Edward H. Kaplan* and Candler Rich

Decomposing Pythagoras

https://doi.org/10.1515/jqas-2017-0055

Abstract: The Pythagorean win expectancy model developed by Bill James remains one of the most celebrated results in sports analytics. Many have extended the application of this model from its original use in baseball to other sports. Others have shown technical scoring conditions that imply the equivalence of win probability and the Pythagorean model. However, no explanation has been offered for *why* different sports yield different results beyond "that's what the data say." This article presents a theoretical analysis of the Pythagorean model by first deducing an exact within-team equation relating win percentage to seasonal scoring records, and then reconciling mathematically this result with the Pythagorean model which is cross-sectional across teams in a league. We derive a complete decomposition of the Pythagorean coefficient γ in terms of the exact model, and show that γ captures two key quantities – average points per game, and the average margins of victory and defeat – that together explain why different sports yield different results. We demonstrate this decomposition using the past decade of seasonal results from MLB baseball, NBA basketball, NFL football, and NHL hockey, and show that the data do reflect the properties deduced in our analysis.

Keywords: applied probability; Pythagorean method; regression decomposition; sports analytics.

1 Introduction

The Pythagorean win expectancy model, introduced by Bill James as a method for converting total runs for and against a baseball team to an estimate of that team's seasonal win percentage (James 1980), is one of the most celebrated and well-studied models in sports analytics. This model is covered in many sports analytics texts (e.g. Winston 2012; Severini 2015), and has been adapted to translate seasonal scoring to win/loss records in many sports other than baseball including basketball (Kubatko et al. 2007; Winston 2012; Kubatko 2013; Statis Ticator 2015), football (Schatz 2003; Winston 2012), and hockey (Cochran and Blackstock 2009). The Pythagorean win-expectancy model has also been adapted to study overtime in various sports (Rosenfeld et al. 2010). While the majority of published research regarding this model has focused on empirical issues such as parameter estimation and goodness-of-fit (e.g. Braunstein 2010), some have pursued more theoretical inquiries. Miller (2007) was the first to show that if the probability distributions of the number of points scored by opposing teams in games follow independent Weibull distributions, the resulting probability of winning a game corresponds to the Pythagorean model for the fraction of games won (see also Miller et al. 2014), while Dayaratna and Miller (2012) and Miller et al. (2014) produced a simple yet very accurate linear approximation of the Pythagorean model via its first-order Taylor series expansion. However, no explanation has been offered for *why* different sports yield different Pythagorean model results beyond "that's what the data say." This is unsatisfying. While scoring records do not translate to win percentages the same way in different sports, and though the Pythagorean model tracks such differences across sports empirically, the model should explain *why* such differences result in a manner that reflects known differences between sports.

The contribution of this paper is to offer just such an explanation, and provides the following insight: for a given sport, the single parameter of the Pythagorean model depends upon the typical scoring margins of victory and defeat in a game, in addition to the average number of points scored in total. Both scoring margins and point totals differ by sport. For example, while it is common to see a final score of 4-2 in baseball, it is extremely rare to see a score of 100-50 in basketball. The typical margins of victory and defeat in baseball are numerically close to the average number of runs scored in a game, whereas in basketball, the winning margins are much, much smaller than the number of points scored per game. In this paper, arguing from first principles, we will show how this argument is embedded in the Pythagorean model. In particular, we will show that the ratio of the Pythagorean coefficients for two different sports is approximately equal to the ratio of average points over average winning margin for the first sport divided by the average points-to-winning-margin ratio for the second sport. The different Pythagorean coefficients

***Corresponding author: Edward H. Kaplan,** William N. and Marie A. Beach Professor of Operations Research, Yale School of Management, New Haven, CT, USA; Professor of Public Health, Yale School of Public Health, New Haven, CT, USA; and Professor of Engineering, Yale School of Engineering and Applied Science, New Haven, CT, USA, e-mail: edward.kaplan@yale.edu
Candler Rich: Applied Mathematics Program, Yale University, New Haven, CT, USA, e-mail: lawrence.rich@yale.edu

for different sports thus tell a story, in that they reveal how scoring margins together with total points combine to produce winning records.

The remainder of the paper proceeds as follows: in the next section, we briefly review the mathematics of the Pythagorean model including its linearization via Taylor series. Following this, in Section 3 we derive an *exact* linear model from first principles for the win percentage of an individual team as a function of scoring for and against that team. This team-specific model requires no assumptions governing the probability distribution of scoring for or against, nor must we assume that total scoring by a team is independent of total scoring against a team. The only assumption required is that games cannot end in a tie. While this model is exact, it is team-specific, yet the Pythagorean model we seek to explain is of course cross-sectional across teams in a league. In Section 4, we reconcile this exact analysis with the first-order Taylor expansion of Pythagoras, which leads to a decomposition of the Pythagorean coefficient in terms of scoring margins and their relation to total points scored. We present examples of this decomposition for professional baseball, basketball, football and hockey in Section 5, while Section 6 concludes.

2 Pythagorean win expectancy model

The Pythagorean model as formulated by James (1980) related a team's seasonal win percentage (*WP*) to that team's total runs scored (*RS*) and runs against (*RA*) via

$$WP = \frac{RS^2}{RS^2 + RA^2}. \tag{1}$$

Note that one can divide both *RS* and *RA* by 162 (the number of games in a baseball season) without changing the left hand side, which allows interpreting *RS* and *RA* as the average number of runs per game scored for and against a team.

While James found that this formula worked well for baseball, he did not provide a theory or mechanism that resulted in this formula; rather his result stemmed from his remarkable ability to observe empirical regularities in baseball data. Lacking a justification for squaring both *RS* and *RA*, many realized that it was a simple matter to "tune" this formula to provide a better fit to observed data. This more general form of the Pythagorean model can be written as

$$WP = \frac{RS^\gamma}{RS^\gamma + RA^\gamma} \tag{2}$$

which of course reduces to the original model when $\gamma = 2$. Over time, baseball analysts have concluded that 1.83 represents a better value for γ (including Bill James apparently, see Davenport and Woolner 1999, and Miller et al. 2014 among others).

Equation (2) invites application to other sports, with the understanding that *RS* and *RA* now refer to the average number of *points* per game scored for and against a team in sports like basketball (Kubatko et al. 2007; Winston 2012; Kubatko 2013; Statis Ticator 2015) and football (Schatz 2003; Winston 2012), or goals per game in hockey (Cochran and Blackstock 2009). Not surprisingly, the relationship between scoring and winning is best described by different values of γ for different sports. For example, in basketball, $\gamma \approx 14$ (Kubatko 2013), while in football, $\gamma \approx 2.37$ (Schatz 2003). That scoring in baseball, basketball and football are different is clear to all, thus it is not surprising that the resulting estimates of γ also differ. Not clear, however, is *why* the γ's differ the way that they do. For example, why should basketball's γ be about 7.5 times higher than baseball's? We return to this question in Section 4.

2.1 Taylor series approximation

A key tool in our approach to understanding γ is the first-order Taylor series expansion of equation (2), previously reported by Dayaratna and Miller (2012) and Miller et al. (2014). Letting R_{total} denote the average number of runs per team per game over all games in a season (which roughly equals 4.3 for baseball, Miller et al. 2014), the first-order expansion of equation (2) is given by

$$WP \approx \frac{1}{2} + \frac{\gamma}{4 \times R_{\text{total}}} \times (RS - RA). \tag{3}$$

Dayaratna and Miller (2012) show that simple linear regressions of win percentage versus the difference between runs for and against across teams (see Jones and Tappin 2005 for such regressions) result in slope estimates that are extremely close to value of $\gamma/(4 \times R_{\text{total}})$, as must be the case if the Pythagorean model accurately captures the relationship between winning and scoring. Equation (3) also implies an important method for approximating γ. Imagine running the following simple linear regression using all teams in a season

$$WP = \alpha + \beta(RS - RA) + \varepsilon, \tag{4}$$

and obtaining the slope estimate $\widehat{\beta}$ (note that assuming all games are played, it must be that $\widehat{\alpha} = 1/2$, for the total number of runs scored by all teams equals the total number of runs scored against all teams, and over all

teams the average win percentage must equal 1/2). Then the Pythagorean coefficient can be estimated as

$$\widehat{\gamma} \approx 4 \times R_{\text{total}} \times \widehat{\beta}, \tag{5}$$

which will help unlock the puzzle of what the Pythagorean model is doing when it translates scoring to winning.

3 Exact win expectancy model

The Pythagorean model and its first-order approximation apply cross-sectionally across teams in a league across regular season play. In this section we shift our attention to an exact model for the win percentage of an individual team over the course of a season. For any particular team, consider a randomly selected game, and let random variable X denote the *spread*, that is, the difference between runs (or more generally points) for and against that team in a game. Note that we can estimate the mean spread per game at the end of a season by

$$E(X) \approx RS - RA \tag{6}$$

where as before we interpret RS and RA as the average runs (points) per game for and against the team in question.

In a game selected at random, the team of interest wins the game if and only if $X > 0$ (the team outscores its opponents), and conversely the team loses if $X < 0$. Our single assumption is that ties are not possible, that is, $X \neq 0$ (note that this assumption is also invoked in the Pythagorean model). Consequently, the probability that a team wins the game is given by

$$\Pr\{\text{Win}\} = \Pr\{X > 0\}, \tag{7}$$

and this win probability can be estimated by the teams seasonal win percentage, that is,

$$\Pr\{\text{Win}\} \approx WP. \tag{8}$$

Now, define a team's expected *margin of victory* by

$$MOV = E(X|X > 0), \tag{9}$$

and similarly define a team's expected *margin of defeat* by

$$MOD = -E(X|X < 0). \tag{10}$$

These definitions tell us, on average, by how much a team wins when it wins, and by how much a team loses when it loses. Each can be estimated simply from seasonal data: to estimate MOV for a given team, simply tally total runs scored minus runs against in games that the team of interest wins, and divide by the number of wins. To estimate MOD, tally total runs against minus runs scored, and divide by the number of losses.

With these definitions, we invoke the law of total expectation to write

$$\begin{aligned} E(X) &= E(X|X > 0)\Pr\{X > 0\} + E(X|X < 0)\Pr\{X < 0\} \\ &= MOV \times \Pr\{\text{Win}\} - MOD \times (1 - \Pr\{\text{Win}\}) \\ &= (MOD + MOV) \times \Pr\{\text{Win}\} - MOD \end{aligned} \tag{11}$$

and after dividing by $(MOD + MOV)$ and rearranging terms, we arrive at the desired result:

$$\Pr\{\text{Win}\} = \frac{MOD}{MOD + MOV} + \frac{1}{MOD + MOV} \times E(X). \tag{12}$$

This linear equation exactly relates the probability of a team winning to its expected point differential. Note that the derivation is completely general, and in particular does not require assuming particular probability distributions for the number of points scored for or against a team, or that such scoring be independent. The only assumption is that games do not end in a tie (that is, $\Pr\{X = 0\} = 0$).

Equation (12) is also exact for each of the n teams in the league after substituting team-specific seasonal estimates for the various parameters, that is

$$WP_i = \frac{mod_i}{mod_i + mov_i} + \frac{1}{mod_i + mov_i} \times (RS_i - RA_i) \tag{13}$$

where WP_i, RS_i and RA_i are the observed win percentage, runs scored and runs against while mod_i and mov_i are the observed average margins of victory and defeat respectively for the ith team, $i = 1, 2, \ldots, n$. Equation (13) is illustrated in Figure 1 for the 2016 MLB season (data from http://baseball-reference.com). In the figure, there are $n = 30$ straight lines, each one representing a different team. The intercept a_i and slope b_i for the ith team are given by

$$a_i = \frac{mod_i}{mod_i + mov_i} \tag{14}$$

and

$$b_i = \frac{1}{mod_i + mov_i}. \tag{15}$$

There are also 30 points, one on each line, that represent the exact win percentage and average run differential per game for each team. Note that while the intercepts differ across teams, the average of these intercepts is clearly close to 1/2. Also note that the slopes are quite close numerically, which implies that $mod_i + mov_i$ is roughly constant across teams, as implied by the nearly-parallel lines in Figure 1.

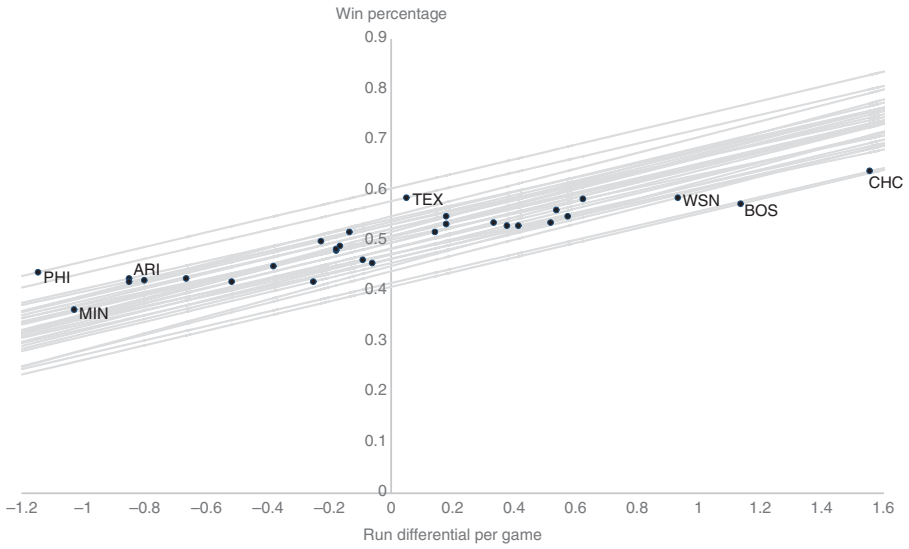

Figure 1: Exact win percentage for the 2016 major league baseball season.

It is tempting to compare equation (13) to the first order Taylor expansion of the Pythagorean model of equation (3); doing so suggests that

$$\begin{aligned}\frac{1}{2} + \frac{\gamma}{4 \times R_{\text{total}}} \times (RS_i - RA_i) \\ \approx \frac{mod_i}{mod_i + mov_i} + \frac{1}{mod_i + mov_i} \times (RS_i - RA_i)\end{aligned} \tag{16}$$

which in turn suggests that

$$\frac{mod_i}{mod_i + mov_i} \approx \frac{1}{2} \tag{17}$$

and

$$\frac{\gamma}{4 \times R_{\text{total}}} \approx \frac{1}{mod_i + mov_i} \tag{18}$$

for $i = 1, 2, \ldots, n$. However, this is not correct, for while equation (13) is exact on a team-by-team basis, equation (3) applies cross-sectionally across the teams. Indeed, as argued earlier, equation (3) can be thought of as the regression line through the 30 individual points in Figure 1; Figure 2 superpositions this regression line on Figure 1. As is clear, while the intercept of this line equals 0.5 as indeed it must, the slope is attenuated from the team-specific values. Still, equation (18) suggests that the Pythagorean parameter γ depends upon scoring margins in addition to total scoring. The question is how to move from the within-team exact model to the Pythagorean model which is cross-sectional across teams. We address this in the next section.

4 Decomposing Pythagoras

Our approach to reconciling the exact and linearized Pythagorean model proceeds by substituting the exact equation (13) for each team on the left-hand side of equation (4), solving analytically for the estimated regression slope $\widehat{\beta}$ in terms of exact model properties, and using equation (5) to arrive at the Pythagorean parameter estimate $\widehat{\gamma}$. To simplify notation, let: $x_i = RS_i - RA_i =$ observed average run differential per game over the course of a season for the ith team; $y_i = WP_i =$ seasonal win percentage for the ith team; and recall the definitions of a_i and b_i from equations (14, 15).

Now, as is well known, the estimated regression slope $\widehat{\beta}$ in the model $E(Y) = \alpha + \beta X$ is given by

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2} \tag{19}$$

where

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{20}$$

is the sample covariance between run differential and win percentage, and $s_x^2 = s_{xx}$ is the sample variance of run differential. However, owing to the exact model, for any team i we have

$$y_i = a_i + b_i x_i \text{ for } i = 1, 2, \ldots, n \tag{21}$$

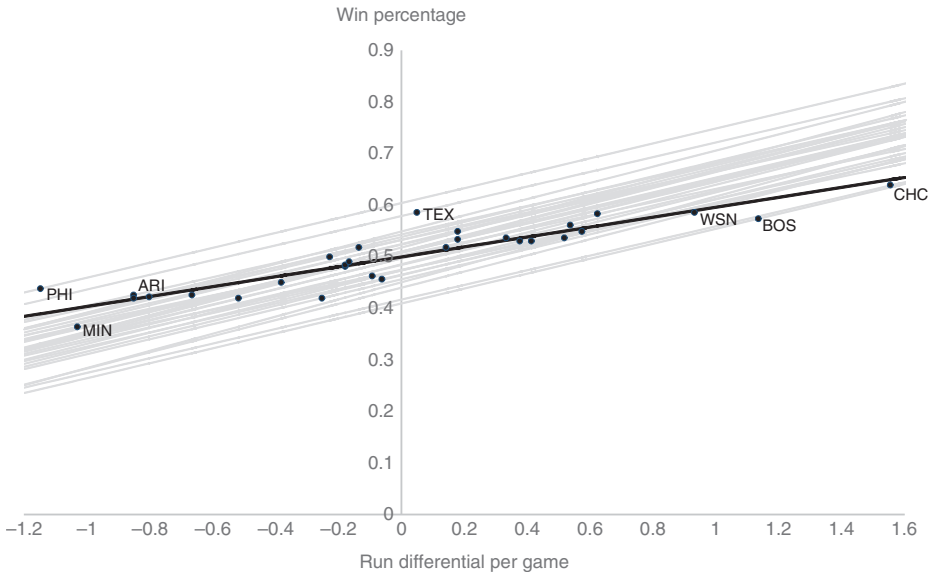

Figure 2: First-order Pythagorean and exact win percentage models.

which, upon substitution into equation (19), yields

$$\begin{aligned}\widehat{\beta} &= \frac{s_{x,a+bx}}{s_x^2} \\ &= \frac{s_{ax} + \overline{b}s_x^2 + s_{b,x^2}}{s_x^2} \\ &= \overline{b} + \hat{\beta}_{a|x} + \hat{\beta}_{b|x^2}\frac{s_{x^2}^2}{s_x^2}\end{aligned} \tag{22}$$

where in deriving this result we have used the fact that $\overline{x} = 0$ owing to the equality over all teams of points scored for and against. To understand the terms on the right-hand side of equation (22), note that

$$\overline{b} = \frac{1}{n}\sum_{i=1}^{n} b_i, \tag{23}$$

the empirical average of the exact model slopes across all teams. The second term on the right-hand side of equation (22) follows from recognizing that the ratio s_{ax}/s_x^2 is exactly the estimated slope for the regression of the a_i's against the x_i's, and captures how the ratio of $mod_i/(mod_i + mov_i)$ changes with run (score) differential x_i across teams. The third term follows from noting that

$$\frac{s_{b,x^2}}{s_x^2} = \frac{s_{b,x^2}}{s_{x^2}^2} \times \frac{s_{x^2}^2}{s_x^2} \tag{24}$$

and that $s_{b,x^2}/s_{x^2}^2$ is exactly the estimated slope for the regression of the b_i's against the x_i^2's. Substituting equation (22) into equation (5) finally yields

$$\widehat{\gamma} \approx 4 \times R_{\text{total}} \times \left(\overline{b} + \hat{\beta}_{a|x} + \hat{\beta}_{b|x^2}\frac{s_{x^2}^2}{s_x^2}\right). \tag{25}$$

Equation (25) clarifies how the Pythagorean parameter depends upon the scoring characteristics of whatever sport is in question. Other things being equal, we see that not only does $\widehat{\gamma}$ increase with the average number of points scored per game (R_{total}), it also increases with $\overline{b}$, which itself declines with scoring margin. The term $\widehat{\beta}_{a|x}$ is the rate with which the ratio of mod_i to $mod_i + mov_i$ changes with score differential x_i across teams, and as can be seen from Figures 1 and 2, teams with higher average net scores (higher values of $x_i = RS_i - RA_i$) have *lower* values of $mod_i/(mod_i + mov_i)$. This implies that $\widehat{\beta}_{a|x} < 0$. Simply stated, better teams have higher average net scores, larger margins of victory (so when they win they win by more), and smaller margins of defeat (so when they lose they lose by less). Finally, as we will demonstrate numerically in the next section, but as can also be inferred from Figures 1 and 2, the term $\widehat{\beta}_{b|x^2}$ essentially equals zero. This follows from noting that the lines in Figure 1 are essentially parallel, meaning that the slopes b_i in the exact model are essentially invariant with score differential (x_i), and hence also invariant with x_i^2.

These points are illustrated graphically in Figure 3 for the 2016 MLB season. There are three gray lines in Figure 3, each denoting a different contribution to the (first-order) Pythagorean win percentage. The horizontal line $\alpha = 0.5$ sets the average win percentage. The increasing line $\overline{b}x$ shows how win percentage increases with average run differential at a slope that depends upon margins of victory and defeat; this is essentially the average of the 30 lines in Figures 1 and 2 subtracted from 1/2. The decreasing line $\widehat{\beta}_{a|x}x$ shows how win percentage declines on account of the declining ratio of $mod_i/(mod_i + mov_i)$ with run

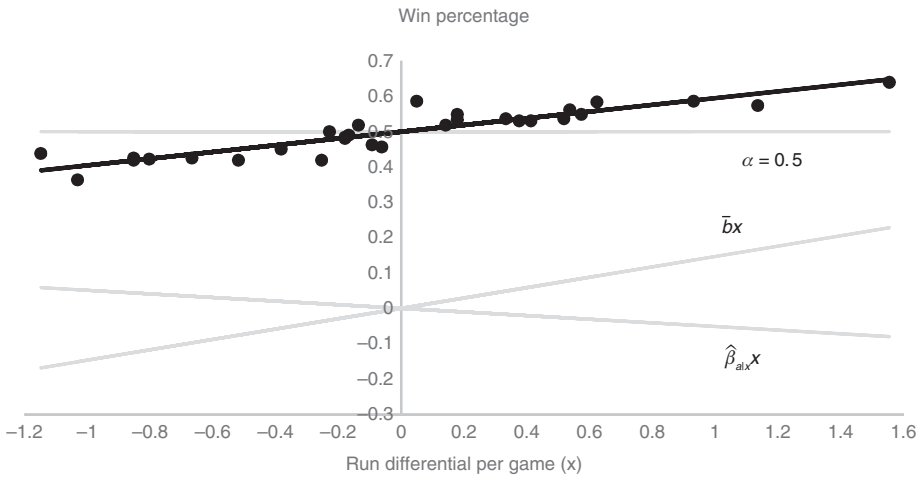

Figure 3: Decomposing Pythagoras.

differential. This line captures the effect of hopping from higher to lower lines in Figures 1 and 2 as run differential increases. As discussed above, we ignore $\widehat{\beta}_{b|x^2} \approx 0$. Adding the three three gray lines together yields the black line that represents the (first-order) Pythagorean model, and as was already shown, this line provides an excellent fit to the observed win percentages, shown as black dots in Figure 3.

5 Application to baseball, basketball, football and hockey

Having worked through the theory of decomposing Pythagoras, it is a simple matter to apply this decomposition to different seasons in different sports. Our purpose in doing so is to both apply our decomposition results to different sports, but also to see what the Pythagorean coefficients tell us about scoring in different sports. We will discover empirically that the ratio of the Pythagorean coefficients across two sports approximately equals the ratio of $R_{\text{total}} \times \overline{b}$ across these same sports. As $R_{\text{total}} \times \overline{b}$ depends upon both average points scored and margins of victory and defeat, we have a way of understanding why some sports have larger Pythagorean coefficients compared to others.

Table 1 reports results for the past ten MLB seasons (data from http://baseball-reference.com). There are several points worth noting in these results. First, all components of the decomposition are quite stable: plotting $\widehat{\beta}$, $\overline{b}$, $\widehat{\beta}_{a|x}$ and $\widehat{\beta}_{b|x^2}\frac{s_{x^2}^2}{s_x^2}$ over time would yield four nearly flat lines. Second, as argued in the last section, the rate $\hat{\beta}_{A|X}$ with which $mod_i/(mod_i + mov_i)$ changes with net scoring x_i is negative. Third, the term $\widehat{\beta}_{b|x^2}\frac{s_{x^2}^2}{s_x^2}$ is very small in absolute value compared to $\widehat{\beta}$ (at most 5%), and is often within two standard errors of zero. Fourth, in absolute value, $\widehat{\beta}_{a|x}$ is about 30% of $\overline{b}$, which suggests a further approximation for MLB:

$$\widehat{\gamma} \approx 4 \times R_{\text{total}} \times 0.7 \times \overline{b}. \tag{26}$$

Table 1: Pythagorean Decomposition Results for Major League Baseball.

Year	$\widehat{\beta}$	se	$\bar{b}$	se	$\widehat{\beta}_{a\|x}$	se	$\widehat{\beta}_{b\|x^2}\frac{s_{x^2}^2}{s_x^2}$	se	R_{total}	$\widehat{\gamma}$	se
2007	0.0853	0.0079	0.1408	0.0020	−0.0526	0.0077	−0.0029	0.0022	4.80	1.64	0.15
2008	0.1039	0.0082	0.1439	0.0016	−0.0424	0.0081	0.0025	0.0016	4.65	1.93	0.15
2009	0.1063	0.0085	0.1428	0.0014	−0.0360	0.0092	−0.0004	0.0014	4.61	1.96	0.16
2010	0.0932	0.0041	0.1485	0.0017	−0.0552	0.0043	−0.0001	0.0021	4.38	1.63	0.07
2011	0.1035	0.0077	0.1534	0.0019	−0.0446	0.0072	−0.0053	0.0022	4.28	1.77	0.13
2012	0.1124	0.0072	0.1529	0.0021	−0.0385	0.0072	−0.0020	0.0021	4.32	1.94	0.12
2013	0.1025	0.0064	0.1540	0.0019	−0.0467	0.0054	−0.0048	0.0019	4.17	1.71	0.11
2014	0.1099	0.0088	0.1543	0.0019	−0.0412	0.0089	−0.0031	0.0020	4.07	1.79	0.14
2015	0.0972	0.0097	0.1492	0.0018	−0.0520	0.0093	0.0000	0.0024	4.25	1.65	0.17
2016	0.0936	0.0081	0.1465	0.0017	−0.0512	0.0084	−0.0016	0.0022	4.47	1.67	0.15

Table 2: Pythagorean Decomposition Results for NBA Basketball.

Year	$\widehat{\beta}$	se	$\bar{b}$	se	$\widehat{\beta}_{a\|x}$	se	$\widehat{\beta}_{b\|x^2}\frac{s_{x^2}^2}{s_x^2}$	se	R_{total}	$\widehat{\gamma}$	se
2007	0.0326	0.0019	0.0485	0.0008	−0.0144	0.0020	−0.0015	0.0009	98.74	12.87	0.77
2008	0.0299	0.0012	0.0461	0.0008	−0.0161	0.0013	0.0000	0.0007	99.92	11.96	0.48
2009	0.0355	0.0009	0.0477	0.0010	−0.0099	0.0009	−0.0022	0.0010	99.95	14.19	0.37
2010	0.0337	0.0015	0.0470	0.0009	−0.0126	0.0016	−0.0007	0.0009	100.45	13.56	0.59
2011	0.0331	0.0015	0.0494	0.0007	−0.0159	0.0017	−0.0004	0.0007	99.55	13.17	0.60
2013	0.0327	0.0015	0.0472	0.0006	−0.0135	0.0014	−0.0010	0.0007	98.22	12.85	0.61
2014	0.0322	0.0016	0.0478	0.0009	−0.0141	0.0015	−0.0016	0.0010	101.01	12.99	0.64
2015	0.0330	0.0013	0.0470	0.0008	−0.0139	0.0015	−0.0001	0.0010	100.01	13.21	0.54
2016	0.0322	0.0013	0.0471	0.0008	−0.0118	0.0016	−0.0031	0.0010	102.67	13.22	0.51

The importance of this approximation is that it reveals the dependence of the estimated Pythagorean parameter on two key sports quantities: the average number of points per game (R_{total}) and the winning margin as expressed via $\overline{b}$ (which is the average of the reciprocals of $mod_i + mov_i$).

Table 2 reports the results of applying our decomposition to NBA seasons since 2007 (data from http://basketball-reference.com; we have omitted the strike-shortened 2012 season when only 66 instead of 82 games were played). Similar to the baseball results from Table 1, we see stability in all elements of the decomposition over time, $\widehat{\beta}_{a|x} < 0$ for all years, and $\widehat{\beta}_{b|x^2}\frac{s_{x^2}^2}{s_x^2}$ is again very small in absolute value compared to $\widehat{\beta}$ and often statistically not different from zero. But perhaps most interesting, we see that in absolute value, $\widehat{\beta}_{a|x}$ is again about 30% of $\overline{b}$ (the average ratio is 28.5%), which means that equation (26) is also quite accurate for basketball. This suggests something quite fundamental about how scoring translates to winning in basketball versus baseball, for via equation (26), the ratio of the Pythagorean parameter for basketball to baseball should approximately equal the ratio of $R_{\text{total}} \times \overline{b}$ for basketball to baseball. While R_{total} is the average number of points scored per game, $\overline{b}$ can heuristically be thought of as the reciprocal of the average winning margin (it is really the average of the reciprocals rather than the reciprocal of the average). This means that the Pythagorean gammas should roughly be in proportion to the ratio of scoring to scoring margin for baseball and basketball. From Table 1, we see that for baseball the average values for $\widehat{\gamma}$, $\overline{b}$ and R_{total} equal 1.77, 0.1486, and 4.40 respectively. From Table 2, the same quantities for basketball average 13.11, 0.0475, and 100.06. The ratio of the average $\widehat{\gamma}$'s for basketball to baseball thus equals $13.11/1.77 = 7.41$. The ratio of the product of the average $\overline{b}$ and R_{total} for basketball to baseball equals $(0.0475 \times 100.06)/(0.1486 \times 4.4) = 7.27$, which is very close. What the Pythagorean model seems to be saying is that the way scoring translates to winning depends upon two key quantities: the average number of points per game, and the average winning margin (more precisely, the sum of the average margins of victory and defeat).

Table 3 presents results for NFL football seasons since 2007 (data from http://pro-football-reference.com). The same observations again hold: stability in all components of the decomposition over time, $\widehat{\beta}_{a|x} < 0$ for all years, and truly insignificant values of $\widehat{\beta}_{b|x^2}\frac{s_{x^2}^2}{s_x^2}$ in all years. However, the absolute ratio of $\widehat{\beta}_{a|x}/\overline{b}$ equals about 37% for football, which is larger than the 30% found for baseball and basketball. Still, it is inviting to see how the Pythagorean coefficient ratios for football and other sports compare to the ratios of average $\overline{b}\times$ average R_{total}. For football, the average values for $\widehat{\gamma}$, $\overline{b}$ and R_{total} equal 2.51, 0.0464 and 22.37 respectively. Comparing football to baseball, we see that the ratio of the average $\widehat{\gamma}$'s equals $2.51/1.77 = 1.42$, while the ratio of the product of the average $\overline{b}$ and R_{total} for football to baseball equals $(0.0464 \times 22.37)/(0.1486 \times 4.4) = 1.59$, which is close to the ratio of the Pythagorean coefficients. Comparing basketball to football, the ratio of the average $\widehat{\gamma}$'s equals $13.11/2.51 = 5.22$, while the ratio of the average $\overline{b} \times$ average R_{total} for basketball to football is given by $(0.0475 \times 100.06)/(0.0464 \times 22.37) = 4.58$, which is less close but still in the ballpark (or the court).

Finally, Table 4 presents results for NHL hockey seasons since 2007 (data from https://www.hockey-reference.com/). Again one sees stability in all components of the decomposition over time, $\widehat{\beta}_{a|x} < 0$ for all years, and insignificant values of $\widehat{\beta}_{b|x^2}\frac{s_{x^2}^2}{s_x^2}$ in all years. The absolute ratio of $\widehat{\beta}_{a|x}/\overline{b}$ averages about 26% for hockey, which is closer to the 30% ratio found for baseball and basketball than the 37% ratio for football. Taking the ratio of the average Pythagorean coefficient for hockey (2.03) to the same for baseball, basketball and football respectively yields 1.15, 0.15 and 0.81. Now taking the ratio

Table 3: Pythagorean Decomposition Results for NFL Football.

Year	$\widehat{\beta}$	se	$\bar{b}$	se	$\widehat{\beta}_{a\|x}$	se	$\widehat{\beta}_{b\|x^2}\frac{s_{x^2}^2}{s_x^2}$	se	R_{total}	$\widehat{\gamma}$	se
2007	0.0268	0.0018	0.0448	0.0014	−0.0192	0.0019	0.0012	0.0020	21.69	2.33	0.15
2008	0.0280	0.0023	0.0443	0.0017	−0.0154	0.0027	−0.0009	0.0022	22.03	2.47	0.20
2009	0.0251	0.0018	0.0431	0.0019	−0.0178	0.0022	−0.0002	0.0021	21.47	2.15	0.16
2010	0.0280	0.0024	0.0469	0.0021	−0.0163	0.0025	−0.0026	0.0025	22.04	2.47	0.21
2011	0.0273	0.0021	0.0460	0.0018	−0.0186	0.0021	−0.0001	0.0023	22.18	2.43	0.19
2012	0.0255	0.0020	0.0463	0.0020	−0.0198	0.0022	−0.0010	0.0023	22.76	2.32	0.18
2013	0.0280	0.0018	0.0496	0.0018	−0.0198	0.0017	−0.0018	0.0019	23.41	2.62	0.17
2014	0.0301	0.0019	0.0416	0.0012	−0.0115	0.0022	0.0000	0.0013	22.59	2.72	0.17
2015	0.0293	0.0020	0.0497	0.0019	−0.0169	0.0021	−0.0035	0.0019	22.81	2.68	0.18
2016	0.0322	0.0034	0.0513	0.0018	−0.0157	0.0035	−0.0035	0.0026	22.78	2.93	0.31

Table 4: Pythagorean Decomposition Results for NHL Hockey.

Year	$\widehat{\beta}$	se	$\bar{b}$	se	$\widehat{\beta}_{a\|x}$	se	$\widehat{\beta}_{b\|x^2}\frac{s_{x^2}^2}{s_x^2}$	se	R_{total}	$\widehat{\gamma}$	se
2007	0.1755	0.0141	0.2449	0.0037	−0.0689	0.0141	−0.0005	0.0054	2.78	1.95	0.16
2008	0.1830	0.0123	0.2497	0.0026	−0.0701	0.0123	0.0034	0.0033	2.91	2.13	0.14
2009	0.1852	0.0124	0.2497	0.0028	−0.0642	0.0128	−0.0004	0.0038	2.84	2.10	0.14
2010	0.1727	0.0135	0.2448	0.0024	−0.0673	0.0137	−0.0048	0.0026	2.79	1.93	0.15
2011	0.1697	0.0146	0.2501	0.0030	−0.0758	0.0146	−0.0046	0.0030	2.73	1.86	0.16
2012	0.1783	0.0133	0.2546	0.0035	−0.0724	0.0139	−0.0039	0.0059	2.73	1.95	0.15
2013	0.1865	0.0095	0.2562	0.0036	−0.0682	0.0094	−0.0015	0.0046	2.74	2.04	0.10
2014	0.1857	0.0116	0.2536	0.0035	−0.0627	0.0122	−0.0052	0.0058	2.73	2.03	0.13
2015	0.2092	0.0114	0.2481	0.0029	−0.0369	0.0116	−0.0020	0.0026	2.71	2.27	0.12
2016	0.1810	0.0075	0.2502	0.0033	−0.0707	0.0078	0.0016	0.0049	2.77	2.00	0.08

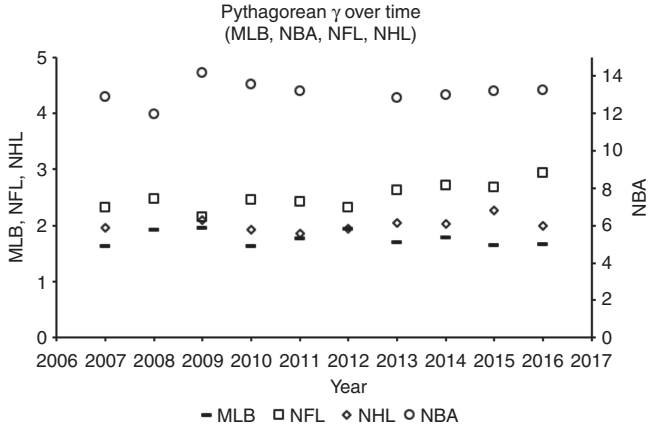

Figure 4: Pythagorean coefficients (γ's) by sport over time.

of average $\bar{b}$ × average R_{total} for hockey (0.69) to the same for baseball, basketball and football respectively yields 1.06, 0.15 and 0.67. These calculations suggest that while the Pythagorean "story" relating win percentage to both scoring totals and margins of victory and defeat works well when comparing baseball, basketball and hockey, the results are less satisfying for football.

Figure 4 plots the estimated Pythagorean γ's for baseball, football, hockey (left vertical axis) and basketball (right vertical axis) over time. While there is some year-to-year variation, by sport the magnitudes of these coefficients are quite stable over time. As explained by our decomposition, the Pythagorean model does capture the differences between scoring (and scoring margins) in different sports.

6 Summary

The Pythagorean win expectancy model remains one of the most celebrated tools in sports analytics, and while many have documented its ability to approximate win percentages from seasonal scoring records, very little has been written in sports-specific terms regarding *what* this model really does, and *why* it produces different results for different sports. Here we have offered an original explanation: the single coefficient in the Pythagorean model effectively captures the ratio of average scoring to the sum of the margins of victory and defeat. These characteristics clearly differ by sport, and the Pythagorean coefficient estimates for different sports capture this difference. We discovered this result from first principles – we first derived an exact within-team model relating win

percentage to scoring differential, and mathematically reconciled this model with the Pythagorean model which is cross-sectional across teams – and showed that at least for baseball, basketball, football and hockey, observed data give rise to Pythagorean coefficient estimates that agree with our analytical claims.

It is remarkable that Bill James deduced the Pythagorean model for baseball based solely on observing empirical patterns in the data. James is well known for having intuited many things about sports more generally. In deference to his insights, we close with the following anecdote reported by Weinbaum (2013): when Daryl Morey applied the Pythagorean model to basketball, he was working part-time for STATS Inc., which was co-founded by Bill James. Upon learning of Morey's result, James remarked (Weinbaum 2013): "I would never have guessed that you could adapt the Pythagorean to basketball. Basketball has very small margins, relative to the score. A top baseball team scores 25 percent more runs than it allows, but a top basketball team outscores its opponents by only 6 to 7 percent." Viewing his statements in light of the results in this paper, it appears that James understood the sports fundamentals of his Pythagorean model more than even he realized.

References

Braunstein, A. 2010. "Consistency and Pythagoras." *Journal of Quantitative Analysis in Sports* 6(1):1–16.

Cochran, J. J. and R. Blackstock. 2009. "Pythagoras and the National Hockey League." *Journal of Quantitative Analysis in Sports* 5(2):1–13.

Dayaratna, K. and S. J. Miller. 2012. "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* 22:15–19.

Davenport, C. and K. Woolner. 1999. "Revisiting the Pythagorean Theorem." *Baseball Prospectus*. http://www.baseballprospectus.com/article.php?articleid=342. Accessed on June 11, 2017.

James, B. 1980. 1980 *Baseball Abstract*. Lawrence, KS: Self-published.

Jones, M. and L. Tappin. 2005. "The Pythagorean Theorem of Baseball and Alternative Models." *The UMAP Journal* 26(1):23–34.

Kubatko, J. 2013. "Pythagoras of the Hardwood." *Statitudes*. http://statitudes.com/?s=Pythagoras+of+the+Hardwood. Accessed on June 10, 2017.

Kubatko, J., D. Oliver, K. Pelton and D. T. Rosenbaum. 2007. "A Starting Point for Analyzing Basketball Statistics." *Journal of Quantitative Analysis in Sports* 3(3):1–24.

Miller, S. J. 2007. "A Derivation of the Pythagorean Won-Loss Formula in Baseball." *Chance* 20:40–48.

Miller, S. J., T. Corcoran, J. Gossels, V. Luo and J. Porfilio. 2014. "Pythagoras at the Bat." Pp. 89–113 in *Social Networks and the Economics of Sports*, edited by P. M. Pardalos and V. Zamaraev. Cham, Switzerland: Springer International.

Rosenfeld, J. W., J. I. Fisher, D. Adler and C. Morris. 2010. "Predicting Overtime with the Pythagorean Formula." *Journal of Quantitative Analysis in Sports* 6(2):1–19.

Schatz, A. 2003. "Pythagoras on the Gridiron." *Football Outsiders*. http://www.footballoutsiders.com/stat-analysis/2003/pythagoras-gridiron. Accessed on June 10, 2017.

Severini, T. A. 2015. *Analytic Methods in Sports*. Boca Raton, FL: CRC Press.

Statis Ticator. 2015. "Morey's Law: How Do Points Scored and Points Allowed Tie to Win Percentage?" *Statisticator*. http://statisticator.blogspot.com/2015/02/moreys-law-how-do-points-scored-and.html. Accessed on June 10, 2017.

Weinbaum, W. 2013. "Moreyball." *Northwestern*. http://www.northwestern.edu/magazine/winter2013/feature/moreyball.html. Accessed on June 13, 2017.

Winston, W. L. 2012. *Mathletics*. Princeton, NJ: Princeton University Press.