

Approximate Conditional Inference for Evaluating Managerial Performance

David H. Annis*

Abstract

The so-called Pythagorean Theorem of Baseball presumes a relationship between a team's wins and losses and its runs scored and runs allowed. This method and related work fail to capture the nuances of the conditional behavior of wins given run totals. We present an alternative method which captures this dependence and allows for appropriate conditional inference. Because exact computations are impractical, we present an approximate method using Markov Chain Monte Carlo (MCMC). Comparisons and contrasts are drawn between the proposed method and established ones.

Keywords: Markov Chain Monte Carlo, Gibbs Sampler, Permutation Distribution, Baseball

*David H. Annis is Assistant Professor, Operations Research Department, Naval Postgraduate School, Monterey, CA 93943 (Email: annis@nps.edu).

1 Introduction

James (1986, p. 701) introduced a so-called *Pythagorean Theorem for Baseball* which states that the estimated win-loss record of a team is related to its runs scored and runs allowed in the following manner,

$$\text{Winning Percentage} = \frac{(RS)^2}{(RS)^2 + (RA)^2}$$

or, equivalently,

$$\log\left(\frac{W}{L}\right) = 2 \log\left(\frac{RS}{RA}\right) \quad (1)$$

where W and L are wins and losses, respectively, and RS and RA denote the total number of runs scored and runs allowed by the team in question over the course of the season. Equation (1) has a number of appealing properties, foremost among them *monotonicity* – i.e., for a fixed number of runs scored, the estimated winning percentage is decreasing in runs allowed; and for a fixed number of runs allowed, the estimated winning percentage is increasing in runs scored.

An intuitive consequence of this formulation is that a team that scores exactly as many runs as it allows would be expected to win as many games as it loses. However, while this holds for individual teams, it does not hold for the league as a whole. Despite the fact that league-wide, runs scored must equal runs allowed, Equation (1) does not constrain games won to equal games lost. For example, consider the 2002 Major League Baseball (MLB) season, for which the estimated number of wins under (1) is 2432 and the estimated number of losses is 2420, resulting in the peculiar conclusion that despite the zero-sum aspect of the game, collectively all teams are above average.

As an alternative to (1), Horowitz (1994a; 1994b) suggests fitting

$$\frac{W}{L} = \beta_1 \left(\frac{RS}{RA}\right) + \beta_2 \left(\frac{RS}{RA}\right)^2 + \epsilon \quad (2)$$

via ordinary least squares and using the sum of the estimated coefficients ($\hat{\beta}$) as a measure of a manager's effectiveness. Scully (1994) presents a slight variation of (1),

$$\log\left(\frac{W}{W+L}\right) = \beta_1 + \beta_2 \left(\frac{RS}{RA}\right) + \epsilon. \quad (3)$$

Both authors assume the errors are independent and identically distributed normal variates, $\epsilon \sim N(0, \sigma^2)$.

This Pythagorean winning percentage has become so well-accepted, in fact, that Baseball Reference publishes standings based on fitting (1) on its website (<http://www.baseball-reference.com>).

Each of these methods is a reasonable attempt to explain the propensity of a team to win (production output) based on its demonstrated ability to score and prevent its opposition from scoring (inputs to production). However, all suffer from the same shortcomings:

- Fitting equations such as (1), (2) and (3) to teams individually ignores certain collective constraints (such as equality of wins and losses).
- Using only the ratio of runs scored to runs allowed in predicting wins is clearly inappropriate. For example, suppose two teams play each other twice and, given the totals, runs are equally likely to be scored in either game. Consider the case when aggregate runs scored are 2 and 1. Since ties are not possible, each team must have won exactly one of the games. Therefore there is a 0% chance that the higher scoring team won both games. However, if the same two teams played twice and scored an aggregate of 20 and 10 runs (producing the same ratio of runs scored to runs allowed), the higher scoring team will win both games over 95% of the time! The Pythagorean estimate misses the mark in both cases, suggesting that in either situation, the higher scoring team should win 1.6 games (80%).

These deficiencies are the result of a philosophical disconnect between intent and result. Though these procedures attempt to explain wins *conditional on run totals*, none makes explicit use of this conditional relationship. In light of this, we propose a conditional procedure for win/loss inference based on runs scored and allowed. In contrast to Ruggiero et al. (1997), who present an identity, our procedure assumes only the aggregate runs for each team as well as the league-wide schedule.

The remainder of the paper is organized as follows. Section 2 explores the conditional behavior of wins and losses when runs scored and allowed are fixed at their observed values. Using this framework, methods are developed to assess the observed performance (wins) of teams based on their run totals. Results of the analysis are given in section 3. We conclude with an interpretation and discussion in section 4.

2 Conditional Inference

Since the goal of the Pythagorean method is to determine how many games a team *should* win based on its total runs scored and runs allowed, we propose a conditional method for inference. Specifically, contingent on all teams' runs scored, runs allowed and the league schedule, one can conceive of allocating both totals over all games played by all teams. Each possible permutation

of runs completely defines the number of wins and losses for each team. Conceptually, therefore, one could evaluate any quantity of interest based on this joint permutation distribution for game-specific runs. We choose to examine the average number of wins and the percentile of the observed wins (i.e., the probability that the team in question wins no more games than what was observed).

Not surprisingly, however, exhaustive enumeration of this distribution is unwieldy and impractical. Therefore, we propose an approximation to this distribution based on a Markov chain Monte Carlo (MCMC) (see, e.g., Robert and Casella, 1999) estimate of the parameters of interest.

Considering each game individually is problematic on two levels – first, the sheer number of games is overwhelming; and second, and more difficult to deal with, because ties are not allowed (they are broken by playing extra innings), permutations in which both opponents in a game accumulate the same number of runs are not allowed. Both problems are resolved by sampling from the permutation distribution of *total* runs scored and allowed for each pair of opponents. In addition to reducing the dimension of the problem, the issue of ties is resolved because each pair of opponents, if they play at all, plays at least two games against each other. Thus, it is perfectly acceptable for two opponents to accumulate the same aggregate run totals against each other.

In this setting, we consider total runs scored by team i against team j over the course of the season to be a Poisson random variable with mean proportional to the number of games between the two teams. It is well known (see, e.g., Agresti, 1990, p. 37) that when the sum over all opponents is held constant, this formulation is equivalent to multinomial sampling. All that remains is to sample from all multinomial models subject to each team’s runs scored and allowed constraints.

Although intuitive, this approach is difficult to implement due to the structural zeros (some teams never play others and therefore must not score or allow any runs against that particular opponent) and unbalanced schedule (other things being equal, a team will score and allow more runs against a frequent opponent than one faced only a few times). Features of the exact conditional joint distribution are exceedingly difficult to evaluate. However, by making use of a Gibbs sampler (see, e.g., Casella and George, 1992, for an explanation), random runs tables can be simulated from the joint distribution via Markov Chain Monte Carlo (MCMC).

Though tedious, it is straightforward to determine a set of “constraint” cells (cells, which given values of all others, are mathematically determined based on the marginal run totals). For each step in the Gibbs sampler, a new state is generated by sampling from the conditional joint distribution of a single unconstrained cell and all those constrained cells which it affects. McCullagh and Nelder (1989, p. 257) show that this is a non-central hypergeometric distribution. Since the support of

this distribution can be enumerated, generating a new value from the one-dimensional conditional distribution is easy. This procedure is repeated for all unconstrained cells, resulting in a new runs table.

The aforementioned procedure results in a multivariate Markov Chain with the desired stationary probabilities. Once the chain has been constructed, inference is conducted by treating the sample path as a (correlated) sample from the distribution of opponent-specific run totals. Subsequently, conditioned on the total runs scored and allowed between each pair of opponents, the expected number of wins and percentile of observed wins are both functions of the Markov chain, and can be evaluated numerically.

3 Results

We consider Major League Baseball (MLB) data from the 1998 through 2004 seasons. We chose 1998 as a starting point, as it was the last time MLB expanded (introducing the Tampa Bay Devil Rays and Arizona Diamondbacks). All seasons since then have involved the same competitors in the same league configuration.

For each season, permutations of runs between each pair of opponents were generated. Because selecting a feasible permutation is difficult, the chains were initialized at their observed values, and run for 10,000 steps, with the first 5,000 states being discarded as burn-in. (Assessing a chain's convergence to steady-state is non-trivial, and it is impossible to guarantee stability. We base our choice on that of Gelman et al. (1995, p. 295) who recommend discarding the first half of the runs and give as an example discarding 100 of 200 iterations.) Inference was based on the remaining 5,000 season permutations. Given the run totals, the average number of wins and the percentile of the observed wins can be evaluated numerically. Our approximation is based on 100 simulated seasons for each of the 5,000 permutations, resulting in a (correlated) sample of size 500,000. Given these samples, expected values are straightforward and standard errors can be calculated by examining "batch means" (see, e.g., Carlin and Louis, 1996, p. 172) to mitigate correlation. (Various batch sizes were examined, though we report those based on batches of length 50.)

The proposed conditional inference can be compared to the model-based approaches by comparing the excess number of wins over the predicted level for each team. Those winning more games than their run totals would suggest can be thought of as "well-managed" while those winning fewer are "managed poorly." In addition to win differential, because the conditional approach samples

from a distribution of possible win totals, it allows for estimating a percentile (probability of the observed wins or fewer) for each team. Teams with high observed percentiles (near one) get the most production (wins) from their materials (run scoring/defensing potential) while those with low observed percentiles (near zero) are less efficient.

On a gross level, predicted won/loss records based on Equations (1)-(3) are highly correlated to our conditional approach, owing to the fact that all are reasonable attempts to distill wins from scoring. Furthermore, it is not surprising that predictions based on Equations (1)-(3) are nearly identical as each takes a similar approach of fitting a (perhaps ad hoc) parametric model using the ratio of runs scored to runs allowed as the sole predictor of success (winning). Due to their similarity, comparisons will use James's Pythagorean formula as a surrogate for all such parametric methods. In addition to the striking similarities between methods based on Equations (1)-(3), there is perhaps an even stronger agreement between our metrics based on excess conditional wins and percentile of observed wins. In four of the seven seasons (1998, 1999, 2003 and 2004) both methods agree exactly on the ranking of team efficiency. In the three others (2000-2002) only two teams differ in rank and their rank difference is ± 1 in all cases. While we prefer the percentile method for assessing managerial performance, we will focus on win differential as it makes comparisons with other methods straightforward.

We hasten to note that there is a substantial disagreement between our method and its predecessors, *when it comes to evaluating efficiency* (i.e., number of wins above or below expected). Each season provides a noticeable number (between 3 in 2001 and 8 in 1998, with a mode of 6) of teams for which the Pythagorean method predicts more (or less) wins than observed, while the conditional inference indicates the opposite. Thus, if the difference between observed wins and expected wins is used to gauge managerial performance, the parametric models proposed by James, Horowitz and Scully are in disagreement with our conditional inference.

For each of thirty teams over seven seasons (1998-2004), we compute the difference between the observed and expected win totals, resulting in 210 team-seasons. Tables 1 and 2 gives the five best and five worst seasons in terms of excess wins. Teams in bold appear in the top (bottom) five based on both the Pythagorean and conditional methods.

One way to measure the effectiveness of a team's managerial structure is to determine where their win differential ranks among other teams. Because there is often managerial turnover – especially for poor teams – this measure evaluates an organization rather than an individual in the dugout. Table 3 gives the best and worst teams according to average rank of win differential. It is

Table 1: Comparison of the 5 Most Efficient Seasons.

Conditional		Pythagorean	
Team	Differential	Team	Differential
2004 Reds	+13.89	2004 Yankees	+11.57
1998 Royals	+13.78	2004 Reds	+10.21
2003 Reds	+11.78	1998 Royals	+9.73
2001 Mets	+10.70	2001 Mets	+9.47
2002 Tigers	+10.68	2002 Twins	+7.42

Table 2: Comparison of the 5 Least Efficient Seasons.

Conditional		Pythagorean	
Team	Differential	Team	Differential
2002 Red Sox	-11.80	1999 Royals	-10.62
1998 Astros	-10.57	2001 Rockies	-9.51
2003 Astros	-10.33	2000 Astros	-8.48
2001 Rockies	-9.92	2002 Red Sox	-8.29
1999 Diamondbacks	-9.18	2002 Cubs	-8.15

interesting that relatively successful teams (such as the Red Sox and Astros) are at the bottom of these lists, while underachievers such as the Devil Rays and Expos (now the Washington Nationals) are near the top.

It is important to stress that these rankings are based on win differential not absolute performance. For example the 2003 Tigers nearly set the record for most losses in a season, finishing 43-119. However, conditional inference suggests that they should have won only 40 games, so this represents “good management.” Similarly, the 2004 Red Sox won 98 games (and subsequently the World Series), however our approach indicates that 103 wins would be reasonable. Therefore, our inference (and many Red Sox fans) suggests that they won the World Series despite (manager) Terry Francona, not because of him. Another observation concerns the New York Yankees. The “Bronx Bombers” have been known in recent years for their high scoring offense. However, while the Pythagorean and Horowitz approaches conclude that they consistently win more than they should (even high scoring teams give up hundreds of runs and these methods are based on a ratio),

the conditional method pegs the Yankees at 14th place, suggesting that they win about as many as they should given their consistently high level of talent.

Table 3: Comparison of Average Win Differential Rank 1998-2004.

	Conditional	Pythagorean	Horowitz
1	Expos	Yankees	Yankees
2	Twins	Twins	Twins
3	Devil Rays	Braves	Braves
4	Brewers	Expos	Expos
5	Marlins	Rangers	Athletics
⋮	⋮	⋮	⋮
26	Braves	Red Sox	Phillies
27	Cardinals	Pirates	Red Sox
28	Mariners	Rockies	Rockies
29	Astros	Orioles	Astros
30	Red Sox	Astros	Orioles

4 Discussion

Many authors (James, 1986; Horowitz, 1994a,b; Scully, 1994) have conjectured relationships between a baseball team’s winning percentage and the ratio of its runs scored to runs allowed. While seemingly reasonable, these ad hoc procedures fail to capture the conditional behavior of wins given run totals. In addition, their formulation virtually ensures predictions for which league-wide wins and losses are not equal, though such situations are clearly impossible.

We have proposed a method for approximate conditional inference which accurately describes the conditional behavior of wins given run totals. As the exact permutation distribution is cumbersome, we give, instead, an approximation based on Markov Chain Monte Carlo techniques which allows us to sample from the distribution of interest.

References

- Agresti, A. (1990). *Categorical Data Analysis* (First ed.). New York: John Wiley & Sons.
- Carlin, B. P. and T. A. Louis (1996). *Bayes and Empirical Bayes Methods for Data Analysis* (Second ed.). Boca Raton, FL: Chapman Hall.
- Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician* 46(3), 167–174.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- Horowitz, I. (1994a). On the manager as principal clerk. *Managerial and Decision Economics* 15(5), 413–419.
- Horowitz, I. (1994b). Pythagoras, tommy lasorda and me: on evaluating baseball managers. *Social Science Quarterly* 75.
- James, B. (1986). *The Bill James Historical Baseball Abstract*. New York: Villard Books.
- McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (Second ed.). Boca Raton, FL: Chapman Hall.
- Robert, C. P. and G. Casella (1999). *Monte Carlo statistical methods*. Springer.
- Ruggiero, J., L. Hadley, G. Ruggiero, and S. Knowles (1997). A note on the pythagorean theorem of baseball production. *Managerial and Decision Economics* 18(4), 335–342.
- Scully, G. W. (1994). Managerial efficiency and survivability in professional sports teams. *Managerial and Decision Economics* 15(5), 403–411.