By the Numbers

Volume 22, Number 1 The Newsletter of the SABR Statistical Analysis Committee May, 2012

Comment

Academic Research: Correcting the Aging Bias

Charlie Pavitt

The author reviews a recent academic paper that analyzes aging patterns while trying to adjust for talent biases.

Hakes, Jahn K. and Chad Turner (2011). Pay, productivity and aging in major league baseball. Journal of Productivity Analysis, Vol. 35, pp. 61-74

Hakes and Turner (2011) describe a bias others have described (including our esteemed editor Phil in his essay in the *2009 Hardball Times Baseball Annual)* that occurs when pooling across players while estimating career trajectories.

Let us assume that players are major leaguers when their value to their team is above replacement level and are not major leaguers when their value is less than replacement. Assuming any kind of inverted-U relationship between age and value, one obvious implication of this assumption is that less skilled players will spend fewer years above replacement level and thus have shorter careers than the more skilled. In other words, the careers of most less skilled players will begin later and end earlier than the more skilled.

In this issue

The most recent past issue of this publication was November, 2011 (Volume 21, Number 2).

This leads to a less obvious assumption: when calculating mean performance of very young major leaguers (early 20s) and very old major leaguers (late 30s), the analyst will be working with a sample that is disproportionately highly skilled compared with calculating mean performance of major leaguers intermediate in age (late 20s). This wouldn't matter if the shape of the career trajectories for the high and low skilled were the same. But if the shape is different, than calculating career trajectories based on the performance of all skill levels grouped together will result in estimates biased toward the performance of the highly skilled at the front and back ends. The assumption that all career trajectories are either the same or unaffected by any player characteristic (skill would be an example, but other possibilities that come to mind include type of physique and fielding position) is probably not correct.

Hakes and Turner address the issue of calculating career trajectories based on performance data from 1985 through 2005 for hitters with a minimum of 130 at bats in given seasons. As a consequence of the bias problem just described, Hakes and Turner divided their sample into quintiles based on OPS adjusted both for season and for position (as replacement level is of course, far lower for shortstops than for first basemen). The data demonstrated that career trajectories do differ substantially based on skill.

First, the mean peak (estimated as the third highest career OPS, assuming that the first and second highest represent playing above one's head) ranged from 25.6 years of age for the lowest quintile up to 28.2 for the highest, implying that career peak is later for the higher skilled.

Second, beta coefficients for the slope of the estimated trajectories during early years increased along with higher OPS, meaning that upon starting their careers, young, highly skilled players improved their performance more quickly than did the young, lowly skilled.

Third, squared beta coefficients representing curvature of the estimated trajectories increased along with higher OPS, meaning that the more skilled players have greater variation in performance across their careers than the less skilled, as the latter's performance doesn't rise much above replacement level during their shorter tenures.

Finally, Hakes and Turner examined the implications of the career-length bias more directly by determining that the lower skilled enter the majors at ages just a bit older than the highly skilled but leave quite a bit earlier. As a consequence, the impact of overestimation is greater for old major leaguers than for young ones.

In attempting to address the skill-difference bias, Hakes and Turner may introduced some issues of their own. Using the third-highest adjusted OPS to represent peak performance at the very least makes it hard to compare their findings to other studies that have used the very best season. In addition, there is a problem that may be endemic to all this research but that Hakes and Turner's approach makes particularly salient. Suppose a player falls below replacement level for a season at a relatively young age and is banished to the minors. Then suppose that player not only reverts to greater than replacement level but performs better than ever, but has been saddled with the reputation of "minor leaguer" and is not given the opportunity to show his stuff in the majors during his peak (think Esteban German here). If this were the case for a substantial number of lower skilled players, then Hakes and Turner would be underestimating the age of that skill group's peak.

Charlie Pavitt, chazzq@UDel.Edu ♦

Review: "Basic Ball"

Phil Birnbaum

A review of "Basic Ball," a recent book in which Pete Palmer and Dave Heeren explain sabermetric principles of baseball, football, and basketball, while rating players in all three sports.

Pete Palmer's most famous work is "*The Hidden Game of Baseball*," co-written with John Thorn. It appeared in 1984, about the same time as the *Bill James Baseball Abstract*, then in its third commercially-published season, was picking up steam. In that era, those two were virtually the only sabermetrics books available, and James and Palmer were often cited together as twin pioneers.

Pete has been quieter than James since then; he's had a couple of analytical books out, and he's been active publishing "Total Baseball" and other baseball encyclopedias. He's also put out a few articles, in BTN and elsewhere, but, overall, he's had a lot less output than James and many of the others who came along since.

I've always thought that Palmer's work never got enough attention. He's the inventor of "Linear Weights," which basically calculates the run values of various events. Those values are one of the more important components of the sabermetrics toolkit. But Bill James doesn't use them much, and academic papers tend to credit others, so Palmer's work winds up with a lot less attention than I think it deserves.

So, his latest book, "Basic Ball," is very welcome. It actually covers three sports -- baseball, basketball, and football. For all three sports, it's co-written with Dave Heeren. Actually, it's an anthology more than a collaboration, as the two authors are responsible for different chapters.

Most of the baseball section is Palmer's, and his chapters are a must read. They're essentially a summary of the work Pete has done over his career.

Basic Ball: New approaches for determining the greatest baseball, football, and basketball players of all-time

By Dave Heeren and Pete Palmer

St. Johann Press, 293 pages,
$29.95 (US),
ISBN: 187828276X

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

Think of "The Hidden Game," but for engineers instead of literary types -- written in the style of Pete Palmer instead of John Thorn. Pete's approach is … well, it's right to the point. Pete knows what he wants to say, and he just says it, in a way that seems to be aimed at intelligent readers who aren't scared of math, and who may already somewhat familiar with sabermetrics. For instance:

> "Every time a batter does not make an out, not only does he get on base himself, but he also allows another batter to come up with the same number of outs. And that batter could get on base, bringing up still another batter. This becomes an infinite series of 1 plus 1/3 plus 1/9 plus 1/27, etc., that converges at 1.5, which is 1 divided by the quantity 1 minus 1/3. A batter adds one more batter than average every time he gets on base and loses ½ a batter less than average when he makes an out."

Spoken like a true engineer! It explains the concept perfectly and understandably, in only as few words as necessary to cover it.

Pete's style does have its idiosyncrasies; the first paragraph of the book is, I think, about 500 words, and there's another paragraph that's about 800. It might put you off a little if you're a beginner, but readers of BTN should have no problem with it.

And, don't be scared by the infinite series explanation … it is, I think, the only part of the book that goes beyond simple arithmetic.

Another nice thing about the style is that it lets Pete throw in almost everything he knows. Well, of course, not everything -- I've always thought that, in general, it's so hard for an expert to explain what he knows that at least 80 percent of it is left untold. . It's clear that Pete has thought of a lot of things that didn't make it to the book, but you're still getting a great summary of the highlights.

And, even though most of the material will be familiar to Pete's devoted readers, there are still a few things that I hadn't seen before, that Pete's done in the three decades since "The Hidden Game."

For instance, Chapter 6 is about "Luck and Skill," which is one of my areas of particular interest. In the past few years, discussion on Tom Tango's blog, and elsewhere, has resulted in a consensus of understanding how luck works in the various sports. One important principle is Tango's method for calculating the variance of team talent. He observed that you can calculate the theoretical variance of team records that would be due to luck, using the binomial approximation to normal. And, of course, you can easily calculate the observed variance of W-L records. Therefore, if you subtract the first variance from the second, you're left with an estimate of the variance of talent.

Well, Pete summarizes some of that stuff in the chapter, but it seems he'd figured it out well before the community did. His chapter covers some of the material he first wrote about in a guest article in "Baseball Hacks," which was written in 2005 (Tango's post, the first that I know of, was 2006 -- I will probably credit both of them from now on).

My guess is that Pete had figured all this out years before, but 2005 was the first time he'd written about it. Which is another reason this book is a must -- for the first time in a long while, you get to read Pete Palmer talking about how sabermetrics works. Even when it's not something that's new, you get to hear Palmer reason it out, and you wind up thinking about it in a different way.

As I was reading Pete's chapters, I was thinking about how differently Bill James would approach the same questions. Bill is more the sage, while Pete is more the engineer. Bill would explain things with convincing analogies and relevant examples. And Pete will reply, "well, yes … but, really, it's because the infinite series sums to 1.5."

I'm exaggerating a bit -- as I said, there's not a lot of fancy math in the book. But Pete's direct approach had me smiling as I read through it. He's a science geek explaining stuff to other science geeks like me. And we need a book like that. It's a book that doesn't treat sabermetrics as a novelty, or assume that the reader needs help with basic principles of arithmetic. It treats sabermetrics as a science, and explains it to the reader directly, the same way a physicist might explain Newton's Three Laws.

If you're interested in understanding the sabermetric thinking on strategy questions, for instance, I'd send you right to Chapter 9. It's like a six-page introductory course, a perfect first step you'd take before looking deeper (by, for instance, going to the more detailed analysis found in Tango/Lichtman/Dolphin's "The Book").

It's not all math, by the way. There are chapters on the history of statistics in all three sports. The baseball side was covered in "The Hidden Game," but I'd never seen the football or basketball timeline before, and those are very nice work. Dave Heeren, who is most famous for his basketball expertise, wrote the history of that sport.

Actually, I don't think any of Heeren's chapters are particularly mathy. Rather, his focus is mostly on rating players and teams, using his own formulas, for all three sports, but especially basketball. (In fact, the entire basketball section of the book belongs to Heeren.)

He's most noted for his TENDEX system for evaluating basketball players, and he uses that to pick some top tens in each category. That's stuff that will appeal to basketball fans. That doesn't really include me, but I still enjoyed his argument claiming the 1970-71 Milwaukee Bucks as the greatest team of all time.

But, one thing I had wished Heeren had done more of, is … science. That is, I wish he had explained why his system works, and how it compares to the others. There's a lot of new basketball research and statistics out there. Dean Oliver is well noted for his systems in "Basketball on Paper," and David Berri's recent books have their own systems, too. There's also an active basketball "APBRmetric" community that debates the ins and outs of some of these methods.

So, I wish there had been some of that. My gut feeling seems to be that Heeren's formulas are kind of arbitrary -- based on a certain amount of reasonableness, but not thoroughly tested for flaws and inaccuracies. The baseball section has Heeren evaluating players by formulas that are very inaccurate by established sabermetric standards. I don't know if the basketball and football ones compare better. However, in a nice chapter on the NBA draft, Heeren argues that his TENDEX statistic has outpredicted the scouts on draft day.

The football section is a combination of Palmer and Heeren. Pete lays out the analytics and reasoning, just as he did for baseball. It's the best starting point I've seen for anyone who wants to learn about football sabermetrics. Palmer is also a pioneer in that field; he co-wrote "The Hidden Game of Football" in 1989, with an update in 1998. I've always thought that book deserved more attention than it's received.

If there's a weakness to the book, it's perhaps that neither author seems too aware of the substantial progress that's been made in the field, especially in the last decade or so.

For instance, Palmer uses the same rating system as he has for the past three decades, which is denominated in wins above average. He does mention Bill James' "Win Shares" (2002), and discusses the difference that James' stat is denominated in runs above zero. But he doesn't mention "Wins Above Replacement," which (in my experience) is almost universally accepted as the best method by the sabermetric community.

And Palmer's outside references are a little dated. He mentions Baseball Prospectus, but none of the other websites that have seen at least as much current research (especially "The Book"). And, for football, he doesn't mention Brian Burke's research blog, which is one of the most respected and cited.

Of course, it could be just bias on my part, thinking that the sites I frequent the most are also the most relevant. Still, I wish the reader had been left with a feeling of how much the field has advanced since the biggest part of Pete's (substantial) contributions.

So, if you buy this book, know what you're getting -- a summary of the work of Pete Palmer and Dave Heeren, and not too many others. But that's not a bad thing. Both have been plying their trade since the 1950s, and they have a lot to offer. You get the impression that Heeren knows every basketball player over the last fifty years. And you get the impression that Pete can derive any one of his sabermetric principles, on a blackboard, from first principles, in thirty seconds, blindfolded.

No matter how well you know something, you always see it a bit better after you've heard it explained by a master practitioner like Pete. I hope he does more of this kind of writing.

Phil Birnbaum, 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 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.

Trends in “Strategy” in Major League Baseball

Donald A. Coffin

The author looks at how rates of sacrifices, stolen base attempts, and intentional walk rates have changed over the history of baseball, and how those changes correlate with increases and decreases in power stats.

I. Introduction

I think it is fair to say that most people who have done statistical analyses of run scoring and run prevention in Major League Baseball (MLB) have concluded that certain choices (two on offense and one on defense, in particular) are not, in general, particularly useful. On offense, most analysts have not had much good to say about sacrifice bunts or stolen base attempts; on defense, the intentional base on balls has similarly not found much favor. What I have not seen, although I concede that there has been much written on this subject that I am not aware of, and that is not easily accessible, is a discussion in the trends in how these particular strategic choices have changed over time, or in whether such changes are related to other changes in the game. In this piece, I take a look at both questions.

To anticipate my conclusions: I find, using data from Baseball Reference, that both sacrifice bunts per game for the 1912-2011 period) and intentional bases-on-balls per game for the 1955-2011 period; IBB data are not available before 1955) have been used less frequently over time. Stolen base attempts per game declined precipitously between 1912 and 1950, then increased significantly between 1950 and about 1986. That was followed by a period of decline (to about 2004, with a modest increase since).

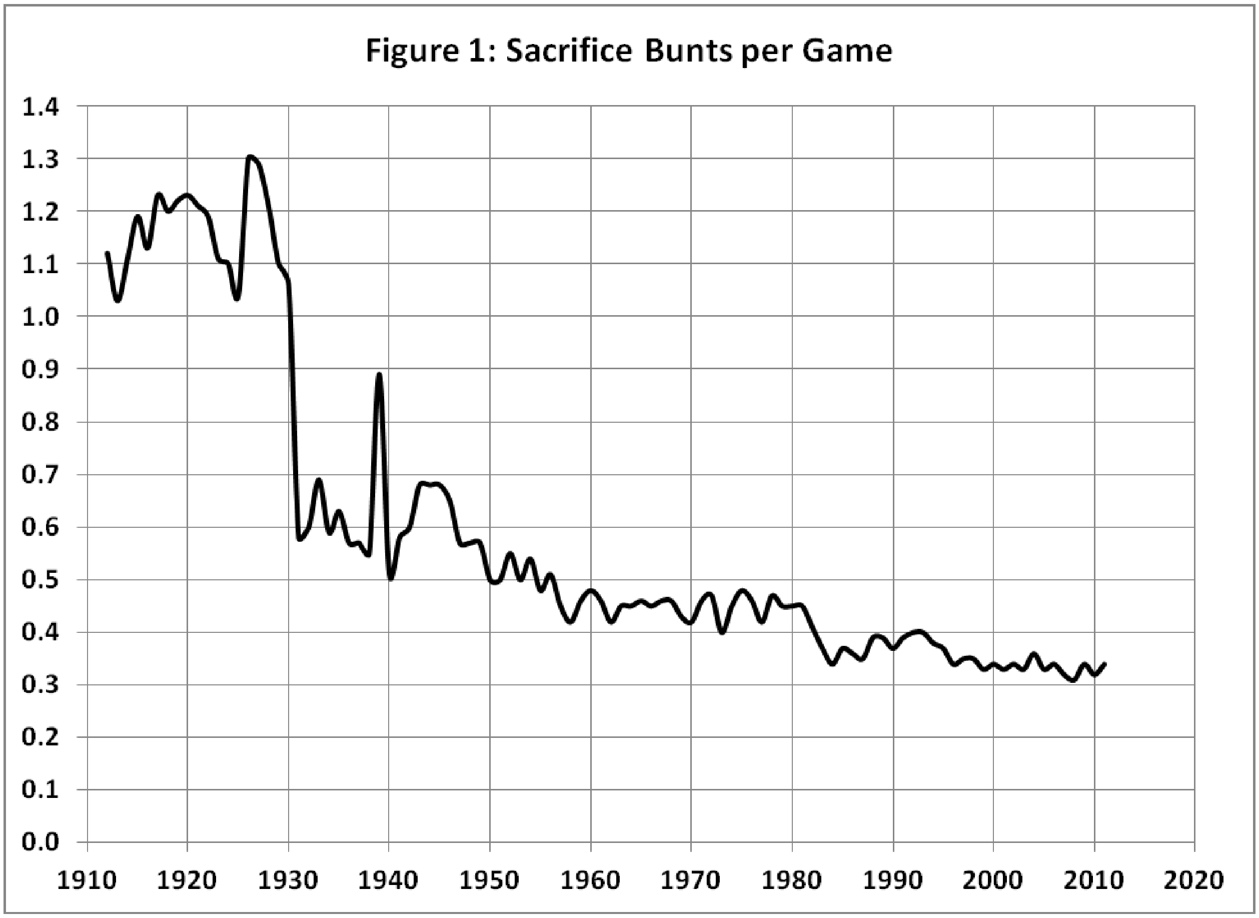
I also find that there are significant negative correlations between the use of these managerial choices and measures of power hitting. There is a clear tendency for the use of the sacrifice bunt, the stolen base attempt, and the intentional base on balls to decline as home runs per game and isolated power increase.

II. Trends in the Use of Sacrifice Bunts, Stolen Base Attempts, and Intentional Bases-on-Balls[1]

Sacrifice Bunts

It is, I think, fairly widely known that the use of the sacrifice bunt was extraordinarily common in MLB in the early 20th century. Between 1912 (when Baseball Reference first reports the data) and, to my surprise, 1926, sacrifice bunts per team per game increased from 1.12 to 1.3 (see Figure 1, next page). Had I been required to guess, I would have guessed that decline in the use of the sacrifice bunt would have started sooner. In any event, by 1931, sacrifices were down to 0.58 per team per game.

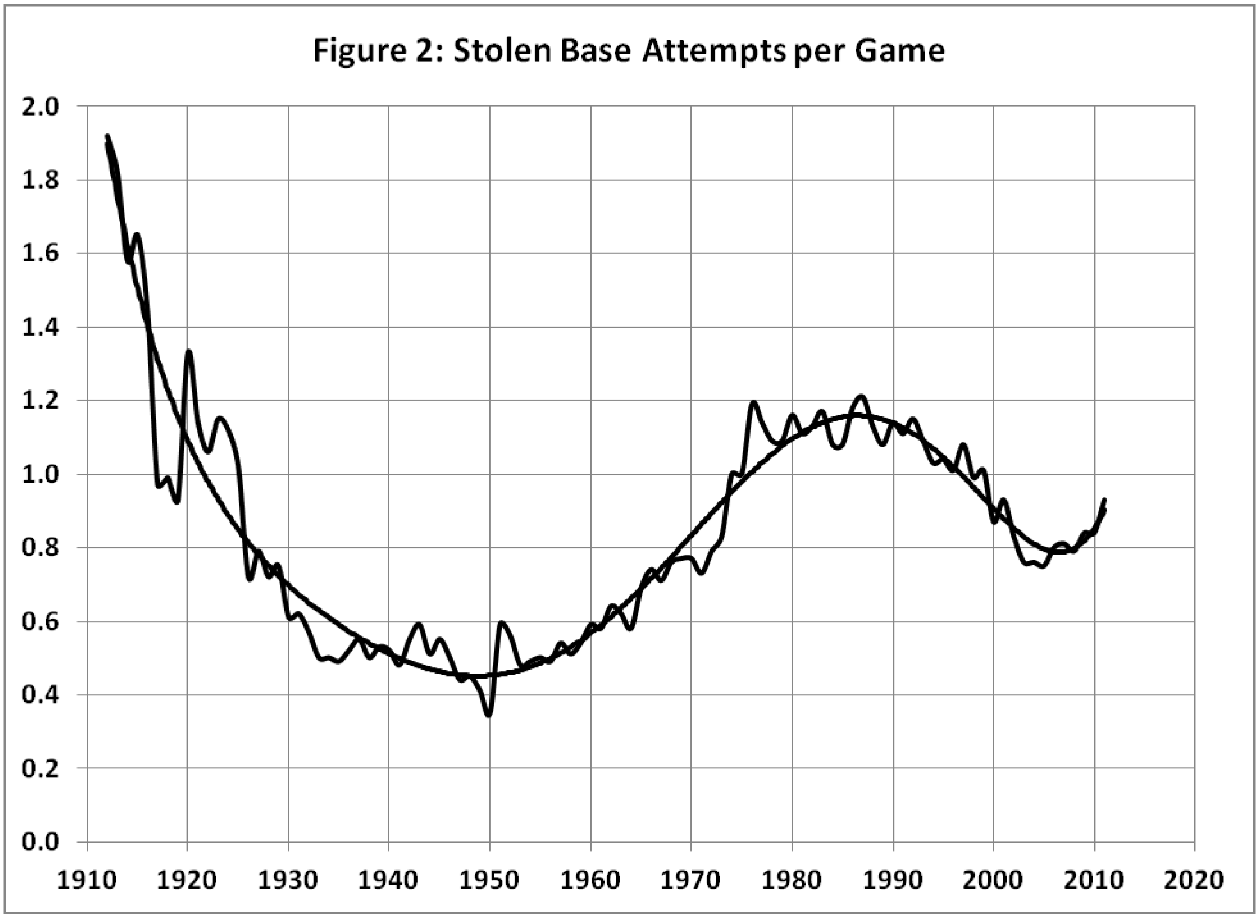
[1] I measure the use of these tactics on a per-game basis, largely to control for expansion. For stolen base attempts per game and intentional bases on balls per game, the diagrams that follow include a trend line. I do not include a trend line for sacrifices, because of the somewhat unique shape of the curve. All data are from www.baseballreference.com.



The spike in 1939 is a consequence of a rules change; in 1939, but not before or afterward, SFs were included in sacrifices. [Later (in the 1950s), SFs began to be counted as a separate category.] So the use of the bunt declines fairly steadily (albeit with a bump upward during World War II) since 1931. Between 1931 and 2011, the rate declines from 0.58 per game to 0.34 per game, an average annual rate of decrease of 0.67% per year. To phrase it somewhat differently, in 1931, a random team sacrificed slightly more than once every two games; now, it's once every three. Compared with 1912, the rate of use of sacrifice bunts had declined by nearly 70% (from 1.12 per team per game to 0.34 per team per game) by 2011.

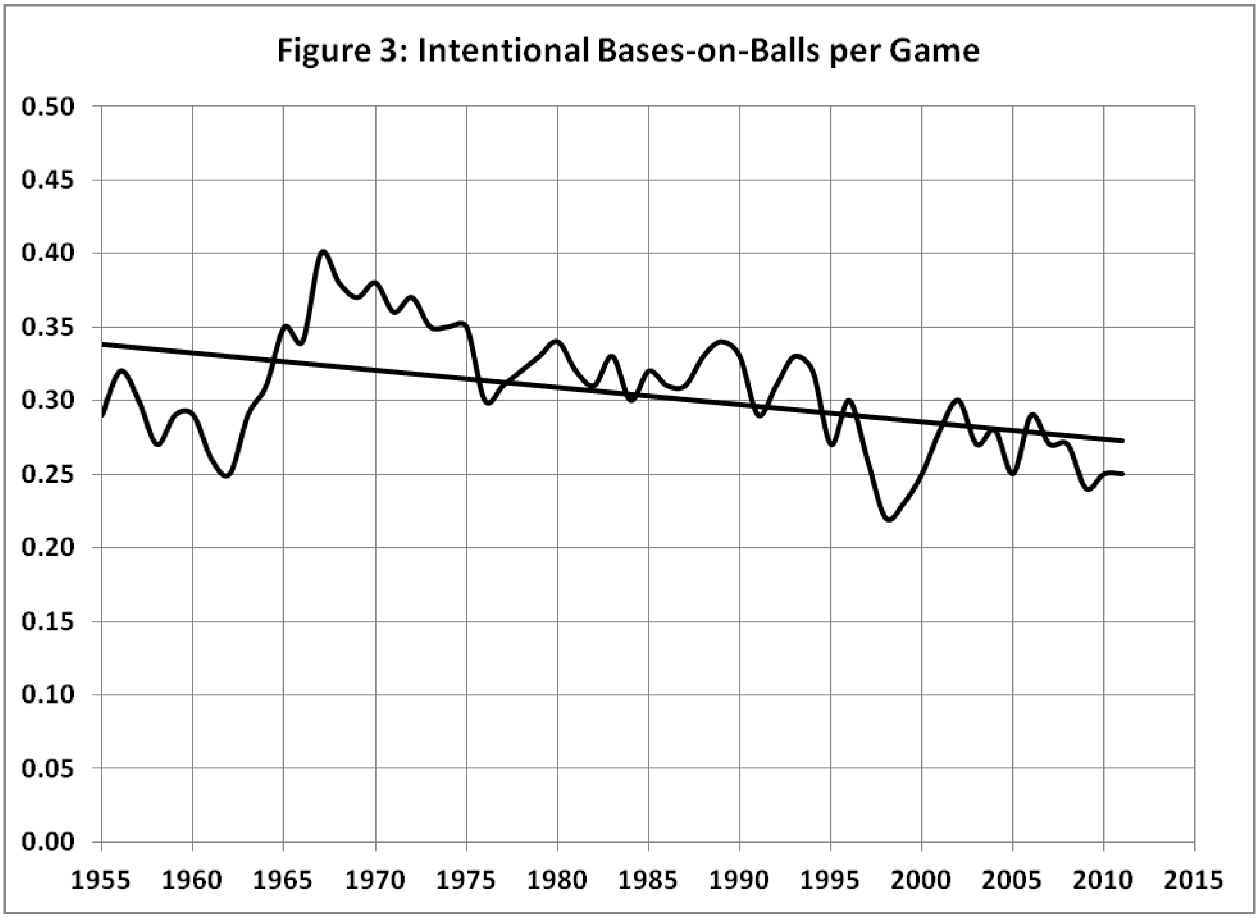
Stolen Base Attempts

In 1912, teams attempted 1.92 steals per game. At the low point of base stealing (1950), that had dropped to 0.70 per game, a decline of nearly 62%. But steal attempts recovered somewhat from the 1950s through the late 1980s, rising to 1.21 per team per game in 1987, before falling once again to 0.75 per team per game in 2007. In 2011, teams attempted 0.965 steals per team per game. That's still about half the rate of 1912. A polynomial trend fits the changes over time very closely, as can be seen in Figure 2.



Intentional Walks

Baseball Reference does not present data on the intentional base on balls before 1955, and, in 1955, IBBs were issued at a rate of 0.29 per team per game. Use of the IBB declined, somewhat erratically,[2] from its 1967 peak of 0.4 per team per game to the 2011 level of 0.25 per team per game, a rate of decline of 1.1% per year. (In 1967, the typical team issued about 1 intentional walk every 2.5 games; now, it's once every 4 games.)



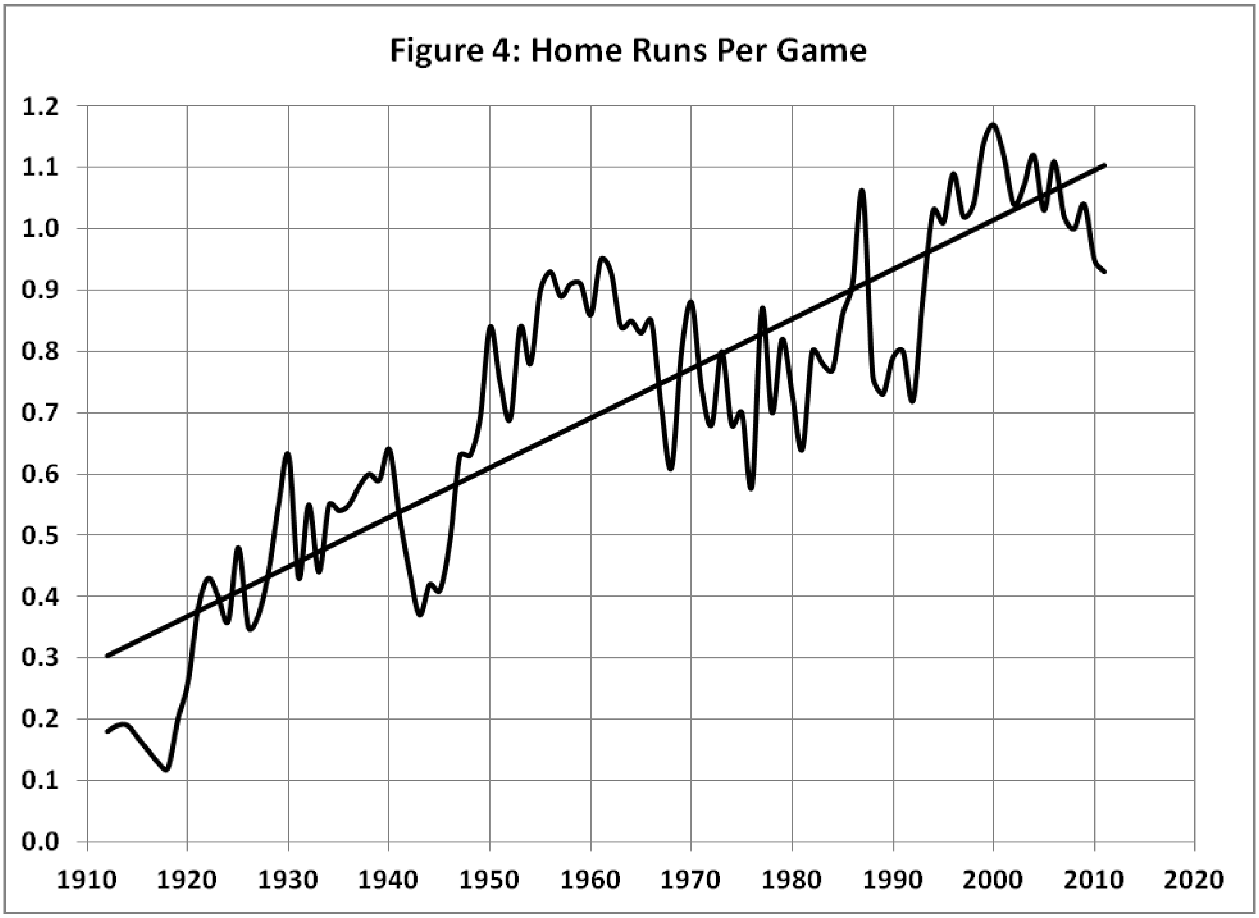
I will also note that even the incredible number of IBBs drawn by Barry Bonds did not reverse this downward trend. I have not removed his numbers, but, were I to do so, the decline would be much larger. Consider that in 2007, Bonds drew, by himself, 3.3% of all IBBs in MLB; in 2004, his 120 IBBs were 8.9% of all IBBs.

So far, we can see that broadly the use of all three of these managerial tools has declined since the early 20th century, and only the use of the stolen base has not been in a state of more-or-less continuous decline. It should be clear that, over time, managers have become less and less inclined to use these as means of scoring (in the case of sacrifices and stolen base attempts) or preventing (IBBs) runs.

[2] I regard the increase in the use of the intentional walk during the late 1960s—from 0.25 per team per game in 1962 to 0.4 per team per game in 1967—as odd. During that time, runs per game (per team) fell by 15%, from 4.46 per game to 3.77 per game. This does not seem, to me, to be an offensive environment in which one would want to give a team a baserunner.

III. The Long-Run Increase in Power[3]

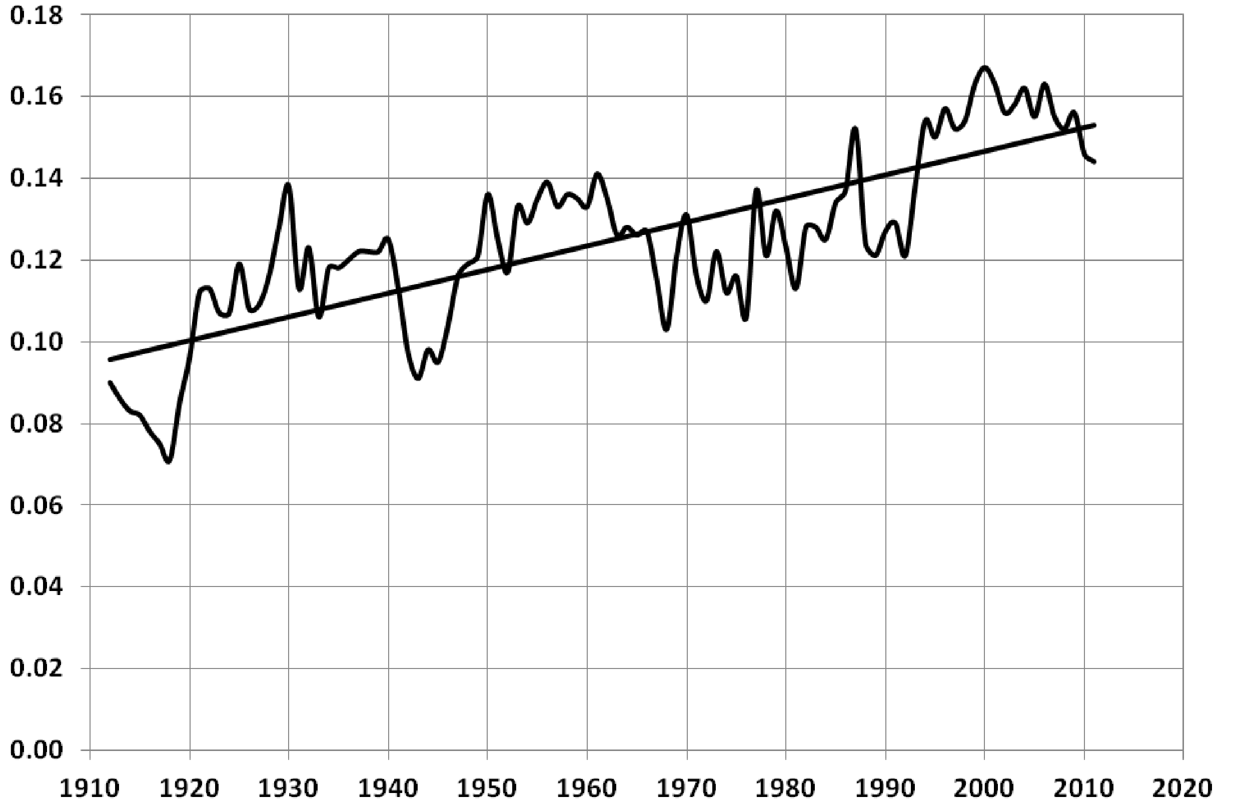
If one thing about offense in MLB is clear, it is that power (however one might choose to measure it) has shown a long term increase. We can see this very clearly by looking at home runs per game and at isolated power.



Home runs were rare in the early 20th century, which I'm sure I didn't have to say. In 1912, teams averaged about 0.18 home runs per game; this means the average team hit about 30 per season. By 2011, the average was about 0.93 per team per game (down from 1.12 per team per game in 2000). In a sense, this quintupling of HRs per game is just the converse of the 62% decrease in stolen base attempts that occurred even more rapidly, between 1912 and 1950. While the increase in HRs per game has not been as steady as some of the changes we have already seen, it has been persistent. Home-run hitting stagnated (and bounced around a lot) in the 1970s, but, otherwise, the trend has been strongly upward. Between 1912 and 2011, HRs per game has increased at an average annual rate of 1.67% per year. Despite the decline in the most recent decade, I don't see any reason to expect that HRs per game will not shortly resume its 100-year-long rise.

[3] In this section, both charts—for Home Runs per Game and Isolated Power--include linear trend lines.

Figure 5: Isolated Power



Isolated Power (slugging percentage minus batting average) has shown an equally strong and persistent increase. From 0.090 in 1912, it has increased to 0.144 in 2011 (again, down from a peak of 0.167 in 2000). The increase—an average annual rate of about 0.5% per year—has been steadier, and the downturns more modest, than has been the case with home runs. Again, there is no reason I can see to expect that this long-run upward trend in power-based offenses will not resume.

Table 1– Power and Strategy

	Correlation Coefficient (ρ)
ρ, HR/G with SAC/G	-0.851
ρ, HR/G with SBA/G	-0.222
ρ, HR/G with IBB/G	-0.691
ρ, ISO with SAC/G	-0.704
ρ, ISO with SBA/G	-0.208
ρ, ISO with IBB/G	-0.728

SAC/G = Sacrifices per game
SBA/G = Stolen base attempts per game
IBB/G = Intentional walks per game
All ρ are significant at the 1% level.

IV. The Apparent Relationship Between Increasing Power and Decreasing Use of Sacrifices, Steals, and IBBs

I say "apparent," because, as we all know, correlation does not prove causation. But there is certainly correlation.

I won't include the diagrams showing the scatterplots of HRs per game and the three "strategy" variables and of isolated power with them. The correlations can be seen in Table 1.

It's quite clear that the correlations, especially between power measures and sacrifices and IBBs, are large. There is a relationship between power and these two measures of managerial strategic choices that is hard to ignore. And while it's possible that these correlations exist solely by chance, or that both are caused by some other factor, that seems to me unlikely. It would seem clear that managers, who could not avoid seeing and thus could not ignore the rise of power as an offensive weapon, would become more reluctant to give up an out for a single base (in the case of the sacrifice bunt) or to present the other team with a baserunner (in the case of the IBB).

Stolen base attempts present the hard case. Here, the long-term pattern is not a simple decline, but distinct periods of decline, increase, decline, and increase (a symphony, if you will, in four movements—so far). But overall, stolen base attempts are down significantly, and the correlation with power numbers, while only around -0.2, is statistically significant. Whatever relationship there is here, it is clearly not as strong, but it also cannot, I think, be completely ignored.

V. Conclusions

We all know that offense in MLB has changed over time, moving more toward power and away from "small-ball." One can applaud that trend, or regret it, but it is there, and it is hard to ignore. Further, the trend toward increased power, while it has been occasionally been interrupted, has not yet been permanently halted. I can see no reason to expect the recent downturn in HRs per game or in isolated power to be anything but temporary.

The change in offense has changed the risks and rewards of strategies for offense and for defense. It has made giving up outs less rewarding, and thus using a strategy that might give up an out riskier. It has made the defensive strategy of giving the offense a "free" baserunner also riskier and, thus, also less rewarding.

What I have not seen noted or commented on (again, there's much I have not seen) is that managerial choices seem to have responded to the changes in this risk/reward structure. Sacrifices have declined dramatically, IBBs are used significantly less often, and even stolen base attempts are down (although this could use some further examination). Managers, apparently, can, and do, learn, and managers can, and do, adapt when circumstances change.

Donald A. Coffin, dcoffin@iun.edu ♦

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 .

First-Order Approximations of the Pythagorean Formula

Kevin D. Dayaratna and Steven J. Miller

The authors show, mathematically, that a linear approach converting runs to wins is simply a first-order approximation to Bill James' Pythagorean Projection, thus helping to explain why the linear model works as well as it does. In addition, the authors estimate the linear model on actual MLB data, to verify the best value for the exponent.

I. Introduction

First postulated by Bill James in the early 1980s, the Pythagorean Won-Loss formula indicates the percentage of games (winning percentage, WP) a baseball team should have won at a particular point in a season as a function of average runs scored (RS) and average runs allowed (RA):

$$\mathrm{WP} \approx \frac{\mathrm{RS}^{\gamma}}{\mathrm{RS}^{\gamma} + \mathrm{RA}^{\gamma}}.$$

James initially postulated the exponent γ to be 2 (hence the name "Pythagorean" from a sum of squares). Empirical observation suggested that $\gamma \approx 1.8$ was more appropriate.

For decades, the Pythagorean Won-Loss formula gave a strong indication of the percentage of games a baseball team should have won at a particular point in a season. Until just a few years ago, however, the formula had no statistical verification. Miller (2007) provided such verification by assuming that runs scored and runs allowed follow separate independent continuous Weibull distributions. Upon making these assumptions, he was able to derive James's formula in the form of the probability that the runs a particular team scores is greater than the runs it allows. He estimated this model via least squares and maximum likelihood estimation on 2004 American League data and determined that the appropriate value of γ was indeed around 1.8, consistent with empirical observation.

Jones and Tappin (2005) presented a simple linear model that also serves as a predictor of the team's winning percentage. In the following section, we prove that this formula is actually nothing but a first order approximation to the Pythagorean Won-Loss formula:

$$\mathrm{W}P = .500 + \beta(\mathrm{R}S - \mathrm{R}A);$$

here $\mathrm{W}P$ is the team's winning percentage, $\mathrm{R}S$ is the average points scored (goals in hockey, runs in baseball, et cetera) and $\mathrm{R}A$ is the average points allowed. Notice that if $\mathrm{R}S = RA$ then the team is predicted to win half its games. Typically β is a small number. As a result, for observed values of $\mathrm{R}S$ and $\mathrm{R}A$ we do not need to worry about the above expression exceeding 1.000 or falling below 0.000. For example, in baseball in 2010 runs scored ranged from 513 to 859 and runs allowed from 581 to 845. For these ranges, the winning percentages are all "reasonable," ranging from 0.352 to 0.599. (MLB.com)

II. Derivation of Linear Predictor

We now show how the above linear predictor follows from the Pythagorean formula. We assume there is some exponent γ such that

$$\mathrm{W}P = \frac{\mathrm{R}S^{\gamma}}{\mathrm{R}S^{\gamma} + \mathrm{R}A^{\gamma}}.$$

We provide a simple statistical derivation of the linear formula utilizing multivariable calculus.[4]

Proof

In this subsection we assume the reader is familiar with multivariable calculus. Recall the second order Taylor series expansion of a function $f(x,y)$ about the point (a,b) is

$$\begin{aligned} f(x,y) = f(a,b) &+ \left.\frac{\partial f}{\partial x}\right|_{(a,b)} (x-a) + \left.\frac{\partial f}{\partial y}\right|_{(a,b)} (y-b) + \frac{1}{2}\left.\frac{\partial^2 f}{\partial x^2}\right|_{(a,b)} (x-a)^2 \\ &+ \left.\frac{\partial^2 f}{\partial x \partial y}\right|_{(a,b)} (x-a)(y-b) + \frac{1}{2}\left.\frac{\partial^2 f}{\partial y^2}\right|_{(a,b)} (y-b)^2 \\ &+ \textit{higher order terms}. \end{aligned}$$

Here, the higher order terms involve products of $(x-a)$ and $(y-b)$ to the third and higher powers. The tangent plane approximation, which means keeping just the constant and linear terms, is

$$f(x,y) = f(a,b) + \left.\frac{\partial f}{\partial x}\right|_{(a,b)} (x-a) + \left.\frac{\partial f}{\partial y}\right|_{(a,b)} (y-b).$$

Let R_{ave} denote the average number of runs scored in the league. We let

$$f(x,y) = \frac{x^{\gamma}}{x^{\gamma} + y^{\gamma}}.$$

[4] An alternative proof using single variable calculus is included in an expanded version of this paper, available online at: http://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/DM_LinPythag_App.pdf

We now expand about the point $(a,b) = (\mathrm{R}_{ave}, \mathrm{R}_{ave})$, with $x = \mathrm{R}S$ and $y = \mathrm{R}A$, so

$$f(\mathrm{R}_{ave}, \mathrm{R}_{ave}) = .500$$

$$\frac{\partial f}{\partial x} = \frac{\gamma x^{\gamma-1} y^{\gamma}}{(x^{\gamma} + y^{\gamma})^2} \Rightarrow \left.\frac{\partial f}{\partial x}\right|_{(\mathrm{R}_{ave}, \mathrm{R}_{ave})} = \frac{\gamma}{4\mathrm{R}_{ave}}$$

$$\frac{\partial f}{\partial y} = -\frac{\gamma x^{\gamma} y^{\gamma-1}}{(x^{\gamma} + y^{\gamma})^2} \Rightarrow \left.\frac{\partial f}{\partial y}\right|_{(\mathrm{R}_{ave}, \mathrm{R}_{ave})} = -\frac{\gamma}{4\mathrm{R}_{ave}}.$$

Noting that the predicted winning percentage is $f(\mathrm{R}S, \mathrm{R}A)$, we see that the first order, multivariate Taylor series expansion about $(\mathrm{R}S, \mathrm{R}A)$ gives

$$\mathrm{W}P \approx .500 + \frac{\gamma}{4\mathrm{R}_{ave}}(\mathrm{R}S - \mathrm{R}_{ave}) - \frac{\gamma}{4\mathrm{R}_{ave}}(\mathrm{R}A - \mathrm{R}_{ave}) = .500 + \frac{\gamma}{4\mathrm{R}_{ave}}(\mathrm{R}S - \mathrm{R}A).$$

III. Model Estimation

Michael Jones and Linda Tappin (2005) used this linear model for baseball. They wrote $\mathrm{W}P = .500 + \beta(\mathrm{R}S - \mathrm{R}A)$, and by looking at the seasonal data from 1969 to 2003 found the best values of β ranged from .00053 to .00078, with an average value of .00065. Taking their average value of .00065 and using $\gamma = 1.81$ leads to a predicted value of 696 runs scored per team per year, or about 4.3 runs per game. Conversely, using the average number of runs scored in 2010 by American League teams (721) and their average value of β, one gets a prediction of 1.88 for γ.

Our analysis in Section II provides theoretical support for the linear model. In particular, the slope is no longer a mysterious quantity, but is naturally related to the exponent and average scoring in the league. Here, we also provide empirical support by estimating the model via the method of least squares:

$$\mathrm{W}P \approx \alpha + \beta(RS - RA), \text{where } \beta = \frac{\gamma}{4R_{ave}}.$$

On the following page are our estimates via the method of least squares:

Coefficient Estimates and Model Fit Statistics

Season	$\hat{\alpha}$	$\hat{\beta}$	R_{ave}	$\hat{\gamma}$	95% Lower Bound on $\hat{\gamma}$	95% Upper Bound on $\hat{\gamma}$	R^2
1991	0.500	0.119	4.308	2.058	1.807	2.310	0.922
1992	0.500	0.126	4.117	2.076	1.710	2.442	0.851
1993	0.500	0.109	4.598	2.001	1.645	2.359	0.851
1994	0.500	0.084	4.923	1.658	1.366	1.951	0.836
1995	0.500	0.094	4.847	1.826	1.466	2.185	0.807
1996	0.500	0.091	5.036	1.825	1.564	2.085	0.889
1997	0.500	0.087	4.767	1.668	1.345	1.991	0.813
1998	0.500	0.098	4.790	1.881	1.667	2.095	0.920
1999	0.500	0.099	5.085	2.010	1.794	2.226	0.929
2000	0.500	0.092	5.140	1.893	1.626	2.160	0.883
2001	0.500	0.104	4.775	1.978	1.743	2.215	0.913
2002	0.500	0.103	4.618	1.908	1.682	2.134	0.914
2003	0.500	0.103	4.728	1.949	1.716	2.181	0.913
2004	0.500	0.109	4.814	2.108	1.843	2.374	0.905
2005	0.500	0.095	4.586	1.737	1.436	2.040	0.833
2006	0.500	0.098	4.858	1.901	1.567	2.235	0.829
2007	0.500	0.085	4.797	1.640	1.330	1.951	0.807
2008	0.500	0.104	4.651	1.931	1.619	2.244	0.851
2009	0.500	0.106	4.613	1.963	1.642	2.284	0.848
2010	0.500	0.094	4.366	1.634	1.489	1.780	0.950
2011	0.500	0.104	4.283	1.775	1.506	2.045	0.867

After choosing a standard significance level of 0.05 and instituting Bonferroni corrections, which reduces our significance level to 0.0025, each of our coefficient estimates, as well as overall model fit, is highly significant. This statistical significance, coupled with our coefficients of determination being reasonably close to one, signify that our linear model fits quite well.

Furthermore, with the exception of the 2010 season, the commonly accepted value of 1.82 for the exponent for the Pythagorean Won Loss formula (Miller 2007) falls within all of our 95% confidence intervals. Bonferroni corrections increase the size of all of our confidence intervals, including for the estimates pertaining to the 2010 season (to an interval of [1.399, 1.870]). These facts provide us with empirical verification that the Jones and Tappin (2005) linear model of winning percentages is simply just a first order approximation to the Pythagorean Won-Loss formula.

IV. Conclusions and Future Research

We have provided a theoretical justification for an existing linear model that allows for an interpretation of the slope parameter in terms of the Pythagorean Won-Loss formula's coefficient. Our theoretical work, along with our model estimation, helps explain why this simple and elegant linear model is such a strong linear predictor.

There are a number of potential avenues of future research we hope our work will encourage. We have presented a first order approximation of the Pythagorean Won-Loss Formula. In future research, one could compare higher order approximations to the one presented here. Secondly, one could examine slight variations in γ as a result of changes over time such as steroid use, height of the pitcher's mound, players' diets, and the introduction of inter-league play among others. Thirdly, one could apply this model to other sports such as basketball, hockey, football, and soccer. Finally, it could be fascinating to apply this model to a much larger span of data and compare resulting coefficient estimates for teams of different eras.

References

- Baseball Almanac, http://baseball-almanac.com. Retrieved January 2012
- Casella G. and Berger R., Statistical Inference, Second Edition, Duxbury Advanced Series, 2002.
- Ciccolella, R. Are Runs Scored and Runs Allowed Independent, By the Numbers 16 (2006), no. 1, 11--15.
- ESPN.com, “MLB - Major League Baseball Teams, Scores, Stats, News, Standings, Rumors – ESPN” http://espn.go.com/mlb/. Retrieved January 2012
- Hogg, R.V.; Craig, A.T.; and McKean, J.W., Introduction to Mathematical Statistics, Sixth Edition, Prentice Hall Inc, 2004.
- James, B. The Bill James Baseball Abstract, self-published, 1979.
- James, B. The Bill James Baseball Abstract, self-published, 1980.
- James, B. The Bill James Baseball Abstract, self-published, 1981.
- James, B. The Bill James Baseball Abstract, Ballantine Books, 1982.
- James, B. The Bill James Baseball Abstract, Ballantine Books, 1983.
- Jones, M. A. and Tappin, L. A, The Pythagorean Theorem of Baseball and Alternative Models, The UMAP Journal 26.2 (2005), 12 pages.
- Major League Baseball, Regular Season Standings | MLB.com: Standings (2010, October 3rd 3), MLB.com. Retrieved April 22, 2012, from http://mlb.mlb.com/mlb/standings/#20101003
- Miller, S.J., A Derivation of the Pythagorean Won-Loss Formula in Baseball, Chance Magazine 20 (2007), no. 1, 40--48. An abridged version appeared in The Newsletter of the SABR Statistical Analysis Committee 16 (February 2006), no. 1, 17--22, and an expanded version is available at http://arxiv.org/abs/math/0509698.

Steven J. Miller was partially supported by NSF Grant DMS0970067.

Kevin D. Dayaratna, kevind@math.umd.edu
Steven J. Miller, Steven.J.Miller@williams.edu , Steven.Miller.MC.96@aya.yale.edu ♦