By the Numbers

Volume 25, Number 1 The Newsletter of the SABR Statistical Analysis Committee May, 2015

Review

Academic Research: Umpire Status Bias

Charlie Pavitt

A recent academic study concludes that umpires are biased in favor of pitchers with established "status" -- those with numerous All-Star appearances, and those with an established history of good control.

Jerry W. Kim and Brayden G. King, Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring, Management Science, Vol. 60 No. 11, pp. 2619-2644

This is probably the best analysis of umpire bias to date. The basic argument is that umpires are predisposed toward favoring "high-status" pitchers. Umps are more likely to call "real" balls as strikes ("overrecognition" in the authors' terminology) and less likely "real" strikes as balls ("underrecognition") the higher the pitcher's status, with the bias accentuated for pitchers known to have good control.

To examine the argument's validity, all 2008 and 2009 pitches without batter swings were categorized via PITCHf/x data, with a long list of control measures gathered from various sources including Retrosheet. Status was based on number of All-Star appearances (which strikes me as a good index). Walks per plate appearance served as the measure of pitcher control.

The results: in total, overrecognition occurred on 18.8 percent of real balls and underrecognition on 12.9 percent of real strikes. Both over- and underrecognition were more likely for the home team, counts favoring the batter, later innings, high leverage plate appearances, more experienced pitchers, and, as hypothesized, pitchers with more All-Star appearances and better control.

The status effects were still apparent after adjusting for count, specific umpire, pitch location, and pitch type. All-Stars received a relative 6.7% reward in overrecognition and 5.7% bonus in underrecognition. Overrecognition also occurred for lefty batters and games with higher attendance.

In my view, the authors' argument seems to generalize to more experienced pitchers, who would have status for that reason alone. In addition, the results for attendance and home team are consistent with the most strongly supported explanation for the home-field advantage: crowd noise.

Similar biases were uncovered in favor of batters -- those with high status and demonstrated batting eyes. An analysis for catchers revealed different skill levels in pitch framing ability, which was not associated with All-Star catcher appearances; skill in pitch framing does appear less appreciated than it deserves.

Finally, overcoming a problem in past umpire bias research, an on-line unpublished version of the paper included individual differences among umps in both over- and underrecognition. The authors concluded that 80% of umps are guilty of the former and 64% of the latter. Interestingly, the two biases were largely independent, correlating at only -.16; additionally, the correlation goes the "wrong way."

Charlie Pavitt, chazzq@udel.edu ♦

In this issue

The previous issue of this publication was May, 2014 (Volume 24, Number 2).

BRJ: Do Hitters Overperform In Walk Years?

Phil Birnbaum

A recent paper in SABR's "Baseball Research Journal" found a boost to batter performance in the year preceding the negotiation of a new contract, presumably because the player is applying extra effort. Here, the author argues that the observed effect is likely spurious -- the result of a self-referential model, as well as a statistical artifact of the way the study handled aging.

A few years ago, I did a little study that looked to see if players boost their performance in their "contract year," the last year of their current contract. The idea is, players may be turning it up a notch, seeing as how their next year's salary depends heavily on their performance.

I concluded that the evidence showed no effect. Some players did seem to perform better, but others performed worse, and the overall effect was roughly a wash.[1]

In the Fall, 2014 issue of SABR's "Baseball Research Journal" (BRJ), Heather O'Neill challenges that conclusion.[2] She doesn't discuss my study directly -- I'm only one brief mention out of many citations, some of which also found no effect.

In her own study, O'Neill does find a significant effect -- that players in their walk year perform better than expected by 6 points of OPS+. I think that works out to somewhere around 3 runs for a full-time batter, or about a third of a win.[3]

Retiring Players

O'Neill actually finds that players in their contract year do *worse* than the others, with an OPS of 85.9, compared to 97.2. Of course, you can't rely on those raw numbers -- you need to correct for various factors. Obviously, you need to adjust for the talent of the players and their age, both of which are included in the model.

One additional adjustment that O'Neill chooses to make is to adjust for players about to retire from baseball. You can see how that might make sense. If a hitter is trying harder only in anticipation of an upcoming contract, you'd find the effect more easily by eliminating players who don't have that motivation.

But how do you figure that out? You can't just eliminate players who retired after that season, because they might be retiring only because they performed badly and nobody will offer them a contract. That would create the appearance of a false "contract year" effect -- the players who performed better than expected are included in the sample, but those who performed worse are eliminated.

To mitigate that bias, O'Neill adjusts not for whether a player actually retired, but for an estimated *probability* of that player retiring. Her paper actually includes two regressions -- the first estimates the probability of retirement, and the second uses that estimate in the equation that predicts OPS+ for the contract year.

But there's a problem with that approach. The bias winds up still there, because one of the factors that goes into the estimated probability of retiring is ... performance in that contract year season!

---

[1] My study appeared in Baseball Research Journal 35 (2006); a .pdf is available for download at http://research.sabr.org/journals/pdfs-brj/537-baseball-research-journal-volume-35. In addition, slides from my 2006 SABR presentation on the same study can be found at http://www.philbirnbaum.com/freeagent.ppt .

[2] Heather M. O'Neill, Ph.D, "*Do Hitters Boost Their Performance During Their Contract Years?*" The article can be downloaded from SABR at http://sabr.org/research/fall-2014-baseball-research-journal, in .pdf or HTML format.

[3] The paper is denominated in "OPS100," but that statistic appears to be identical to "OPS+" as listed at baseball-reference.com.

This makes for a model with an implicit circular argument, which goes something like this (all numbers made up by me, to illustrate the point):

> A 36-year-old career .800 OPS hitter hits for an .900 OPS in his contract year. That's a 100-point increase.
>
> Another 36-year-old career .800 hitter hits only .700 in his contract year. But, wait -- when a 36-year-old hitter hits only .700, he has a 50 percent chance of retiring after the season. And, from the regression, it turns out that players with a 50% chance of retiring after the season perform 50 points worse than others (perhaps because they're not motivated). Therefore, we're only going to count the .700 player as a 50-point shortfall, not a 100-point shortfall.
>
> Overall, then, the first player was +100, and the second was -50. The average is +50. So we have evidence that players hit better in their contract year!

That's the circular argument. Players are assumed to have hit badly because they were going to retire. How do we know they were going to retire? Because they hit badly!

Retirement, Age, and Performance

So, part of the reason the BRJ study found an effect is that the circular reasoning arbitrarily dismisses some of the counterevidence.

It's part of the reason, but actually not the biggest part. The main reason for the effect, I think, is the age adjustment. That's because the regression equation shows age (actually, years of MLB experience) as a much stronger indication of retirement than performance.

The effect really only kicks in for older players, but, when it does, it kicks *hard*.

The study tells us that Moises Alou had an estimated 60 percent chance of retiring after the 2008 season, when he had 19 years experience.[4]

In the second regression equation, the one that uses retirement in its estimate of OPS+, a 60 percent chance of retirement corresponds to an OPS+ drop of 60 points.[5] In other words, the regression says that a veteran of 19-years should be expected to hit for an OPS+ that's 60 points worse than his career mark. For a player who was league-average for his career, that's an expected OPS+ of only 40.

Well, 40 is truly awful ... it's obviously far, far too pessimistic a projection. You can't keep a job with an expectation of 40. Not even close. In 2014, the worst OPS+ for full-time hitter (enough PA to qualify for the batting title) was Zack Cozart's 61. (His OPS was .568.)

For a true 40, you have to go deep down the bench. Try Curt Casali, whose 2014 batting line looks like this:

	AB	R	H	2B	3B	HR	RBI	BB	SO	avg	obp	slg	OPS	OPS+
Casali	72	10	12	3	0	0	3	8	23	.167	.268	.208	.477	40

That's what the model is saying happens to an 19-year veteran of league-average career talent.

Of course, like most 19-year veterans, Moises Alou isn't average. For his career, he was a 126, which means the regression estimates his 19th year at 66. But 66 is obviously still unrealistically low -- it would still give Alou an expectation close to the worst regular in baseball.

For another example, take Derek Jeter. His career OPS+ was 115. With 19 years' experience, the regression would expect him to be at 55. In 2014, with 20 years' experience, it would project him even lower, probably below 50.

So, in 2014, when Jeter hit .256/.304/.313 for an OPS+ of 75, that would go into the "positive evidence that players hit better in contract years" bucket. That doesn't make sense.

---

[4] In 2008, Alou batted fairly well (OPS+ of 107), but only in 15 games. Neither of those numbers matter much -- the playing time and OPS barely affect the retirement estimate, which is almost all about age.

[5] I think it's coincidence that the equation works out so neatly, to almost exactly one point of OPS+ for each percentage point of retirement probability.

Age and Contract Status

What all this means is: the BRJ study badly underestimates the future performance of longtime veterans. By the regression models used, older players will almost always appear to have exceeded expectations.

The problem is: older players are significantly more likely to be in a contract year. That's because the older the player, the more likely he's already signed to a short-term contract, even a single-year contract. Teams don't give seven-year contracts to 38 year olds.

The paper's summary table confirms the confounding of age and contract status. The average age of contract-year players was 33.59 (11.6 years experience), while for non-contract years it was only 32.25 (10.6 years experience).

Does one year of experience make that much difference? It does. The study tells us that, from 14 to 16 years of experience, Bobby Abreu's probability of retirement went from 0 percent, to 13 percent, to 33 percent. That's a decline of about 13 points per season (some of the excess from 2010 to 2011 was related to a drop in performance). For players with more than 16 years, the decline would be significantly steeper, because the equation is quadratic.

Players should be expected to decline as they age, but by nowhere near as many as 13 points of OPS+. The observed effect was only 6 points. In that light, it seems reasonable, and even likely, that the observed "contract year" effect is really caused by a poor-fitting aging curve.

Summary

In a nutshell, I think this is what's going on:

- Older players are compared to unrealistically low projections.
- Older players are more likely to be in a contract year.

- Therefore, contract year players are compared to unrealistically low projections.

And that, I think, is why the study found what it did.

Phil Birnbaum, birnbaum@sympatico.ca ♦

Relieving and Readjusting Pythagoras

Victor Luo and Steven J. Miller

In a previous paper, it was found that modelling runs scored and allowed as Weibull distributions provides a statistical derivation of the Pythagorean projection. Here, the authors model the run distributions as a combination of two Weibull distributions, in an attempt to get a more accurate estimator while at the same time preserving a theoretical underpinning.

Introduction

The Pythagorean Win/Loss Formula, also known as the Pythagorean formula or Pythagorean expectation, was invented by Bill James in the late 1970s to predict a team's winning percentage using only its observed runs scored and runs allowed. The original formula is

$$winpct = \frac{RS^2}{RS^2 + RA^2}$$

The Pythagorean earned its name from the similarity of the denominator to the sums of squares in the Pythagorean formula from geometry. Later versions found better agreement by replacing the exponent 2 with values near 1.83; this optimal exponent was obtained by examining what value of the exponent had the best fit with the observed winning percentages of teams. On average, the formula is accurate to three or four games per team-season.

We will refer to the formula with exponent 2 as "Pythag(2.00)", the 1.83 version as "Pythag(1.83)", and the generalized version, with exponent γ, as "Pythag(γ)."

The formula Pythag(γ) is remarkably simple, requiring only runs scored and allowed, and the arithmetic is easily done on any calculator or phone. It is one of the most commonly cited estimators. One reason for its prominence is its accuracy in predicting a team's winning percentage through a simple calculation and not through computationally intense simulations. Additionally, it allows sabermetricians and fans to assess a manager's impact on a team, and estimate the value of new signings by projecting how their run production would change a team's overall record.

Because of the formula's widespread use and utility, increased accuracy would be especially beneficial. In his senior thesis, the first named author, supervised by the second named author, explored various attempted improvements to the Pythagorean formula. These included replacing observed runs scored and allowed with adjusted numbers, with the adjustments coming from a variety of sources (such as ballpark effects, game state, WHIP, ERA+, and WAR of the pitcher).

As these led to only minor improvements (we will discuss some of them later in this paper), we turned our attention to the successful theoretical model used by Miller et al. [DaMil1, DaMil2, Mil, MCGLP], who assumed runs scored and allowed were independently drawn from binned Weibull distributions with the same shape parameter. Recall the three parameter Weibull density is given by

$$f(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

The effect of α is to control the spread of the probability density, while β translates the distribution. The most important parameter is γ, which controls the shape. See Figure 1 for examples of how the shape of the Weibull distribution changes for several values of γ.

The success of the Weibull model is due to the fact that the three-parameter Weibull is a very flexible family of distributions, capable of fitting many unimodal distributions, including, to a statistically significant degree, the observed runs scored and allowed data.

Miller chose to use Weibulls for two reasons.

First, they lead to double integrals for the probabilities that can be evaluated in closed form . This is extremely important if we desire a simple expression such as Pythag(2.00) (or Pythag(γ)) posited by James[6] and is why we cannot use an arbitrary distribution. Not all distributions lead to integrations resulting in closed-form expressions; an important theoretical contribution of these earlier works is precisely the fact that utilizing a Weibull for theoretical modeling does produce closed-form answers.

Second, in addition to being flexible, certain values of the Weibulls correspond to well-known distributions. For instance, $\gamma = 1$ produces an exponential distribution, while $\gamma = 2$ is the Rayleigh distribution.

We explore the possibility of improving the predictive power of the theoretical model by modeling runs scored and allowed as being drawn from linear combinations of independent Weibulls. The advantage of this approach is that we are still able to obtain tractable double integrals which can be solved in closed form. There is a cost, however, as now more analysis is needed to find the parameters and the correct weights.

We will refer to this linear combination model as "Weib2," and the original single-Weibull model as "Weib1."

While Weib2 results in a more complicated formula than Pythag(γ), it is well worth the cost, as, on average, it is better by one game per team per season than Pythag(2.00) and roughly as accurate as Pythag(1.83). Specifically, a typical error for Weib2 is 3 games per team per year, as opposed to 4 games for Weib1.

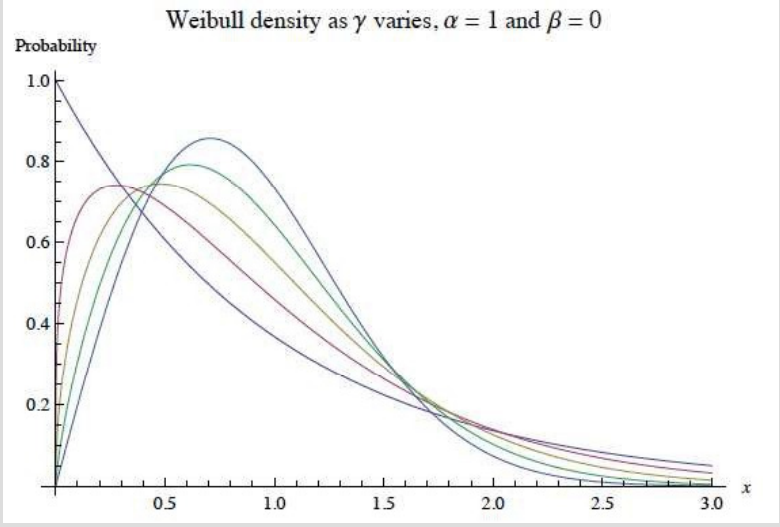

Figure 1 – Weibull distribution for various values of γ

Comparing the linear combination Weibull model to baseball-reference.com's calculated Expected W/L (equivalent to Pythag(1.83)) from 1979 to 2013, we found the Weib2 model to be approximately .06 of a game better. While this difference is not statistically significant, it is important to note the difference in how the respective predictions are made . Briefly (we will expand on this in a moment), in Pythag(1.83) one looks at the observed won-loss data over many seasons, and chooses the exponent that gives the best prediction; thus we are assuming a functional form. In Weib2, we model the runs scored and allowed as distributions rather than single datapoints, and use those distributions to predict a functional form, and eventually the winning percentages.

Though Miller showed in his paper that the Weib1 produces the Pythagorean formula Pythag(γ), using a combination of Weibulls (Weib2) produces a more general form in determining a team's winning percentage. At worst, it does as well Weib1, since we can produce Weib1 from Weib2 by setting the linear combination coefficient of one of the Weibulls to zero.

In addition, our work provides a theoretical justification for the general form. Although it is only slightly better at modeling observed values of teams' winning percentages, it is important to note that the win estimates given on sites such as ESPN and baseball-reference use the Pythagorean formula without theoretical rationale; by contrast, we supply a concise way of generalizing the Pythagorean formula with clear theoretical justification. We are also looking at more general combinations of Weibulls and thus expanding the quality of our modeling, as opposed to other methods such as Markov Chains and Taylor Series that are just fitting the data without giving the theoretical justification that we supply.

[6] See [HJM] for alternative simple formulas.

The increased accuracy of Weib2 leads to the conclusion that using a combination of a distribution in modeling runs provides an increased ability in predicting teams' winning percentages as opposed to using just a single distribution.[7] There are also compelling reasons to consider two Weibulls specifically, to model teams facing a left or right handed pitcher, or even on a cruder level, facing a strong or weak team.

Theorem

We now state our main result for the Weib2 model:[8]

Let runs scored and runs allowed per game be two independent random variables drawn from linear combinations of independent Weibull distributions with the same β's and γ's. Specifically, if W(t;α,β,γ) represents a Weibull distribution with parameters (α,β,γ), and we choose non-negative weights c_1, c_2, c'_1, and c'_2 where $c_1+c_2=1$ and $c'_1+c'_2=1$, then the density of runs scored, X, is

$$f\left(x;\alpha_{RS_1},\alpha_{RS_2},\beta,\gamma,c_1,c_2\right)=c_1W\left(x;\alpha_{RS_1},\beta,\gamma\right)+c_2W\left(\alpha_{RS_2},\beta,\gamma\right)$$

and runs allowed, Y, is

$$f\left(y;\alpha_{RA_1},\alpha_{RA_2},\beta,\gamma,c'_1,c'_2\right)=c'_1W\left(y;\alpha_{RA_1},\beta,\gamma\right)+c'_2W\left(\alpha_{RA_2},\beta,\gamma\right).$$

In addition, we choose α_{RS_1} and α_{RS_2} so that the mean of X is RS_{obs} and choose α_{RA_1} and α_{RA_2} so that the mean of Y is RA_{obs}. For $\gamma>0$, we have

$$\begin{aligned}&WonLostPercentage\left(\alpha_{RS_1},\alpha_{RS2},\alpha_{RA_1},\alpha_{RA_2},\beta,\gamma,c_1,c_2,c'_1,c'_1\right)\\&=c_1c'_1\frac{\alpha_{RS_1}^{\gamma}}{\alpha_{RS_1}^{\gamma}+\alpha_{RA_1}^{\gamma}}+c_1c'_2\frac{\alpha_{RS_1}^{\gamma}}{\alpha_{RS_1}^{\gamma}+\alpha_{RA_2}^{\gamma}}+c_2c'_1\frac{\alpha_{RS2}^{\gamma}}{\alpha_{RS_2}^{\gamma}+\alpha_{RA_1}^{\gamma}}+c_2c'_2\frac{\alpha_{RS_2}^{\gamma}}{\alpha_{RS_2}^{\gamma}+\alpha_{RA_2}^{\gamma}}\\&=\sum_{i=1}^{2}\sum_{j=1}^{2}c_ic'_j\frac{\alpha_{RS_i}^{\gamma}}{\alpha_{RS_i}^{\gamma}+\alpha_{RA_j}^{\gamma}}.\end{aligned}$$

This result also holds if γ <0; however, in that situation, the effect works in the wrong direction -- the more a team's runs scored exceeds their runs allowed, the worse the team's predicted record. This is due to the different shape of the Weibull. Here, as in applications of Weibulls in survival analysis, the shape parameter γ must be held to be positive.

[7] We also tried three Weibulls in combination, but there was no significant gain in doing so as opposed to using two; we suppose that as more distributions are added past two, a diminishing returns scenario takes place.

[8] A proof is available in [LM]. We leave the straightforward generalization to combinations of more Weibulls to the reader.

Curve Fitting

We now turn to finding the values of the parameters leading to the best fit. We require $\beta = -1/2$ (for binning purposes; this way all the integer scores of games are in the center of bins, and not at an edge), but otherwise the other parameters (the various α, γ, and c terms) are free.

Our first approach was to use the Method of Moments, in which we compute the number of moments equal to the number of parameters. Unfortunately, the resulting equations were too involved to permit simple solutions for them in terms of the observed data[9]. We thus turned to the Method of Least Squares (though one could also do an analysis through the Method of Maximum Likelihood).

We looked at all 30 teams from each season from 2004 to 2012. We implemented the Method of Least Squares, which involved minimizing the sum of squares of the error of the runs scored data plus the sum of squares of the error of the runs allowed data. Again, the α, γ, and c terms are free, except that the two c terms must sum to 1, and the two c' terms must also sum to 1.

For each team, we found the best fit linear combination of Weibulls (Weib2). Figure 2 shows the results for 2011.[10]

Using the Method of Least Squares, the mean γ over all 30 teams was 1.83 with a standard deviation of 0.18 (and median 1.79). We can see that the exponent 1.83, considered as the best exponent, is exactly our best fit value (and thus we provide theoretical justification for that exponent!).

Considering the absolute value of the difference between observed and predicted wins, we find a mean of 2.89 with a standard deviation of 2.34 (and a median of 2.68).[11]

These results are significant improvements on those obtained when using the Weib1 model to predict runs (which essentially reproduces James' original formula of Pythag(2.00), though with a slightly different exponent. That model produces a mean (absolute) error of 4.43 with standard deviation 3.23 (and median 3.54).

	Obs W	Pred W	Diff	γ
Boston	90	89.6	0.38	1.65
Yankees	97	100.3	-3.26	2.07
Baltimore	69	69.1	0.00	2.02
Tampa Bay	91	87.8	3.20	1.65
Toronto	81	79.7	1.31	1.99
Minnesota	63	63.5	-0.48	1.82
White Sox	79	77.6	1.42	1.72
Cleveland	80	77.4	2.57	1.76
Detroit	95	89.2	5.80	1.79
Kansas City	71	75.3	-4.33	2.05
Angels	86	84.9	1.14	1.66
Oakland	74	78.4	-4.41	1.63
Texas	96	95.6	0.42	1.64
Seattle	67	69.8	-2.78	1.66
Atlanta	89	85.3	3.73	1.74
Philadelphia	102	98.2	3.77	1.66
Florida	72	74.5	-2.52	2.02
Mets	77	78.8	-1.81	1.82
Washington	80	79.5	0.45	1.75
St. Louis	90	88.5	1.45	1.84
Houston	56	65.0	-8.96	1.82
Cubs	71	70.7	0.34	2.00
Cincinnati	79	82.9	-3.93	2.25
Pittsburgh	72	72.3	-0.29	1.80
Milwaukee	96	90.8	5.23	1.80
Dodgers	82	81.6	0.41	1.62
San Francisco	86	80.0	6.03	1.71
San Diego	71	77.1	-6.06	1.73
Colorado	73	75.8	-2.83	1.85
Arizona	94	86.8	7.17	2.23

Figure 2 – Results of predictions for 2011 team wins using linear combination of Weibull distributions

We compare the two sets of results in Figure 3. It is apparent that the Weib2 model better estimates teams' win/loss percentage; in fact, it is over one game better at estimating than the Weib1 model! The mean error for the Weib1 model was 4.22 (with a standard deviation of 3.03), while that of the Weib2 model was 3.11 (with a standard deviation of 2.33). In addition, there is less variation in the estimates. Thus, it appears that the Weib2 model provides a tighter, better estimate than does Weib1.

[9] For completeness, the equations are given in Appendix B of [LM] (or see [LUO]).

[10] Results from other seasons, and the code used to produce them, can be found in [Luo].

[11] Without considering the absolute value, the mean is 0.104 with a standard deviation of 3.75 (and a median of 0.39). However, in this paper, we concern ourselves only with the absolute value of the difference, which is the appropriate measure of how accurate our predicted values are.

We performed an independent two-sample t-test with unequal variances to see if the difference between the Weib1 model and Weib2 models is statistically significant. With a p-value less than 0.01, the difference is in fact highly significant.

For the seasons 1979 to 2013, we compared the mean errors of Weib2 to baseball-reference.com's Pythagorean Win-Loss statistic (pythWL, which uses the Pythag(1.83) formula). We display the results of our comparisons in Figure 4. The mean number of games off for the Pythag(1.83) formula was 3.09 with a standard deviation of 2.26, numbers only slightly worse than those of the Weib2 model (mean of 3.03 with standard deviation of 2.21). So, we can see that the Weib2 model is doing, on average, about .06 of a game better than the Pythag(1.83) formula. The difference is not statistically significant (again from an independent two-sample t-test with unequal variances).

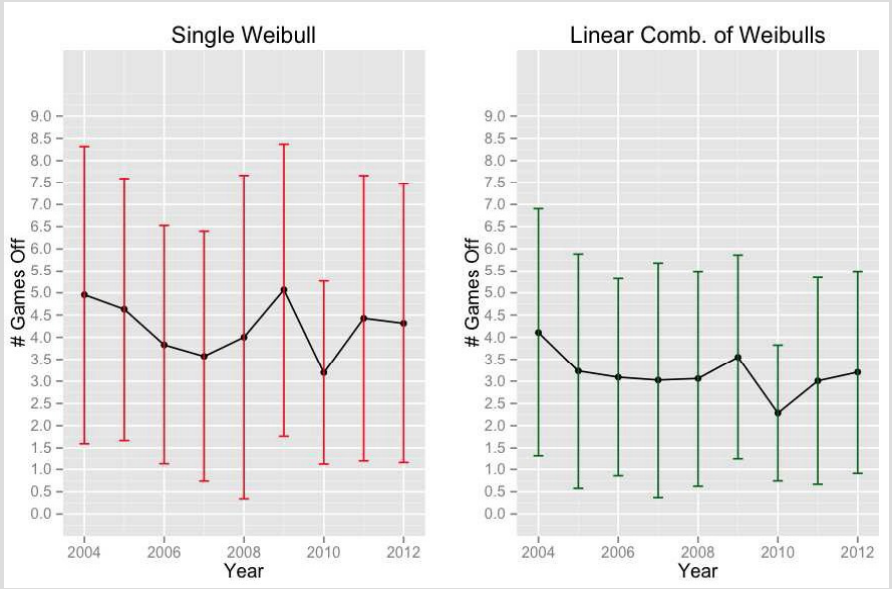

Figure 3 – Comparison of average errors for single Weibull vs. linear combination of Weibulls

In Figure 5, we show the results of a "competition" betwen Pythag(1.83) and Weib2. A bar extending above the axis shows Weib2 more accurate for the particular year; a bar below the axis shows Pythag(1.83) to be more accurate. The length of the bar is the "margin of victory."

The plot in Figure 5 visually suggests that, despite the small magnitude of the .06 difference, there might nonetheless be a slight preference for the Weib2. The difference between the two formulas seems to be a positive value for the most part, suggesting that the Weib2 model is doing slightly better than the Pythag(1.83) formula.

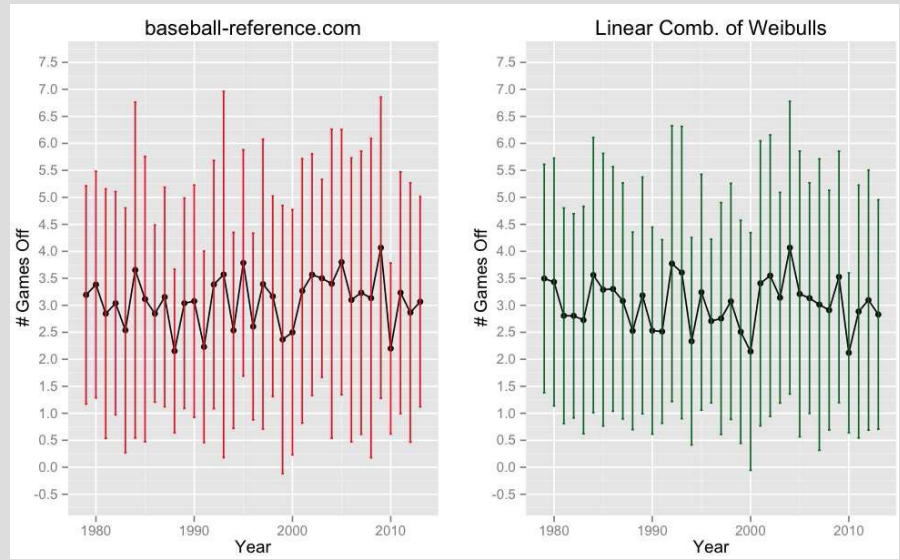

Figure 4 – Comparison of errors (with standard deviation) for Pythag(1.83) (from Baseball Reference) vs. linear combination of Weibulls (Weib2).

Specifically, and again looking at Figure 5, we can see that there are regions of the graph in which the the Pythag(1.83) formula does better, and parts where the Weib2 model does better. From 1979 to 1989, the Pythag(1.83) formula is more accurate, beating the Weib2 model in 7 out of the 11 years. However, from 1990 to 2013, the Weib2 model prevails in 15 out of the 24 years, and does so by around 0.3 games in those years. Furthermore, when the Pythag(1.83) formula does beat the Weib2 model in the years from 1990 to 2013, it does so by around 0.25 games, including the point at 2004, which seems very out of the ordinary. Without the point at 2004, the

Pythag(1.83) formula prevails by about 0.2 games in the years between 1990 and 2013 where it does beat the Weib2 model. Thus, in more recent years, it may make more sense to use the Weib2 model.

In addition, with respect to the standard error -- 2.26 for Pythag(1.83) versus 2.21 for Weib2 -- we can see that the Weib2 model provides on average a tighter fit, i.e., there is less fluctuation in the mean number of games off for each team in each year. From 1990 to 2013, the Pythag(1.83) formula's standard deviation in games off is 2.34 while that of the Weib2 model is 2.22, so we again see that the Weib2 model does noticeably better in recent years.

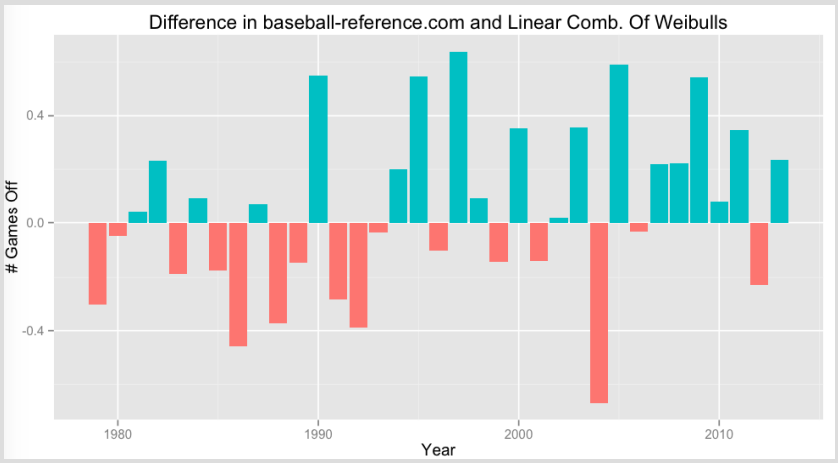

Figure 5 – Annual differences in prediction accuracy between Pythag(1.83) and Weib2

We performed chi-squared tests to determine the goodness of fit of the Weib2 model on the observed data, and test whether runs scored and allowed are independent. With a Bonferroni adjustment, the observed data closely follows the Weib2 model with the proper estimated parameters, and thus runs scored and allowed are acting as though they are statistically independent.[12]

Again, we argue that the choice between the two models should be based not just on fit, but, also on theoretical considerations. As noted previously, the pythWL statistic (Pythag(1.83)) just takes the functional form of the Pythagorean Win/Loss Formula with an exponent (γ) of 1.83, while we give theoretical justification for our formula (the Weib2 model).

Other Adjustments and Future Possibilities

Simplification

A one-game improvement in prediction is very promising, but the Weib2 model does require us to fit the runs scored and allowed distributions. So, we explored possibilities that might provide a simpler process.

Specifically, we tried to simplify the formula arising from the linear combination of two Weibulls, even giving up some accuracy, in order to devise a formula that could be easily implemented using just a team's runs scored and allowed (and the variance of each of these) in order to determine the team's winning percentage. Unfortunately, the individual team weight parameters play too much of a factor; in 2011, the mean of the parameter is 0.21 with a standard deviation of 0.39 (and a median of 0.21). With such large fluctuations in the weight parameters from team to team, the task of finding a simpler formula was almost impossible, as creating a uniform formula that every team could use was not feasible when two of the key parameters were so volatile.

Taking this into account, we tried fixing the γ, c_1, and c'_1 parameters, allowing for us to just solve a quartic involving the first and second moments to find the other parameters. However, while we were able to solve for the other parameters, the new formula gave us a significantly worse prediction of teams' win-loss percentage compared to the Weib2 model and Pythag(1.83).

One of the great attractions of James' Pythagorean formula (Pythag(2.00)) is its ease of use; we hope to return to other simplifications and approximations in a later paper. Our hope is to find a linearization or approximation of our main result, similar to how Dayaratna and Miller [DaMil1] showed the linear predictor of Jones and Tappin [JT] follows from a linearization of the Pythagorean formula.

[12] See [LM] for details.

Platoon

As we briefly mentioned earlier in the paper: if a team has an average hitting profile, what is their profile when facing left-handed versus right-handed pitchers? Could we fit the data for each case, and in turn weight the linear combination of Weibulls by how often we face and/or start a lefty or righty?

We suggest this as an avenue for further research.

Park Effects

One statistic that is easily accessible and has an obvious effect on runs is park effects. In sabermetrics, the established "park factor" statistic indicates whether a ballpark's conditions (shape, distances, altitude) increase or decrease run scoring. Park effects are measured relative to a baseline of 100, so a stadium with a park factor of 105 increases run scoring by 5 percent compared to the average.

To test whether park factors could increase the predictive power of the Pythagorean formula, we looked at the scores of each game in an MLB season. We then divided the scores by the park factor of the home stadium. So, if the Angels scored 4 runs and the Rangers scored 6 runs at Globe Life Park in Arlington (Texas Rangers Stadium), which has a ballpark factor of 118.3, their adjusted runs scored are 4/1.183 = 3.38 and 6/1.183 = 5.07, respectively.

We ran this model for seasons 2005 to 2012. For some seasons, predictive power increased; for others, it decreased. On average, there was a slight decrease in the correlation between predicted wins and actual wins -- from 0.8588 for the regular model with exponent 2, to 0.8558 for the park-adjusted model with exponent 1.83.

While at first this may seem strange, as the ballpark that a team is playing in should play a large role in determining the runs scored and allowed, we must remember that there are 2,430 games in an MLB season. In addition, the home and away team in a game are both affected by the attributes of the park. Taking these two factors into account, eventually it should make sense that the factors wash out over the entire season, thus leading to a predictive power at least on par with that of the original Pythagorean formula.

Game State

The next sabermetric statistic considered was game state. Game state essentially takes the situation that the game is in (inning, number of outs, bases occupied, and current score) and calculates the probability of each team winning the game. The probabilities are primarily based on the empirical record of past decades of baseball games.

The site *baseball-reference.com* calculates such probabilities, labeling them as "Winning Team Win Expectancy (wWE)."

Accounting for game state seems a reasonable way to improve on a win estimator. If a score progresses from, say, 9-0 in the bottom of the seventh to a final score of 12-0, the additional three runs don't contribute much to wins, and constitute mostly an empty inflation of total runs scored. For all practical purposes, when the game is at 9-0 at the bottom of the seventh, the game can almost be considered over, and the 9-0 score might be more reasonable to use when calculating the team's total runs scored and allowed at the end of the year.

In order to test whether this modification might improve predictive power, we decided to pick a "threshold" wWE. If the game progressed to that threshold percentage (for either team), we would consider the game over and use the scores at that state in the game as the final score.

Two thresholds were used, 95% and 98%. After the data was culled, we applied the original Pythagorean formula, and the formula accounting for ballpark factors.

As it turns out, the game-state-adjusted data does not have better predictive power than the original data; in fact, it is notably worse. For the regular Pythagorean formula, we obtained an adjusted R-squared value of 0.8145, while for the Pythagorean formula adjusted for game state at a 98% threshold and ballpark factors, the adjusted R-squared value was only 0.7261. Similar results held for all other years from 2005 to 2012.

This significant decrease in predictive power seems strange, as a threshold of 98% should give values almost similar to the original scores, with the exception of a few cases (the same decrease in power is evident in the case of a 95% threshold as well). This can probably be attributed to teams coming back more often than expected and the fact that we are throwing away good values that could evaluate a team.

So, for example, if a team is down 0-6 in the bottom of the seventh, so the game is considered almost surely over, but comes back to win 7-6, those 7 runs thrown out are good values that are important in evaluating the team. So, we have problems with not looking at the entire game.

We also tried a weighted game state approach, where we used the approach above, but pro-rated the runs scored and runs allowed to a full nine inning game. (For instance, a 10-2 lead after six innings would pro-rate to 15-3).

Unfortunately, again, the modification worsened the predictions, and the Pythagorean formula beat out the adjusted formula.

Specifically, using a weighted game state at 95% threshold gives an r-squared of 0.568, and that of a 98% threshold gives an r-squared of 0.574, a significant decrease from the 0.882 using the regular Pythagorean formula. Again, this can probably be attributed to teams coming back more often than expected and the fact that we are throwing away good values that could evaluate a team.[13]

Other Adjustments

We also tried adjusting for Walks plus Hits per Innings Pitched (WHIP), Adjusted ERA+ (ERA+), and Wins Above Replacement (WAR). However, adjusting for these statistics also did not increase the predictive power of the original Pythagorean formula, instead weakening it fairly significantly.[14] The fact that the predictive power of the regular Pythagorean formula with the original scores is at least on par or better than the adjusted formulas really demonstrates the robustness of the formula, and that in some cases, the simplest formula is in fact the best one.

Summary

Using the Weib2 model rather than the Weib1 model increases the prediction accuracy of a team's W/L percentage. Specifically, we saw that the Weib1 model's predictions for a team's wins were on average 4.22 games off (with a standard deviation of 3.03), while the Weib2 model's predictions for a team's wins were from 2004-2012 were on average 3.11 games off (with a standard deviation of 2.33), producing about a 25% increase in prediction accuracy.

A Chi-squared goodness of fit test for the Weib2 model confirmed the statistical independence of runs scored and allowed (a necessary requirement), and showed that in fact the Weib2 model with properly estimated parameters obtained from least squares analysis closely maps the observed runs scored and allowed.

When compared against the Pythag(1.83) formula, the Weib2 model improves the prediction by .06 of a game for the years 1979 to 2013. This improvement cannot be considered statistically significant. However, in more recent years, it is worth noting that it does appear that the Weib2 model is doing better than the Pythag(1.83) formula.

[13] See [Luo] for more detailed results on both sets of tests.
[14] Again, see [Luo] for details.

References

[DaMil1] K. Dayaratna and S. J. Miller, First Order Approximations of the Pythagorean Won-Loss Formula for Predicting MLB Teams Winning Percentages, By The Numbers – The Newsletter of the SABR Statistical Analysis Committee 22 (2012), no 1, 15–19.

[DaMil2] K. Dayaratna and S. J. Miller, The Pythagorean Won-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey (with Kevin Dayaratna), The Hockey Research Journal: A Publication of the Society for International Hockey Research (2012/2013), pages 193–209.

[HJM] C. N. B. Hammond, W. P. Johnson and S. J. Miller, The James Function, Mathematics Magazine 88 (2015) 54–71. http://xxx.tau.ac.il/ pdf/1312.7627v2.

[Ja] B. James, Baseball Abstract 1983, Ballantine, 238 pages.

[JT] M. A. Jones and L. A. Tappin, The Pythagorean Theorem of Baseball and Alternative Models, The UMAP Journal 26 (2005), no. 2, 12 pages.

[Luo] V. Luo, Relieving and Readjusting Pythagoras, Senior Thesis (supervised by S. J. Miller), Williams College 2014. http://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/st/VictorLuo.pdf.

[LM] V. Luo and S. J. Miller, Relieving and Readjusting Pythagoras (expanded version, 2014), http://arxiv.org/abs/1406.3402.

[Mil] S. J. Miller, A Derivation of the Pythagorean Won-Loss Formula in Baseball, Chance Magazine (2007), no. 1, 40–48 (an abridged version appeared in The Newsletter of the SABR Statistical Analysis Committee 16 (February2006), no. 1, 17–22, and an expanded version is available online at http://arxiv. org/pdf/math/0509698.pdf).

[MCGLP] S. J. Miller, T. Corcoran, J. Gossels, V. Luo and J. Porfilio, Pythagoras at the Bat, in: Social Networks and the Economics of Sports (edited by Victor Zamaraev), Springer-Verlag, 2014 (to appear).

[Mur] G. Muraleedharan, Characteristic and Moment Generating Functions of Three Parameter Weibull Distribution – an Independent Approach, Research Journal of Mathematical and Statistical Sciences 1 (2013), no. 8, 25–27.

Victor Luo, *victordan.luo@gmail.com*
Steven Miller, *sjmiller@math.brown.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.