Journal of Student Research (2018) Volume 7, Issue 2: pp, 25-28
Research Article

DOES BILL JAMES'S PYTHAGOREAN FORMULA APPLY TO ENGLAND'S PREMIER LEAGUE?

Katherine A. Reinmuth[a], Paul M. Sommers[a]

Bill James's Pythagorean formula relates runs scored and runs allowed to team winning in Major League Baseball. Others have applied James's Pythagorean formula to basketball, ice hockey, and tennis, among other sports. Current literature, however, questions the applicability of the Pythagorean approach to soccer. This paper finds that such results may be because the dependent variable is based on the Fédération Internationale de Football Association's (FIFA) point system rather than wins and losses. The authors successfully apply James's Pythagorean formula to soccer and, in particular, England's Premier League (EPL). Between 2000/01 and 2016/17, there is a statistically significant relationship between a team's goal ratio (defined as goals scored divided by goals allowed) and a team's win-loss ratio. A one percent increase in the goal ratio is associated with a 1.70 percent increase in the win-loss ratio.

Keywords: Bill James's Pythagorean Formula, English Premier League, soccer

Bill James, baseball writer and statistician, in 1980 developed a formula for baseball that relates a team's win percentage to the number of runs they score and allow, as follows:

(1) Win Percentage =

$$\frac{(RunsScored)^2}{(RunsScored)^2 + (RunsAllowed)^2}$$

Since the win percentage is the ratio of games won to the total number of games played (games won plus games lost), equation (1) can be re-written as follows:

(2)

$$\frac{Wins}{Losses} = \frac{(RunsScored)^2}{(RunsAllowed)^2} = \left(\frac{RunsScored}{RunsAllowed}\right)^2$$

or in log-linear form:

(3)

$$\ln\left(\frac{Wins}{Losses}\right) = 2\ln\left(\frac{RunsScored}{RunsAllowed}\right)$$

where "ln" is the natural logarithm. Cha, Glatt, and Sommers [1] show that Bill James's Pythagorean formula (so named because of the presence of three squared terms in equation (1)) explains team winning in Major League Baseball between 1950 and 2007.

James's formula has been applied to other sports. Jackson *et al.* [2] show that when points scored and points allowed replace, respectively, runs scored and runs allowed, the James's formula explains team winning in the National Basketball Association with an exponent of "13.91" in lieu of "2" in equation (1). Cochran and Blackstock [3] find that goals scored and goals allowed (and an exponent closer to 1.93, rather than 2) explain team winning in the National Hockey League. Using performance data for the top 100 male singles players between 2004 and 2014, Kovalchik [4] derives a Pythagorean model for match wins in tennis based on the number of break points won.

Attempts to apply the Pythagorean formula to soccer using the sport's traditional point system (3 points for a win, 1 for a draw, and no points for a loss) have been less successful. (See, for examples, Bertin [5] and Hamilton [6].) One complicating factor is that there are ties (hereafter draws) in soccer, but not in baseball, basketball, hockey (since the NHL instituted a shootout), or tennis.

In this paper, we first show that when draws are included, there is no simple Pythagorean formula with a single "exponent" that explains variation in points scored by soccer teams in England's Premier League (hereafter the EPL). However, when draws are excluded, that is, one focuses on only wins and losses, a soccer Pythagorean formula emerges for teams in the EPL.[1,2]

The Data

Data were collected on wins (*W*), draws (*D*), and losses (*L*) for all twenty EPL teams over seventeen seasons (2000/01 through 2016/17) from http://www.oddsportal.com/soccer/england/premier-league-2016-2017/standings/. For each team, we recorded the number of goals scored (*GS*) and goals allowed (*GA*) per season. Teams receive 3 points for a win, 1 point for a draw; there are no points awarded for a loss.[3]

The number of points as a percentage of the team's maximum number of points (hereafter *POINTS*) is $3W + D$ divided by 38 times 3 or 114. For example, Chelsea's 2016-17 record was
30 wins, 3 draws, and 5 losses. Hence, Chelsea's *POINTS* in 2016-17 would be 93/114 or .816.

In leagues (such as soccer) with draws, the win percentage (*winPCT*) is frequently defined as:

(4)

$$winPCT = \frac{2W + D}{2 \times (W + D + L)}$$

In equation (4), a draw is worth one point or *half* the value of a win, as was the case in the pre-1994 FIFA point system. For example, Chelsea's 2016-17 *winPCT* would be .829

a. Department of Economics, Middlebury College, Middlebury, VT, 05753

$$[= \frac{(2\times 30)+3}{2\times(30+3+5)}].$$

The Methodology

Six regression models are estimated. In Model (1), we define $y_{i,t} = \ln(Points)_{i,t}$ and $x_{i,t} = \ln(GS/GA)_{i,t}$ for each team i in each year t. We can estimate the coefficients β_0 and β_1 by applying least squares to y and x in the following regression:

(5) $$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \varepsilon_{i,t}$$

where $\varepsilon_{i,t}$ is a disturbance term.

Model (1) assumes that increasing goals by a factor of k has the same impact on $y_{i,t}$ as decreasing the number of goals allowed by a factor of $(1/k)$.[4] But what if scoring goals was more (or less) important to accumulating points than allowing goals? Model (1) might be revised as follows:

(6) $$\ln(Points)_{i,t} = \beta_0 + \beta_1\ln(GS)_{i,t} + \beta_2\ln(GA)_{i,t} + \varepsilon_{i,t}$$

If scoring goals has a different effect on *POINTS* than allowing goals, then the revised model, hereafter Model (2), would be described by equation (6).

Model (3) is defined as follows:

(7) $$\ln(winPCT)_{i,t} = \beta_0 + \beta_1 \ln(GS/GA)_{i,t} + \varepsilon_{i,t}$$

In the spirit of Model (2), Model (4) replaces $\ln(GS/GA)_{i,t}$ with two regressors: $\ln(GS)_{i,t}$ and $\ln(GA)_{i,t}$.

Models (5) and (6) are similar to Models (3) and (4), respectively, but they *exclude* all matches ending in a draw. Moreover, the dependent variable is now the natural logarithm of the ratio of wins to losses, ln(*Wins*/*Losses*). According to Bill James, the intercept β_0 should be indistinguishable from zero. Moreover, β_1, the coefficient on ln(*GS*/*GA*) in Model (1) [(3) or (5)] should be the same as the coefficients (in absolute value) on both ln(*GS*) and ln(*GA*) in Model (2) [(4) or (6)].

The Results

Table 1. The Regression Results

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
Dependent Variable	*ln(Points)*	*ln(Points)*	*ln(winPCT)*	*ln(winPCT)*	*ln(Wins/Losses)*	*ln(Wins/Losses)*
Constant	-0.83	-1.236	-0.732	-0.933	0.02	0.334
	(-139.90)[a,*]	(-6.39)*	(-138.82)*	(-5.41)*	1.47	0.74
ln(GS/GA)	0.617	—	0.557	—	1.698	—
	(52.14)*		(52.97)*		(61.49)*	
ln(GS)	—	0.668	—	0.582	—	1.659
		(24.81)*		(24.23)*		(26.42)*
ln(GA)	—	-0.563	—	-0.53	—	-1.74
		(-20.01)*		(-21.11)*		(-26.37)*
R^2	0.889	0.891	0.893	0.893	0.918	0.918
[a]Numbers in parentheses are *t*-values.						
*$p < .001$						

Table 1 summarizes the regression results for all six models. The R^2 is better than 0.89 in all six regressions. And, in all six regressions, the explanatory variables, either ln(*GS*/*GA*) or the pair of regressors, ln(*GS*) and ln(*GA*), are significant at better than the .001 level. In Model (2), the coefficients on ln(*GS*) and ln(*GA*) are *not* equal (p = .036), although they are, as expected, opposite in sign. In Model (4), the corresponding coefficients *are* equal (p = .244), but the intercept is discernible from zero. (The intercept is also discernible from zero in Models (1) through (3).) When draws are excluded, as in Models (5) and (6), the intercept term is *not* discernible from zero, as Bill James would expect, *and* we cannot reject the null hypothesis that the coefficient on ln(*GS*) is equal (in absolute value) to the coefficient on ln(*GA*), p = .487, also in accordance with James's Pythagorean formula. Does the addition of fixed effects improve Model (5)? Fixed effects can be relevant for models that use panel data because they help avoid omitted variable bias due to unobservable heterogeneity. Fixed effects allow each cross-sectional unit (for examples, each team or each season) to have a different intercept. At first glance, correcting for this type of heterogeneity appears particularly important in soccer because of factors such as a tradition of relegation that ensures different league memberships from year to year and vastly different tactical approaches taken by each team. Adding fixed effects to Model (5)[5] gives:

$$(8)\quad \ln(Wins/Losses)_{i,t} = \beta_0 + \beta_1 \ln(GS/GA)_{i,t} + \alpha_1 Team1_t + \alpha_2 Team2_t + \ldots + \alpha_{41} Team41_t + \gamma_1 Season2000/01_i + \gamma_2 Season2001/02_i + \ldots + \gamma_{17} Season2016/17_i + \varepsilon_{i,t}$$

The fixed effects increase the model's unadjusted R^2 only marginally from .9182 to .9278 and the new estimated coefficient on $\ln(GS/GA)_{i,t}$ is 1.735, within the 95 percent confidence interval estimated for β_1 without the fixed effects (which extends from 1.644 to 1.752). In short, the 58 additional dummy variables do not improve the model's fit that would be worth the simultaneous loss in degrees of freedom.

Model (5) is therefore employed as the Pythagorean formula for soccer. The coefficient on ln(*GS*/*GA*) is 1.70 (rounded to two decimal places) and is statistically significant at better than the .001 level.[6] Thus the Pythagorean formula for soccer becomes:

$$(9)\quad \ln(Wins/Losses) = 1.70 \ln(GS/GA)$$

or

$$(10)\quad Wins/Losses = \left(GS/GA\right)^{1.70}$$

When the 17-year period is divided into two shorter periods, 2000/01 – 2007/08 and 2009/10 – 2016/17, the estimated coefficient on ln(*GS*/*GA*) is 1.73 (from 2000/01 through 2007/08) and 1.67 (from 2009/10 through 2016/17). In both cases, the estimated coefficient is not discernably different from 1.70 (with p-values of .446 and .419, respectively). Thus, the EPL-based coefficient of 1.70 appears to be a consistent estimate of the soccer variation of James's theoretical exponent.

Future research should address the limitation that draws or ties pose. Ties are non-existent in other sports like baseball, basketball, ice hockey (currently with overtime shootouts to break ties at the end of regulation), or tennis. Thus, draws in soccer (for which teams earn an additional point) present a difficulty in accounting for team success through win-loss ratios alone.

Concluding Remarks

Contrary to popular academic opinion, soccer appears to follow the Pythagorean relationship. More than 90 percent of the variation in a soccer team's win-loss ratio can be explained by the number of goals they score and allow. The results presented here show that the Pythagorean formula for soccer predicts at least as well as non-Pythagorean model specifications that seek to explain variation in the number of team points as a percentage of the maximum number of team points or a win percentage that assigns 2 points for a win, 1 point for a draw, and no points for a loss as a percentage of the maximum number of team points. The soccer coefficient of 1.70 implies that a one percent increase in the goals scored-goals allowed ratio is associated with a 1.70 percent increase in the win-loss ratio

References

1. D.U. Cha, D.P. Glatt, and P.M. Sommers, An empirical test of Bill James's Pythagorean formula, *J. of Recreational Mathematics*, 35:2, pp. 117-124, 2009.
2. A.S. Jackson, J.R. Piper, J.R. Jensen, and P.M. Sommers, Does Bill James's Pythagorean formula apply to basketball?, *Topics in Recreational Mathematics*, 1:1, pp. 7-18, 2015.
3. J.J. Cochran and R. Blackstock, Pythagoras and the National Hockey League, *Journal of Quantitative Analysis in Sports*, 5:2, pp. 1-13, 2009.
4. S.A. Kovalchik, Is there a Pythagorean theorem for winning in tennis?, *Journal of Quantitative Analysis in Sports*, 12:1, pp. 43-49, 2016.
5. M. Bertin, Improving soccer's version of the Bill James Pythagorean, *StatsBomb*, April 26, 2016, http://statsbomb.com/2016/04/improving-soccers-version-of-the-bill-james-pythagorean/.
6. H. Hamilton, Why the baseball Pythagorean doesn't work for soccer, *Soccermetrics Research*, April 21, 2016, http://www.soccermetrics.net/soccer-pythagorean-theory/why-the-baseball-pythagorean-doesnt-work-for-soccer.

Footnotes

1. The English Premier League is composed of 20 clubs (or teams) that play 38 matches each season. Over the sample period 2000/01 through 2016/17 seasons, teams played a total of 12,920 matches of which 3,304 (or 25.6 percent) ended in a draw.
2. Soccer goes to a shootout after overtime in playoff games (where a winner must be determined for the tournament/championship to continue). This is not, however, the case during regular season matches like the ones in our sample.
3. Beginning with the 1994 World Cup, FIFA made wins worth three rather than two points after several European leagues experienced reduced competitiveness when teams would intentionally tie

a. Department of Economics, Middlebury College, Middlebury, VT, 05753

each other towards the end of the season in an attempt to strategically alter playoff brackets and/or relegation results.

4. This is true because $\ln(k\text{GS/GA}) = \ln(\text{GS}/(\text{GA}/k))$.
5. Some readers may be puzzled that there are 41 dummies, when only 20 teams compete in the EPL in any one season. Over the 17-year period of study, the process of promotion and relegation in the EPL led to new teams (21 in particular) that were added to our sample.
6. This coefficient is also statistically significant from James's baseball's theoretical coefficient of "2" ($p < .001$).