quite clear--beginning about 1917, the percentage of plate appearances ending in one of the "Three True Outcomes" has nearly doubled, from about 17 percent to (so far in 2014) about 32 percent. Or, to put it another way, in 1917, about 83 percent of plate appearances ended with a ball in play (actually, more than that, because many home runs then were inside-the-park); now, only about 68 percent end with a ball in play.

But we might also be interested in knowing which of the components of TTO have been responsible for this. The next chart disaggregated the TTOs into the components, HR%, W% (actually W+HBP%) and K%.

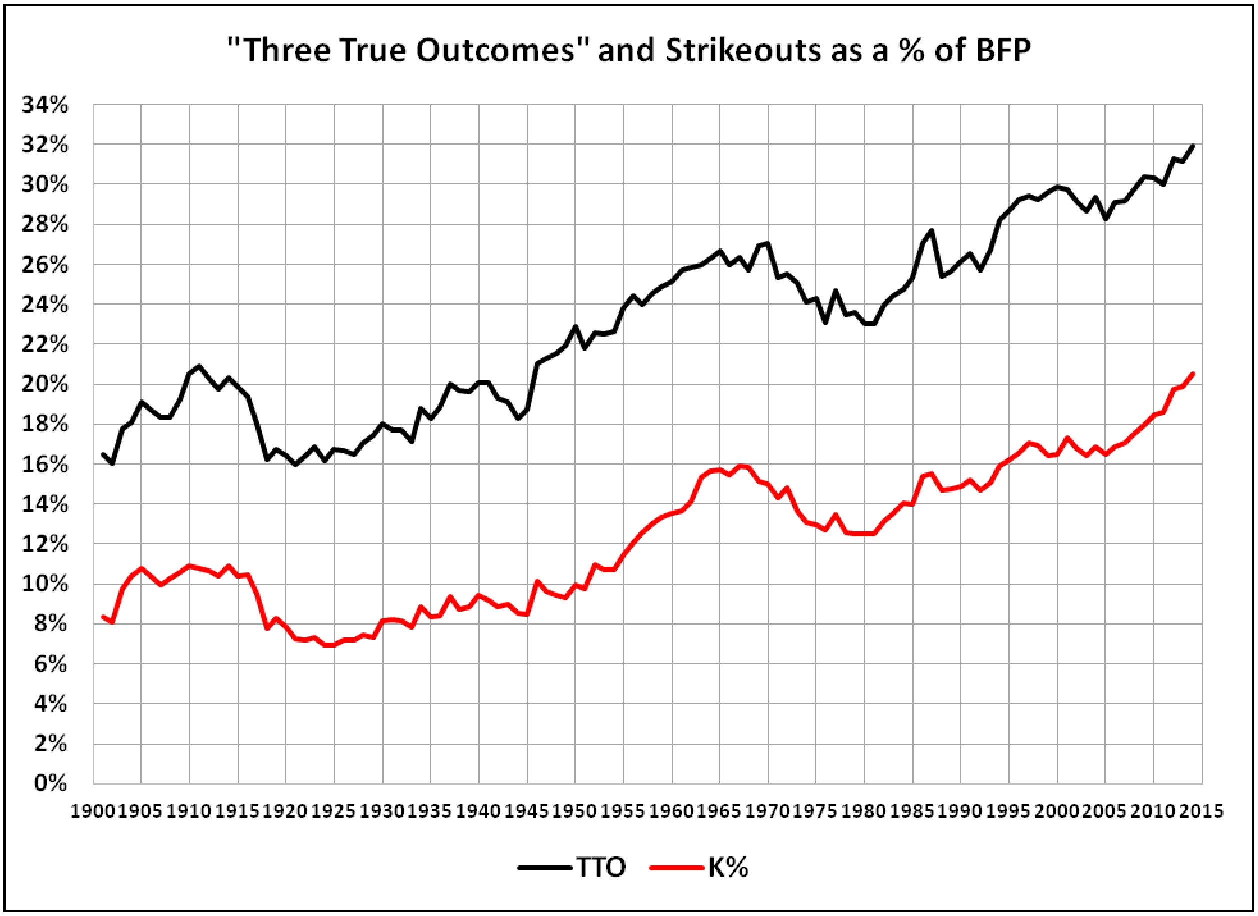


This, I think, makes the case fairly clearly. Home runs as a percentage of plate appearances did increase, quite significantly, between 1917 and 1950 (from 0.37 percent to 2.15 percent). But since 1950, there has been little change in the frequency of home runs. While the HR% did rise (peaking at 2.99 percent in 1999), that is a much smaller increase (over 49 years) than occurred between 1917 and 1950 (only 33 years). In fact, the increase from 1901 to 1950 (also a 49 year period) was from 0.54 percent to 2.15 percent, a 1.61 percentage-point increase, roughly double the increase from 1950 to 1999. The rate at which walks occur has also been fairly stable, in this case peaking in 1950, and having become fairly constant over the past 60 years, at between 8 and 10 percent.

What *has* changed, fairly consistently (with the exception of a 14-year period from 1967 to 1981), is that strikeouts as a percentage of plate appearances have risen--from 6.95 percent in 1925 to (so far) 20.54 percent in 2014--roughly *tripling*. Between 1925 and 1967, the strikeout rate rose from 6.95 percent to 15.93 percent. After falling to 12.50 percent in 1981, the strikeout rate continued its climb, and is now at 20.54 percent (so far in 2014, at date of writing in May).

Bill James, in response to a question at his online site[1], suggests -- to the extent that the rising strikeout rate is a consequence of a rise in the average speed of fastballs -- that hitters will adapt. That may be. But we have (essentially) 88 years of rising strikeout rates, with the only downward adjustment coming from fairly radical changes in the construction of the pitchers' mounds, accompanied by significant changes in the strike zone (as defined, if not as called).

In any event, it's quite clear what has been driving the overall incidence of "Three True Outcomes" upward--it's not home runs, it's not walks--it's strikeouts. From roughly 1-in-15 plate appearances ending in a strikeout in 1925. We're now at 1-in-5 and rising.



There has been no other change in how the game is played that even approaches this in magnitude -- with the possible exception of the number of pitchers used per game.

All data from Baseball Reference. Don Coffin, dcoffin@iun.edu . ♦

[1] http://www.billjamesonline.com/hey_bill/#42381

Extensions of the Linear Runs-To-Wins Model

John F. McDonald

Bill James' Pythagorean Projection is the most famous of the methods that predict winning percentage from runs scored and runs allowed. But there are others, including the linear model, which predicts (roughly) one extra win for each additional 10 runs. Here, the author looks at some of those models, and examines whether modified forms of the linear formula can yield more accurate estimates.

Introduction

A simple linear model to predict a team's winning percentage is:

$$WP = .500 + \beta(RS - RA)$$

where WP is winning percentage, RS is average runs scored per game and RA is average runs allowed. The hypothesis that the run differential is the decisive factor in winning games can be attributed to Bill James and to Thorn and Palmer (1984). This purpose of this paper is to test whether some simple extensions of this linear model can improve accuracy. In particular, the effects of runs scored per game and runs allowed per game may not have the same values, and run differential may exhibit a declining marginal effect on winning percentage.

A competing formula is Bill James' Pythagorean Projection. Miller (2006) provided a theoretical derivation of that formula,

$$WP = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}}$$

where, according to Bill James, γ is an exponent for which values of 1.82 or 1.83 give the best accuracy, but 2 is easiest and almost as good.

Dayaratna and Miller (2012) show that the above linear formula is a linear approximation of Pythagoras. They show that in the linear approximation, β can be estimated as

$$\beta = \frac{\gamma}{4R_{avg}}$$

Here, R_{avg} is the average runs scored per game by all teams.

Dayaratna and Miller estimated the linear equation for the years 1991 to 2011, and found single-year estimates for β varying from 0.084 to 0.126. The average of the estimates over the 21 years was of 0.100 (0.0995 for the earlier 11 years, and 0.1001 for the most recent decade). Average runs per game over the full period were 4.700 per team, so their estimate for γ is:

$$\gamma = \beta\left(4R_{avg}\right) = .100(4)(4.700) = 1.88\,.$$

So Dayaratna and Miller, following Davenport and Wollner (1999) and Jones and Tappin (2005) before them, have shown that the linear approximation to the Pythagorean formula works very well. Note that, because Total Wins = 0.10 (Total RS – Total RA), the coefficient of 0.100 means that scoring 10 more runs in a season will, on average, produce one more win. This is the generally-accepted rule of thumb. Thorn and Palmer (1984) found that runs per win can be computed as 10 times the square root of two times the average runs scored in two half innings. For average runs per game of 4.5, this equals 10.

The Pythagenpat formula, created by David Smyth and Patriot and discussed in Davenport and Woolner (1999), is an extension of the Pythagorean formula in which

$$\gamma = RPG^{0.287}$$

Here RPG is average runs per game for both teams combined (8.332 in 2013), which means that $\gamma = 1.8376$. Several other formulae have been proposed. Fein (2008) provides a tabulation and empirical comparison of all of the proposed formulae. Fein's study actually shows that the Pythagenport formula now preferred by Clay Davenport, with $\gamma = 0.45 + 1.5(\log_{10} RPG)$, provides a slightly better fit to the data compared to Pythagenpat. These two versions provide better results than eleven other formulae applied to data from 1921 to 2007, including the linear run differential model.

Consider a simple example in which Team A scores the league average 4.17 runs per game, every game. Team B also scores 4.17 runs per game, on average, but does so by scoring 4.67 runs half the games, and 3.67 runs the other half. Each team wins 50% of the time in head-to-head competition. So far, the higher standard deviation of runs scored for Team B does not matter.

However, according to Ciccolella (2005), the effects of scoring marginal runs are not the same as the effects of preventing marginal runs. In the simple example, let's keep Team B at 4.17 runs, but remove the symmetry. Specifically, suppose B scores 4.92 runs per game 40% of the time, and 3.67 runs 60% of the time. Now, team B wins only 40% of the time against Team A, despite the same average runs scored.

The distribution of runs scored is in fact not symmetric because runs scored has a lower bound of zero and a long right-hand tail. See Studeman (2005) and Miller (2006) for relevant empirical studies. Miller (2006) fits the Weibull distribution to the distribution of runs scored.

Are the Effects of Runs Scored and Runs Allowed Different?

As noted above, the purpose of this paper is to explore the linear model further, to discover whether greater accuracy can be achieved. One obvious area is "diminishing returns." In particular, one might hypothesize that, at some point, runs scored have a declining marginal effect on winning percentage. Consider the linear equation with .100 as the coefficient:

$$WP = .500 + .100(RS - RA)$$

That formula says that a team will win 55% of its games if it outscores the opponents by one-half run per game, and will win 60% of its games if it outscores the opponents by one run per game. The Pythagorean formula and its extensions are non-linear, but the marginal effects of runs scored and runs allowed are complex equations that are not transparent and intuitive.[1] One purpose of this paper is to conduct statistical tests for non-linear effects of run differential, runs scored, and runs allowed on the percentage of games won that are reasonably easy to comprehend.

Ciccolella modifies the Pythagorean formula by considering the standard deviation of runs scored. He finds that a team with a high standard deviation falls short of expected wins, while a team with a low standard deviation exceeds expected wins. Lower variation in scoring is better than higher variation.

Our first test of the basic linear run differential model will be to examine whether runs scored and runs allowed have numerically different effects on the percentage of games won (one positive, and one negative, of course). We will use all team-seasons from 1996 to 2013.

We will fit the data to this equation:

$$PCTWIN = \alpha + \beta(RS) + \gamma(RA)$$

The results for each of the 18 years are displayed in Table 1. The first column of results shows the coefficient of (RS-RA) for the basic linear model, and the second column provides the R-square for that estimated equation. The third and fourth columns list the estimated coefficients for RS and RA for the above equation, and the fifth column contains the R-square for this model. The coefficients of (RS-RA), RS, and RA are always highly statistically significant.

Note that the coefficient of (RS-RA) is greater than 0.10 nine times and less than 0.10 nine times. The average value for the 18 coefficients is 0.0991 -- that is, very close to 0.10.

[1] See Goldman (2006) for graphs that show the nonlinear effects of runs scored and runs allowed that are built into the Pythagorean formula.

The r-squared for the estimated equation is never less than 0.805. It is greater than 0.90 eight times out of eighteen. It is interesting to note that seven of those eight cases appear in the first nine years (1996-2004), with only one in the second nine years (2005-2013).

When the two variables are estimated separately, the coefficient of RA is larger (in absolute value) than the coefficient of RS in ten out of eighteen seasons. This basic result suggests a 50-50 chance that one of the coefficients is larger than the other in any given year, which is what one would expect if the effects of RS and RA, while random, have equal expected values with opposite signs.

However, a closer look at the results suggests that the effect of RA may be the larger, especially in the more recent years. The average coefficient of RS (over all eighteen years) is 0.0958, while the average coefficient of RA is higher, at 0.1022. Breaking off the last nine years, the effect is RA is larger in six of the nine, 0.1036 to 0.0954.

Multiplying by 10, we see that scoring 100 more runs produced 9.54 more wins and allowing 100 more runs resulted in 10.36 more losses. So, scoring 100 more runs, while at the same time allowing 100 more, meant losing almost 0.82 wins -- almost one game.

The larger effect of RA is even stronger for the last five years: the coefficient of RA is larger in absolute value four times out of the five. And the average coefficient of RS is 0.0974, while the average coefficient of RA is 0.1088 for 2009-2013.

Table 1 – Percentage of Games Won Models:1996-2013

	Single coefficient		Separate coefficients for RS, RA		
Year	Coefficient of RS-RA	r-squared	Coefficient of RS	Coefficient of RA	r-squared
1996	0.0907	0.889	0.0904	-0.0909	0.889
1997	0.0874	0.901	0.0889	-0.0864	0.901
1998	0.0982	0.921	0.1018	-0.0946	0.922
1999	0.0989	0.929	0.1012	-0.0962	0.930
2000	0.0922	0.883	0.0837	-0.1030	0.894
2001	0.1036	0.913	0.0941	-0.1119	0.918
2002	0.1036	0.914	0.0909	-0.1133	0.920
2003	0.1025	0.910	0.1025	-0.1025	0.910
2004	0.1097	0.904	0.1112	-0.1080	0.904
2005	0.0950	0.832	0.0977	-0.0937	0.833
2006	0.0976	0.829	0.0845	-0.1045	0.837
2007	0.0853	0.805	0.0829	-0.0877	0.806
2008	0.1041	0.851	0.1066	-0.1022	0.852
2009	0.1064	0.849	0.0987	-0.1154	0.854
2010	0.0919	0.950	0.0952	-0.0910	0.950
2011	0.1035	0.866	0.0976	-0.1111	0.871
2012	0.1124	0.898	0.1040	-0.1168	0.901
2013	0.1006	0.889	0.0916	-0.1097	0.893

A tentative conclusion is that defense has become slightly more important than offense in recent years. However, the effect is not particularly large and may not persist in the future. The probability that four out of five estimates would go in the same direction is 0.156, far larger than the 0.05 usually required for statistical significance.

Further Tests

As noted above, the standard Pythagorean formula and its extensions build in complex non-linear effects of run differential on the percentage of games won. A simple test for non-linearity is to add the square of run differential in the basic linear model:

$$PCTWIN = \alpha + \beta(RS - RA) + \delta(RS - RA)^2$$

This model was estimated for each of the years from 1996 to 2013, and the coefficient of the squared term is not statistically significant (that is, significantly different from zero) in any of the 18 seasons individually.

What about models in which RS and RA are entered as separate variables? Is the effect of either variable non-linear? The answer is – not often. The squared value of RS is statistically significant in four out of 18 years (2000, 2009, 2011, and 2013), and the squared value of RA comes up statistically significant in only one year (1996). However, the results for the square of RS for the most recent years may suggest something interesting. The coefficient of the square of RS is statistically significant for both 2011 and 2013, and almost statistically significant for 2012 (at the 92% level). In all three years the effect of RS is declining.

Table 2 contains results for four separate models, of which the last three permit the effects of RS and RA to be non-linear. Data is for 2011-2013 team-seasons.

Model 1 is the standard linear model, and the results show that runs scored has a smaller effect (0.0971) than runs allowed (-0.1122).

The next test is to examine whether either runs scored or runs allowed has a non-linear effect on winning percentage.

Model 2 allows RS to be non-linear, by adding a quadratic term. The results show that runs scored does have a declining marginal effect on winning percentage, and that this model has a higher level of explanatory power (r-squared = 0.901) than the linear model.

The square of runs allowed is added in Model 3, but its coefficient is very small and not statistically significant, and the explanatory power of the equation remains at 90.1 percent.

Finally, Model 4 shows the explanatory power of the Pythagorean formula (with no constant term included, as the formula states). The estimated coefficient is 0.9997 (virtually equal to 1.0), and the level of explanatory power is 89.5 percent. That's little better than the linear model, but not quite a high as the nonlinear models with squared terms.

We can use the regression equations to determine the win value of a marginal increase in RS and RA. The results are shown in Table 3.

Model 1 says that if a team adds 0.5 runs per game -- 4.76 instead of 4.26 -- it will achieve a winning percentage of .5490 (a rate of 88.9 wins per 162 games). That's computed as 0.5648 + 0.0971(4.76) – 0.1122(4.26).

Adding 0.5 a second time -- for a total of 1.0 extra marginal runs per game -- brings the winning percentage to .5975 (96.8 wins).

Model 2, which includes a quadratic term for RS, predicts .5437 (88.1) wins for an extra 0.5 runs, and .5699 (92.3 wins) for an extra 1.0 runs.

Comparing the two: the linear model predicts 7.9 more wins for an extra 0.5 runs, and an additional 7.9 wins for the next 0.5 runs. The non-linear model, on the other hand, predicts only 6.9 extra wins for the first 0.5 marginal runs, and an even lower 4.2 wins for the second 0.5 marginal runs.

Table 2 – Models of Winning Percentage, 2011-2013

	Model 1	Model 2	Model 3	Model 4
Constant	0.5684 (16.12)	-0.0318 (0.18)	-0.0205 (0.08)	
RS	0.0971 (16.48)	0.3758 (4.68)	0.3750 (4.58)	
RA	-0.1122 (20.05)	-0.1138 (22.33)	-0.1169 (1.47)	
RS^2		-0.0322 (3.48)	-0.0321 (3.40)	
RA^2			0.0005 (0.06)	
$RS^{1.83}/(RS^{1.83}+RA^{1.83})$				0.9997 (203.91)
R squared	0.877	0.901	0.901	0.895
Sample size	90	90	90	90

Note: Unsigned t-values in parentheses.

Table 3 – Predicted Wins For Marginal Increases in RS

RS/G	Linear in RS (Model 1)	Quadratic in RS (Model 2)	Pythagoras (Model 4)
4.26	81.0	81.2	81.0
4.76	88.9 (+7.9)	87.9 (+6.9)	89.2 (+8.2)
5.26	96.8 (+7.9)	92.2 (+4.2)	96.4 (+7.2)

Notes: Projections Assume RA is a constant 4.26 runs per game. Figures in columns 2-4 are wins per 162 games. Figures in parentheses are marginal differences compared to cell above .

The Pythagorean formula (Model 4) predicts .5506 (89.2 wins) and .5953 (96.4 wins) for these two values of runs per game above average. This formula predicts 8.2 more wins for the first marginal increase, and 7.2 for the second increase. As noted, the Pythagorean formula does have a declining effect of runs scored built in. However, the declining marginal effect is not as large as in the quadratic model for runs scored.

The Pythagenpat formula in this case has an exponent of 1.8494 and produces virtually the same predictions for wins as does the Pythagorean formula. It is not shown in Table 3.

The tests reported in this section show that, for the three most recent years, runs scored per game have a declining marginal effect on wins. Furthermore, for these three years, the declining marginal effect built into the Pythagorean formula (and its extensions) may not be large enough to capture what happened. However, the results for the 15 years prior to 2011 (1996-2010) do not indicate any non-linear effect of runs scored or runs allowed. The bottom line on all of this may be that no specific numerical formula is best for all time. It would seem that the model needs to be updated using the most recent data.

Conclusions

This paper has shown that positive marginal effect of runs scored per game tends to be smaller than the negative effect of runs allowed per game on the percentage of games won. Tests for non-linearity for each year from 1996 to 2013 show that neither run differential nor runs allowed has a non-linear effect, but that data for the three most recent years show that runs scored may have a declining marginal effect (but evidently not in prior years). These results suggest that models need to be updated using recent data.

Future research might conduct additional tests regarding the declining marginal effect of runs scored on percentage of games won – and explore possible reasons for this apparent outcome. This line of inquiry seems to be related to Ciccolella's (2005) result that the shape of a team's probability distribution of runs scored affects its percentage of games won. That probability distribution of runs scored is not symmetric, and therefore cannot be described fully by just mean and standard deviation. For example, Miller's (2006) Weibull distribution of runs scored can be used. Furthermore, will the result that scoring runs has a smaller effect than allowing runs hold up to further testing? Are pitching and defense (slightly) more important than offense?

References

- Ciccolella, Ray, The Pythagorean Projection and the Standard Deviation of Runs, By the Numbers, 15 (2005), No. 2, pp. 5-7.
- Davenport, Clay and Woolner, Keith, Revisiting the Pythagorean Theorem, Baseball Prospectus 1999, New York: Wiley (1999).
- Dayaratna, Kevin D. and Miller, Steven J., First-Order Approximations of the Pythagorean Formula, By the Numbers, 22 (2012), No. 1, pp. 15-19.
- Fein, Zach, Measuring the Accuracy of Baseball's Winning Percentage Estimators, Bleacher Report, online (Sept. 8, 2008), *http://tinyurl.com/pldzzvk* .
- Goldman, Steven, Can a Team Have too Much Pitching? in Baseball Between the Numbers, J. Keri, ed. New York: Basic Books (2006) pp. 272-291..
- Jones, M.A. and Tappin, L.A., The Pythagorean Theorem of Baseball and Alternative Models, The UMAP Journal, 26 (2005), No. 2.
- Miller, Steven J., A Derivation of James' Pythagorean Projection, By the Numbers, 16 (2006), No. 1, pp. 17-21.
- Studeman, Dave, Runs per Game, The Hardball Times (2005), June 28, *http://tinyurl.com/k74ux9c* .
- Thorn, John, and Palmer, Pete, The Hidden Game of Baseball: A Revolutionary Approach to Baseball and Statistics, New York: Dolphin (1984).

Note: this article has been modified from the original, to credit Smyth/Patriot for the "Pythagenpat" formula at the bottom of page 7.

John F. McDonald is just a retired investor living on a pension – from the University of Illinois. He thanks Phil Birnbaum for his assistance above and beyond the call of duty with the data and statistical computations. mcdonald@uic.edu ♦

Back issues

Back issues of "By the Numbers" are available at the SABR website, at http://sabr.org/research/statistical-analysis-research-committee-newsletters, and at editor Phil Birnbaum's website, www.philbirnbaum.com .

The SABR website also features back issues of "Baseball Analyst", the sabermetric publication produced by Bill James from 1981 to 1989. Those issues can be found at http://sabr.org/research/baseball-analyst-archives.

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work are all welcome.

Articles should be submitted in electronic form, preferably by e-mail. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (i.e., informal peer review).

I usually edit for spelling and grammar. If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does.

I will acknowledge all articles upon receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

Send submissions to Phil Birnbaum, at birnbaum@sympatico.ca .

"By the Numbers" mailing list

SABR members who have joined the Statistical Analysis Committee will receive e-mail notification of new issues of BTN, as well as other news concerning this publication.

The easiest way to join the committee is to visit http://members.sabr.org, click on "my SABR," then "committees and regionals," then "add new" committee. Add the Statistical Analysis Committee, and you're done. You will be informed when new issues are available for downloading from the internet.

If you would like more information, send an e-mail to Phil Birnbaum, at birnbaum@sympatico.ca. If you don't have internet access, we will send you BTN by mail; write to Phil at 88 Westpointe Cres., Ottawa, ON, K2G 5Y8.