By the Numbers

Volume 19, Number 1 The Newsletter of the SABR Statistical Analysis Committee February, 2009

Review

Academic Research: Competitive Balance

Charlie Pavitt

The author reviews an academic study showing an increase in competitive balance due to integration and the influx for foreign players.

Martin B. Schmidt, The Nonlinear Behavior of Competition: The Impact of Talent Compression on Competition, Journal of Population Economics, 2009, Volume 22, pp. 57-74

Schmidt has been a huge contributor to the academic sabermetric literature, with ten relevant publications of which I am aware, almost all relevant to competitive balance. As much as this issue has been studied, I still think that there is room for further analysis, as it is at yet unclear whether the steady improvement in balance among teams over the decades has stalled or even reversed since the mid-1990s. What is clear is that steady improvement until at least that time, which needs no further demonstration. What it does need is explanation, and Schmidt's latest makes what I think to be a critical contribution to this issue.

In this issue

In doing this work, Schmidt sits on the shoulders of the famed paleontologist and baseball fan Stephen Jay Gould, whose work I will discuss here for readers unfamiliar with it. Gould wrote a couple of essays demonstrating through the use of batting averages the steady increase in the ability of position players over the course of major league baseball history. Average BAs for regulars have drifted around during the decades. They were in the .250s during the first and second decades of the twentieth century, in the high .280s in the 1920s, and then fell to around .260 between 1940 and 1960. They dipped below that in next decade before popping back toward .260 in the 1970s. In contrast, the within-season variation has decreased steadily over that time. In the first essay, Gould showed that the difference between the five highest batting averages and the league average was in the range of about 90 points during the 19th century, 80 points the first three decades of the 20th, and 70 points since; the difference between the lowest and fifth-lowest began at about 60 or 70 points and has decreased to about 35 since. In the second essay, Gould made the same point in a more statistically trustworthy way, presenting the fact that the standard deviation in within-season batting averages has decreased from about .05 in at the beginning to about .03 at the time of the essay. In both analyses, the improvement was asymptotic, occurring quickly at the beginning but slowing down toward some yet-unknown limiting figure.

Gould's original explanation for this effect was the standardization of play; for example, teams have gotten progressively savvier in positioning their fielders, making it harder to get hits. But by the second essay he began to realize the real explanation; the overall increase in talent. We must assume that the greatest stars have been about equally good over the interim, which would follow from Gould's belief that there is a limit to ability based on the very facts of human physiology (an "outer limit of human capacity") that the best players can approach but never cross. If so, then the average player has gotten closer to the stars, and the greater availability of competent players has increased the replacement level such that .180 hitters are just too poor to play regularly anymore no matter how well they field. There is a critical implication of this conjecture. Although the number of major league teams has almost doubled since 1960, the population from which major league players are procured has far more than doubled. Not only has the population of the United States kept up with expansion, the number of players from the other American nations has exploded, along with East Asians, an ever-increasing number of Australians, and even the occasional European.

Schmidt (2009) transferred this reasoning into the context of teams. As poorer teams have as a whole have poorer players than good teams, if player strength increases mostly at the bottom end, then it will be the poorer teams that get the biggest payoff, thus increasing competitive balance. Further, as player improvement is asymptotic, so should be team improvement. Schmidt used team data from 1911-2005 and concluded from his analyses that the increase in competitive balance over that period of time was indeed asymptotic, but with a significant breakpoint of improvement in the mid 1950s, when the impact of integration on levels of talent would have been at its highest. Further, the degree of competitive balance correlates with the proportion of foreign-born players, which has been rising since 1940 from perhaps 2 percent at best to more than 25 percent. This later result implies that improvements in competitive balance are likely at least partly a result of that influx, consistent with Gould in at least the sense that a bigger player pool to chose from means better overall talent.

Charlie Pavitt, chazzq@UDel.Edu ♦

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis – that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any – I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics: confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Shelly Appleton	slappleton@sbcglobal.net	Statistics
Ben Baumer	bbaumer@nymets.com	Statistics
Chris Beauchamp	cbeaucha@asinc.ca	Statistics
Jim Box	jim.box@duke.edu	Statistics
Keith Carlson	kcsqrd@charter.net	General
Dan Evans	devans@seattlemariners.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	Han60Man@aol.com	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Bill Johnson	firebee02@hotmail.com	Statistics
Mark E. Johnson	maejohns@yahoo.com	General
David Kaplan	dkaplan@education.wisc.edu	Statistics (regression)
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
Chris Long	clong@padres.com	Statistics
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
John Stryker	john.stryker@gmail.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@iw.net	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

A Model for Estimating Run Creation

Richard Schell

There are various run estimators in current use – runs created, linear weights, base runs, etc. Here, the author comes up with a new estimator, deriving it from first principles, and tests its accuracy compared to the others.

The following describes a model for estimating team run scoring based on the idea that run scoring is directly related to the average number of men a team has on base when a batter comes to the plate. The model can be employed for other purposes, described at the end of the paper. It fits available data exceptionally well compared to other run estimators in use today and has an additional advantage, in that it is related to what happens in baseball games.

Background

The goal of run estimation is manifold. The first is to take a set of data specific to a team – familiar baseball statistics -- and use that data to estimate how many runs that team should or will score. A second use is to estimate the number of runs a player on a given team created (or will create) for his team. Related to that is the allocation of runs a team actually scored to its players in order to determine the value of that player to his team in runs (and wins, following Bill James's *Win Shares* statistic [1].

Familiar run estimators include Bill James's well-known Runs Created estimators, of which many versions exist. James introduced this formula nearly thirty years ago [2] and has continued to refine it from time to time to improve its fit to available data. James's formula for estimating runs has a simple form: $Runs = A \times B/C$. What has varied from version to version is what James uses for *A, B,* and *C*. This estimator is non-linear – that is, the equation is not a linear equation. Roughly contemporary with Runs Created is Pete Palmer's Linear Weights [3], whose very name implies that it is a linear estimator – the estimator is a linear equation relating runs to other baseball statistics. Obviously, all run estimators are either linear or non-linear. Among the first category, in addition to Linear Weights, Jim Furtado has developed an improvement called Extrapolated Runs (XR) [4], which is quite effective. David Smyth created a very elegant non-linear model, called Base Runs [5], which is similar to Runs Created in some ways, and different in a couple of very important features.

Dan Fox explains that run estimators can be differentiated in a different way: statistical and intuitive [6]. Most linear estimators, such as Linear Weights and XR, are statistical in nature. They employ linear regression to determine the coefficients, or weights, they assign to various baseball events. Non-linear estimators, such as Bill James's various Runs Created formulas, as well as BsR are intuitive in nature. These categories are useful, even if they are not pure in nature. (Both James and Smyth tweaked their formulas to fit the available data. Paul Johnson's Estimated Run Production [7], which is linear in nature, is driven by an intuitive look at how runs are scored.)

To Fox's two categories, I will add an additional one: constructive models. Constructive models start with an intuition, such as the one I described earlier – namely that run scoring is connected to the average number of men on base, but then use relationships that exist among baseball events to construct a model of how run scoring works.

Constructing Models

Models can be constructed by examining how runs are scored. To obtain David Smyth's model, start from a simple axiom, that the number of runs scored cannot exceed the number of men who reach base minus those who are out on base (via double play, being caught stealing, or out running the bases on a hit). In mathematical terms, $Runs \le Reached - OutOnBase$. Another way to state this is $Runs = f \times (Reached - OutOnBase)$, where *f* is a function that never exceeds the value 1. A third alternative is to write this as $Runs - HR = f \times (Reached - HR - OutOnBase)$, noting that home runs always score one run. What can be said about *f?* It should have the value 1 if no batter makes an out, so a good form for it is $X/(OWB + X)$, where *OWB* stands for "out while batting", including all sacrifice hits and sacrifice flies. That leads to the run estimator

$$Runs = \frac{X \times (Reached - HR - OutOnBase)}{X + OWB} + HR$$

which is the form of Base Runs. Finding *X* is a challenge, but one that can be addressed by using linear regression (carefully). This leads to something very like Base Runs, with a single exception – *Reached should* include reaching on error, a piece of data that is now available for a number of teams for a number of seasons, through Retrosheet. Note that the form of Bill James's original Runs Created can be written in the form $Runs = f \times TB$, where *X* is simply *RBS* (reaching base safely, not including reaching on error). There isn't a fundamental premise that leads to the derivation of this form. As a result, Runs Created is subject to the anomalous behavior mentioned earlier.

Another, more complex, construction starts from the observation that runners who reach base either score, are out on base, or are left at the end of an inning. In mathematical terms, $Runs = Reached - OutOnBase - LeftOnBase$. Add to this two additional assumptions related to the average number of men on base. The first assumption is driven by intuition and observation. It is obvious that a home run or triple will (on average) drive in the average number of men on base; a double will drive in a substantial percentage of that number; and a single will drive in a smaller percentage, a walk a smaller percentage still. Because home runs always drive in the batter, they should be separated. The implication of these two principles is that $Scored = V \times m + HR$, where *V* is (hopefully) a linear function and *m* stands for average men on base.

The second assumption can be motivated as follows. If every batter came up with the same average number of men on base as did the last batter, then the number of men left on base *per inning* would be the same as the average number of men on base. That isn't the case, for a variety of reasons, not the least of which is that the first man in the inning always comes to the plate with no runner on base. It is intuitive that $LeftOn = k \times m \times Innings$, where the number *k* (which could be a variable or a constant) is in the neighborhood of 1. Data indicates (and other models that will be discussed bear out) that this number can be expressed as $BOPI \times (1-z)/2$, where *BOPI* is short for "batter outs per inning" and *z* is some (fairly small, per actual team data) number.

If these assumptions are combined, and using the fact that $Innings \times BOPI = OWB$, the result is

$$[V + (1-z)/2 \times OWB] \times m = Reached - HR - OutOnBase$$

where *OWB* stands for "out while batting". It is convenient to denote the quantity $ReachedBase - HR$ by *"Occupied",* and the quantity $Scored - HR$ by "*TDI"* (for teammates driven in). Then, substituting the result for *m* into $TDI = V \times m$ yields

$$TDI = \frac{2 \times V \times (Occupied - OutOnBase)}{2 \times V + (1-z) \times OWB}$$

The problem is that *V* and *z* are both unknown. They can both be obtained using linear regression techniques (and some tricks with algebra), but that obscures what is going on to actually score runs. The variable *z* is small, and doesn't vary much, according to historical data, so the real problem is in *V*. Empirically, $V = (TB - HR)/3 + U$ where *U* includes terms in doubles, triples, stolen bases, sacrifice hits, and sacrifice flies[1].

A Third Construction

Markov models can be used as a tool to develop a better model, one that has a clearer relationship to what happens in the game. Markov models are based on *stochastic processes* that model real world activities, such as baseball games. Simply put, they explain how an inning is *likely* to progress from one batter to the next. That enables determining the average state of the game, including such numbers as average men on base, average men on first, second, and third, average scoring from any given base, and so on. Their application requires matrix

[1] David Smyth asked in an online forum whether sacrifice flies should be given special treatment in run estimation models. In an early version of the run estimation model described here, sacrifice flies were given the same treatment as home runs; in the current version, they are treated as a variable part of *V*. Both treatments yield roughtly the same "value" for sacrifice flies. Neither treatment is right or wrong.

algebra involving sub-matrices of a 25x25 matrix and row (or column) vectors. Their actual use is beyond the scope of this paper, but they can be useful in deriving certain key expressions and approximations. For a deeper discussion, see [8] and [9].

A Markov model for baseball [10] can be used to derive the relationship

$$m = \frac{2 \times (Occupied - k_1 \times TDI - k_2 \times OOB)}{OWB}$$

The numbers k_1 and k_2 are variable, but can generally be treated as constants for the purposes of modeling the process of run scoring. Markov models suggest that these numbers are roughly .8 and .875 [2]. This last relationship produces a slightly different formula for run estimation. Using $TDI = V \times m$, this version yields

$$TDI = \frac{2 \times V \times (Occupied - k_2 \times OOB)}{OWB + 2 \times k_1 \times V} \qquad \text{(Est)}$$

This latter form is identical to the prior one if k_1 and k_2 are both 1. Note the similarity between this formula and Base Runs. David Smyth did not incorporate "out on base" (equivalent to $k_2 = 0$) and had an identical term in the numerator and denominator (equivalent to $k_1 = 1$ and $A = 2 \times V$).

To finish the process requires devising a linear function, *V*, for which $TDI = V \times m$ holds. To do that, Markov models can be used to derive simple relationships among various averages of men on base – total men on base as well as average numbers of men on each base. A really simple sketch proceeds as follows. Let $V_1, V_2,$ and V_3 be linear functions that approximate the rate of driving in men from first, second, and third base respectively. Then, the simple relationship $m_2 = m - m_1 - m_3$ (where m_1 is the average number of men on first, m_2 the average number of men on second, and m_3 the average number of men on third) yields

$$TDI = V_2 \times m - (V_2 - V_1) \times m_1 + (V_3 - V_2) \times m_3$$

The Markov model also produces good approximate coefficients *a, b, c,* and *d*, such that

$$m_3 \approx a \times m - b \times m_1 + c \times (\frac{3B + d \times OOB}{BFP})$$

This enables the equation for *TDI* to be written in the form

$$TDI = U_1 \times m - U_2 \times m_1 + U_3 \times (\frac{3B + d \times OOB}{BFP})$$

Ultimately, m_1, as well as the final term in the right hand side, can be related to *m*. In this way one obtains a linear function *V* that includes only first-order terms (no constants) for which the relationship $TDI = V \times m$ holds. Using historical values to create these approximations, and making small adjustments to obtain simple coefficients, yields

$$V = .5 \times (TB + ROE - .3 \times H - 2 \times HR + .3 \times 3B + 3.1 \times SF + .7 \times SB + .1 \times (SH + BB + HBP) - .06 \times SO)$$

[2] This is equivalent to computing the value of *z* discussed earlier.

Note also that this formula, unlike others, uses reached-on-error[3], a statistic that is available for many years through Retrosheet. Observing that *m/2* is approximately $(Reached - HomeRuns) / BFP$ for most actual team data, runs scored can be roughly estimated by something of the form

$$(TB + X) \times \frac{Reached}{BFP} + (1 - \frac{2 \times Reached + (TB - 2 \times HR + X)}{BFP}) \times HR .$$

If *X* and the term multiplying *HR* are both negligible (or offset), this is very similar to the original Bill James formula for Runs Created. Runs Created works not because it was derived from first principles, but because it approximates a model that was.

Effectiveness

Is this a good estimator? The answer is quite good, as measured by first, fit to data. The following table shows the results using one hundred years of data from 1907 through 2006. The rows in the table show the average values of actual runs, estimated runs, the correlation between these, and the standard error for the period since the year in the first column. (This method is designed to remove era-specific estimation. Better results for correlation and standard error can be obtained for any specific span of time, but only at the expense of some other span of time.)

Since	Actual	Est.	Correl	Std Err
1907	691.4	691.2	.979	22.68
1920	704.1	703.6	.978	21.77
1930	700.6	700.5	.978	21.70
1940	694.4	695.2	.979	20.97
1950	697.7	698.3	.980	20.40
1960	699.1	699.1	.981	19.95
1970	708.6	709.2	.981	20.08
1980	721.5	721.0	.982	20.49
1990	747.6	746.7	.978	20.40
2000	775.5	775.3	.963	21.67

Since 1960, the estimator performs remarkably well[4]. One should not expect an estimator to do much better, because the average variance in run scoring for two teams that have identical statistics should be between twenty and twenty one as well.

A comparison of the model's fit to data against RC, XR, and BsR is quite favorable. For example, since 1960, the latest technical version of RC produces a correlation of .976 and a standard error of 22.9. Slightly modified versions[5] of XR and BsR produce correlations of .978 and .979, and standard errors of 22.1 and 21.8, respectively. Some of the advantage that the model has over RC, XR, and BsR exist because it incorporates reached on error (ROE), while the others do not. It requires a bit of manipulation, but both BsR and XR can be extended to include ROE; only the most recent version of RC is amenable to this, one with an elaborate weighting scheme. The following table compares RC, BsR, XR, and the new model since 1960.

	RC		BsR		XR		New Model
	W ROE	W/O	W ROE	W/O	W ROE	W/O	
Correl	.979	.976	.980	.978	.979	.977	.981
Stderr	21.1	22.9	21.0	21.8	21.2	22.2	19.9

[3] Adding reached on error is an important feature of this model. Modern teams reach base on error fewer than 70 times (the 2006 league average was 66), versus an average of 180 times in 1907. That difference equates to over fifty runs. Accounting explicitly for outs on base is also important, as that number varies greatly year over year; modern teams hit into more double plays, but are caught stealing fewer times than historical averages.

[4] For example, use of linear regression on the period since 1960 produces only slightly better correlation and standard error, but at the expense of estimating earlier periods well.

[5] The author tweaked coefficients in both models based on findings related to his own model; these tweaks improved both XR and BsR in terms of fit to data. The structure of each model was left intact. Such tweaking was not justified with RC, as Bill James has repeatedly tweaked his own model, unlike Furtado and Smyth.

All these methods clearly improve with the inclusion of ROE, although the new model still outperforms them.

Fit to data is only one criterion for effectiveness. This estimator also "explains" why it works. Intuitively, a relationship exists between scoring runs and the average number of men a team has on base. This model exploits that relationship, yielding good estimates for average men on base as well as run scoring. The fact that the linear function *V* can be obtained by construction (as well as by linear regression) is, I believe, a benefit. As David Smyth remarked about his own Base Runs in [3], it does a "good job of modeling the scoring process". (For Smyth, this was more important than standard error.) The men-on-base model improves on Base Runs, in that it provides an explicit connection among the terms in the estimator.

The model also explains the net values of various statistics in run scoring. Using standard mathematical techniques [11], one can derive linear functions for run scoring from the non-linear formula denoted in the foregoing by (Est). These linear functions can be derived from team-specific data, from yearly league averages, or from averages over longer periods of time. For example, using this technique on historical averages for various statistics yields linear weights of .498, .751, 1.029, 1.421, and .353 for singles, doubles, triples, home runs, and walks (as well as hit by pitch). It also yields a weight of .168 for stolen bases and -.396 for caught stealing.

Measured in terms of simplicity, this model is not nearly as elegant as James's original Runs Created, but compared with any of the more "accurate" versions of RC, or any other estimator, it holds its own. It is possible to simplify the expression for *V* and obtain a reasonable result, but at a fair sacrifice to accuracy. One potential use for this model is allocating team runs to individual players to evaluate their performance (see [111]). Using the full model, rather than an abbreviated version, improves the accuracy of this allocation as well.

Summary

The goal of this paper was to elaborate a new model for estimating runs, one that adheres to the process of run scoring that actually occurs during games. Rather than use linear regression to obtain an estimator, it uses fundamental principles. It ties its structure to other run estimators, both non-linear and linear. With indirect help from a stochastic (Markov) model, it produces a formula for run estimation that is apparently superior in fitting data to any other simple estimator available[6].

References

1. James, Bill, *Win Shares*.
2. James, Bill, *The 1979 Baseball Abstract.*
3. Palmer, Pete and Thorn, John, *The Hidden Game of Baseball*.
4. Furtado, Jim, *The 1999 Big Bad Baseball Annual*
5. Smyth, David, *Base Runs Primer*, no longer available online.
6. In *Dan Agonistes,* October 7, 2004, available online at http://danagonistes.blogspot.com/2004/10/brief-history-of-run-estimation-runs.html. This work references as an original source for the idea, *Curve Ball,* by Jim Albert and Jay Bennett, an excellent source for understanding the role of chance and stochastic processes in baseball.
7. Johnson, Paul, in *The 1985 Baseball Abstract* (James).
8. For those interested, a good starting point is Mark Pankin's tutorial at http://www.pankin.com/markov/.
9. "The Book: Playing the Percentages in Baseball" also makes extensive use of Markov models.
10. Schell, Richard, *Using Markov Models as a Tool for Run Estimation*, to be published.
11. Schell, Richard, *Linear Weights from Non-Linear Run Estimators*, to be published.

Richard Schell, rick@Onset.com ♦

[6] Full Markov models should do better, but are more complex to understand and use.

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, either by e-mail or on CD. 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).

If your submission discusses a previous BTN article, the author of that article may be asked to reply briefly in the same issue in which your letter or article appears.

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
88 Westpointe Cres., Nepean, ON, Canada, K2G 5Y8
birnbaum@sympatico.ca

Get Your Own Copy

If you're not a member of the Statistical Analysis Committee, you're probably reading a friend's copy of this issue of BTN, or perhaps you paid for a copy through the SABR office.

If that's the case, you might want to consider joining the Committee, which will get you an automatic subscription to BTN. There are no extra charges (besides the regular SABR membership fee) or obligations – just an interest in the statistical analysis of baseball.

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 (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. If you don't have internet access, we will send you BTN by mail; write to Neal at
4317 Dayton Ave. N. #201, Seattle, WA, 98103-7154.

McCracken and Wang Revisited

Pete Palmer

The author comments on two issues. First, Voros McCracken's DIPS theory suggests that pitchers have only a small amount of control over whether balls in play turn into base hits. Here, the author shows how much real-life variances in this ability contribute to ERA differences between groups of pitchers. Second, the author comments on Victor Wang's recent study on OBP and SLG ratios, running regressions to provide more data on what linear combination of on-base and slugging best predicts runs.

Voros McCracken suggested that there is no difference among pitchers for outs per balls in play. This has been analyzed and contested and McCracken backed off somewhat, saying that there is very little difference. In actuality, there is a fair difference, but not nearly as much as would be expected based on the change in earned run average. However, there is a good reason for that, because counting outs per balls in play eliminates most of the real differences between pitchers. The table below covering data for 2001-2008, a reasonably homogeneous sample, shows typical stats for pitchers grouped by earned run average. In this sample, only pitchers with at least 1/3 of the appearances as starters were considered. This is because relief pitchers have a slight advantage in ERA because all runners on base when they come in are charged to the previous pitcher, even though the current pitcher is partly responsible for them. Including all pitchers changes the results only slightly.

Statistics for pitchers with at least 1/3 starts, 2001-2008, grouped by ERA

num														
pit	era	tbf	h	2b	3b	hr	bb	hb	so	bip	outs	o/bip	oba	slg
122	2.64	36.40	7.48	1.45	.16	0.69	2.46	.30	8.08	24.37	17.59	.722	.284	.341
162	3.28	37.19	8.15	1.59	.15	0.83	2.67	.28	7.28	25.60	18.29	.714	.301	.372
253	3.76	37.88	8.71	1.76	.17	0.97	2.80	.32	6.44	26.80	19.06	.711	.315	.401
313	4.25	38.52	9.18	1.89	.19	1.06	2.97	.35	6.17	27.41	19.29	.704	.327	.422
284	4.73	39.24	9.70	2.00	.21	1.18	3.14	.37	5.76	28.15	19.63	.698	.340	.446
217	5.22	39.95	10.13	2.14	.22	1.29	3.37	.41	5.71	28.52	19.68	.690	.351	.467
627	6.41	41.38	11.06	2.33	.24	1.51	3.90	.44	5.55	29.28	19.73	.674	.375	.506

Taking the 2nd and 5th lines, the ratio of era is 5.22/3.28 or 1.59, while the ratio of outs/bip is only .714/.690 or 1.03. Taking hits/bip instead, you still get .310/.286 or 1.08. For pitchers, runs allowed are proportional to on-base times slugging. Batters, on the other hand, show runs scored as a function of on-base plus slugging (OPS). This is because a batter gets dropped in a normal lineup every 9 positions, so even a player who hit a home run every time up would add only about 6 runs per game. A pitcher who gave up a homer every time up would allow an infinite number of runs. So for a hits/bip ratio of 1.08, runs allowed would be charged with about 17% more runs rather than 59%.

The correlation between outs/bip and earned run average is still fairly decent at an R-squared of 42%; however, on-base times slugging has a 77% mark. The slope and intercept show the regression line values. Correlation is in terms of R in the table. So the least squares best fit for ERA as a function of outs/bip is ERA equals –38.1409 x outs/bip + 31.6966. Sigma is the standard deviation of ERA using this formula, which for outs/bip is 2.1018 runs. The average ERA is high because each pitcher is counted equally regardless of innings,

Regression equation coefficients for predicting ERA from a single other statistic

Item	Mean	Slope	Int	Corr-r	Sigma
O/BIP	.692	-38.1409	31.6966	-0.6523	2.1018
OXS	.163	37.7772	-0.8879	0.8778	1.3285
OPS	.806	16.3936	-7.9314	0.8475	1.4719
OBA	.349	45.4494	-10.5559	0.8256	1.5647
SLG	.458	21.6610	-4.6283	0.7900	1.7001

What is the difference between a good pitcher and a poor pitcher? If the good pitcher turns a home run into an extra base hit, his outs/bip goes down because there are more BIPs and the same number of outs. If he turns an extra base hit into a single, he gets no gain at all. If he turns an ordinary out into a strikeout, his outs/bip goes down again because of less BIPs and less outs. If he turns a walk into a ball in play, his outs/bip is unaffected, assuming the batter would make outs at the same rate as existing batters. The only time he makes a gain is if he turns a single (or, equally, a non-HR hit) into an out.

Looking at linear weights instead, you can see that only about 25% of the total difference between pitchers in the 2nd and 5th groups is explained by outs/bip. Since doubles and triples count as singles in outs/bip, only the difference between a single and the double or triple is counted. The difference in actual runs is slightly higher than the difference in earned runs.

Components of the ERA difference between the sixth group (5.22) and the second group (3.28)

(all figures per nine innings)

item	2nd	5th	diff	weight	net
ERA	3.28	5.22	1.94	1.00	1.940
R	3.59	5.65	2.06	1.00	2.060
HR	.83	1.29	.46	1.40	.644
3B	.15	.22	.07	.55*	.038
2B	1.59	2.14	.55	.38*	.220
BB	2.67	3.37	.70	.33	.231
HB	.28	.41	.13	.33	.053
SO	7.28	5.71	-1.57	-.28	.440
1B	8.15	10.13	1.98	.47	.931
out	18.29	19.68	1.39	-.28	-.389
					2.168

*as compared to a single

For example: the second group gave up .83 home runs per nine innings. The fifth group gave up 1.29. The difference is .46 home runs. Multiplied by a linear weight of 1.4, we see that differences in home run rates was responsible for .644 runs per nine innings between the two groups.

There has been talk that on-base average is more important than slugging average in correlating with team runs. This was first reported in Michael Lewis' Moneyball, where Oakland had determined that the ratio should be 3 to 1. Mark Pankin gave a talk at SABR which proposed 2 to 1, or to be more exact, 1.8 to 1. Victor Wang in this publication verified 1.8. However, he also showed that the differences for various values in correlation to team runs was very small. OPS is an approximation of NOPS, normalized OPS, which correlates directly with run scoring. The formula is oba/lg + slg/lg – 1, where lg is the league average. This has the effect of weighting oba slightly higher than slg, since the league average oba is around .333 and slugging is around .400. This results in weighting oba 1.20 times slg. An increase of 20% in oba and 20% in slugging results in run scoring at 40% above average. However OPS shows only a 20% increase. NOPS itself is an approximation of linear weight runs.

When you compare OBA and SLG to team runs, you have to use runs per inning batted, because high scoring teams will usually win more games and therefore bat in less innings. This is of course due to the fact that the home team does not bat in the last of the ninth if ahead. Using runs per game will show these teams often to score less runs than predicted.

Here is a trick you can use to amaze your friends. All you need are complete batting and pitching stats for a team, except wins and losses. You can calculate innings batted by taking total plate appearances minus runs and minus left on base, all divided by 3. You already have innings pitched from the team pitching. Wins are simply equal to games over 2 plus innings pitched minus innings batted. This is because for each home win, you have one less inning batted and one each road win you have one more inning pitched. The only problem involves games that are won in the last inning by the home team, but these tend to cancel out. The standard deviation using this method is about 2 wins per year, compared to 4 using normal methods employing team runs scored and allowed. This does not work for teams that have an unbalanced home/away schedule, which occurred during the strikes of 1981 and 1994. Also in 1991, Montreal was forced to play their last 26 games on the road because of structural problems at Olympic Stadium, losing about 13 home games.

An aside on comparing runs to wins: Bill James' Pythagorean Theorem calculates winning percentage by taking runs scored squared divided by the sum of runs scored squared plus runs allowed squared. The best method varies over the years, but in some cases, using an exponent of 1.83 works better than 2. This formula caused Bill to take 20 years to figure the relationship between runs created and wins when he developed Win Shares. There is a much easier formula that makes the relationship between runs and wins much more obvious, namely 10 runs per win. This states that for every 10 runs more scored or 10 runs less allowed, the team will win one more game. So wins equals runs scored minus runs allowed, all over 10, plus games over 2. Bill's formula works better for high scoring teams or teams with a big run differential, like 160 or more. The easy formula can match this using a divisor of 10 times the square root of runs scored per inning by both teams. For example, if the runs per game is 6, the runs scored per inning by both teams would be 12/9, which would give 11.5 runs per win. Since most teams have small run differences, extreme cases where one method might be better than another get swamped, so you have to look at the unusual cases to find a difference in methods.

Typical results for various periods are shown below.

Standard Deviations of various Wins-From-Runs estimators

	Minimum run		standard deviation in wins			
period	difference	teams	runs(10)	runs(sqi)	Pythag(2)	Pythag(1.83)
1900-19	0	328	4.65	4.34	4.51	4.32
1900-19	160	79	4.88	4.30	5.08	4.89
1920-99	0	1622	4.06	3.95	4.05	3.96
1920-99	160	269	4.29	3.91	4.11	3.95
1920-99	240	70	5.00	4.47	4.84	4.65
1920-99	300	18	6.49	5.12	5.04	4.83
2000-08	0	270	4.09	4.06	4.06	4.08
2000-08	160	33	4.20	4.13	4.29	4.27

If you knew exactly how good each team was at the beginning of the season and tried to predict how many games they would win, the expected standard deviation for each team would be square root of (pqn), where p is the probability of winning, q is the probability of losing (which equals 1-p) and n is the number of games.

For 162 games, this is the square root of 40.5 or 6.34. Yet the derived values are much less, around 4. This is because at the end of the season you have more data, namely the actual number of runs scored and allowed. Even though a team was known to be a .500 team, if by luck they scored more runs than allowed, they would usually win a few more games. So what is the absolute minimum that can be achieved? Looking at all teams from 1920 to 2008 who had a run differential of 5 or less (84 teams), the standard deviation of wins was 4.07, which suggests that 4 is about the best that can be done.

Now let's get back to the OBA multiplier. I used values of x between 0.5 and 1.5 and compared OPS as calculated by x times OBA + (2-x) times SLG to team runs per 27 outs. So OBA over SLG equals x over (1-x). There was very little difference for an OBA to SLG ratio of about 1.3 up to 2.3. Looking at 2000 through 2008, the lowest sigma was around 0.16 runs per game or about 25 runs per season, which is about the best you can do without bringing in steals and double plays. The broad minimum varied only from around 0.158 to 0.160 or about 0.3 runs per year.

Results of team run regressions for various weightings of OBA and SLG, 2000-2008

OBA	SLG	RATIO	OPS	RPG	SLOPE	INT	CORR-R	SIGMA
0.50	1.50	0.33	.804	4.82	11.3611	-4.3077	0.9211	0.1959
0.55	1.45	0.38	.799	4.82	11.5871	-4.4373	0.9238	0.1927
0.60	1.40	0.43	.795	4.82	11.8191	-4.5697	0.9265	0.1893
0.65	1.35	0.48	.790	4.82	12.0579	-4.7053	0.9291	0.1860
0.70	1.30	0.54	.786	4.82	12.2997	-4.8411	0.9318	0.1827
0.75	1.25	0.60	.781	4.82	12.5447	-4.9773	0.9341	0.1796
0.80	1.20	0.67	.777	4.82	12.7927	-5.1136	0.9364	0.1766
0.85	1.15	0.74	.772	4.82	13.0545	-5.2583	0.9388	0.1733
0.90	1.10	0.82	.768	4.82	13.3139	-5.3988	0.9409	0.1705
0.95	1.05	0.90	.763	4.82	13.5808	-5.5427	0.9429	0.1676
1.00	1.00	1.00	.759	4.82	13.8493	-5.6854	0.9447	0.1650
1.05	0.95	1.11	.754	4.82	14.1102	-5.8200	0.9460	0.1631

1.10	0.90	1.22	.750	4.82	14.3830	-5.9612	0.9473	0.1612
1.15	0.85	1.35	.745	4.82	14.6527	-6.0977	0.9484	0.1596
1.20	0.80	1.50	.741	4.82	14.9256	-6.2340	0.9491	0.1585
1.25	**0.75**	**1.67**	**.736**	**4.82**	**15.1916**	**-6.3628**	**0.9495**	**0.1579**
1.30	**0.70**	**1.86**	**.732**	**4.82**	**15.4619**	**-6.4924**	**0.9495**	**0.1579**
1.35	0.65	2.08	.727	4.82	15.7255	-6.6147	0.9491	0.1585
1.40	0.60	2.33	.723	4.82	15.9762	-6.7253	0.9481	0.1600
1.45	0.55	2.64	.718	4.82	16.2319	-6.8372	0.9468	0.1619
1.50	0.50	3.00	.714	4.82	16.4627	-6.9291	0.9447	0.1650

The 1920-1999 period had a slightly higher run variation between predicted and actual, but the broad minimum had about the same range. I started at 1920 because that is the first year that left on base stats were kept.

Results of team run regressions for various weightings of OBA and SLG, 1920-1999

OBA	SLG	RATIO	OPS	RPG	SLOPE	INT	CORR-R	SIGMA
0.50	1.50	0.33	.747	4.47	11.1666	-3.8749	0.9197	0.2631
0.55	1.45	0.38	.744	4.47	11.3890	-4.0077	0.9231	0.2577
0.60	1.40	0.43	.741	4.47	11.6256	-4.1498	0.9268	0.2516
0.65	1.35	0.48	.739	4.47	11.8700	-4.2962	0.9304	0.2455
0.70	1.30	0.54	.736	4.47	12.1094	-4.4375	0.9337	0.2398
0.75	1.25	0.60	.733	4.47	12.3567	-4.5832	0.9371	0.2338
0.80	1.20	0.67	.730	4.47	12.6045	-4.7278	0.9402	0.2281
0.85	1.15	0.74	.727	4.47	12.8579	-4.8750	0.9432	0.2225
0.90	1.10	0.82	.724	4.47	13.1182	-5.0258	0.9462	0.2168
0.95	1.05	0.90	.721	4.47	13.3750	-5.1724	0.9489	0.2114
1.00	1.00	1.00	.718	4.47	13.6337	-5.3190	0.9514	0.2063
1.05	0.95	1.11	.715	4.47	13.8978	-5.4679	0.9537	0.2014
1.10	0.90	1.22	.712	4.47	14.1684	-5.6199	0.9561	0.1964
1.15	0.85	1.35	.709	4.47	14.4169	-5.7546	0.9574	0.1935
1.20	0.80	1.50	.706	4.47	14.6867	-5.9029	0.9591	0.1896
1.25	0.75	1.67	.703	4.47	14.9396	-6.0378	0.9602	0.1872
1.30	0.70	1.86	.700	4.47	15.1927	-6.1712	0.9608	0.1858
1.35	**0.65**	**2.08**	**.698**	**4.47**	**15.4390**	**-6.2986**	**0.9610**	**0.1852**
1.40	0.60	2.33	.695	4.47	15.6761	-6.4180	0.9608	0.1856
1.45	0.55	2.64	.692	4.47	15.9015	-6.5279	0.9600	0.1876
1.50	0.50	3.00	.689	4.47	16.1211	-6.6325	0.9587	0.1905

1900 to 1920 was quite a bit worse, and the broad minimum was around a 1.0 OBA to SLG ratio. This was probably due to the fact that there were fewer homers and a higher correlation between batting average and slugging average. Stolen bases, not included in OPS, were more important. There were more errors and more unearned runs. The percent of unearned runs was 32% in 1900, 20% in 1920, 14% in 1930 and 8% today. The lack of left on base stats made the innings batted calculation less accurate, as I had to derive it using team wins and losses instead. Another problem resulted from trying to apply the same regression line over a period where things were changing. If the sample was divided into two parts, then the best results for 1900-1909 was a sigma of 0.2563 runs per game at a ratio of 1.35, and for 1910-19, it was a sigma of 0.2140 runs per game at a ratio of 1.86. Changing to 5 year intervals reduced the runs per game by another 10%, with a ratio around 1.8.

Results of team run regressions for various weightings of OBA and SLG, 1900-1920

OBA	SLG	RATIO	OPS	RPG	SLOPE	INT	CORR-R	SIGMA
0.50	1.50	0.33	.658	4.02	12.7592	-4.3779	0.8774	0.3519
0.55	1.45	0.38	.657	4.02	12.9292	-4.4791	0.8791	0.3497
0.60	1.40	0.43	.656	4.02	13.0992	-4.5801	0.8805	0.3477
0.65	1.35	0.48	.655	4.02	13.2704	-4.6816	0.8820	0.3457
0.70	1.30	0.54	.655	4.02	13.4375	-4.7801	0.8832	0.3441
0.75	1.25	0.60	.654	4.02	13.6058	-4.8791	0.8842	0.3426
0.80	1.20	0.67	.653	4.02	13.7725	-4.9768	0.8852	0.3413
0.85	1.15	0.74	.652	4.02	13.9374	-5.0731	0.8859	0.3402

0.90	1.10	0.82	.651	4.02	14.0990	-5.1669	0.8865	0.3394
0.95	1.05	0.90	.651	4.02	14.2617	-5.2612	0.8870	0.3387
1.00	**1.00**	**1.00**	**.650**	**4.02**	**14.4171**	**-5.3505**	**0.8872**	**0.3385**
1.05	0.95	1.11	.649	4.02	14.5660	-5.4353	0.8870	0.3387
1.10	0.90	1.22	.648	4.02	14.7205	-5.5235	0.8869	0.3388
1.15	0.85	1.35	.647	4.02	14.8577	-5.6003	0.8863	0.3397
1.20	0.80	1.50	.646	4.02	14.9973	-5.6784	0.8856	0.3407
1.25	0.75	1.67	.646	4.02	15.1290	-5.7511	0.8845	0.3422
1.30	0.70	1.86	.645	4.02	15.2525	-5.8183	0.8832	0.3440
1.35	0.65	2.08	.644	4.02	15.3677	-5.8800	0.8815	0.3463
1.40	0.60	2.33	.643	4.02	15.4711	-5.9339	0.8795	0.3491
1.45	0.55	2.64	.642	4.02	15.5680	-5.9835	0.8771	0.3523
1.50	0.50	3.00	.642	4.02	15.6545	-6.0262	0.8745	0.3558

So the ratio of the importance of oba to slg for optimal run scoring is probably higher than unity, but the actual difference in team runs over the course of a season is small.

Pete Palmer, PETEPALMR@aol.com ♦