AN EXTENSION OF THE PYTHAGOREAN EXPECTATION TO SOCCER COMPETITIONS

HOWARD H. HAMILTON

Abstract. In this publication, the Pythagorean expectation used for baseball and other team sports has been extended for use in soccer competitions. The principal extensions are the calculation of the probability of a drawn result and the use of expected points instead of wins. The extended Pythagorean is applied to a selection of professional soccer leagues in Europe, Asia, and North America, and is shown to predict very well the relative placement of teams in a league competition. The results across the leagues examined also give indication of the existence of a universal Pythagorean exponent.

1. Introduction

The Pythagorean formula was first developed by Bill James to predict the number of wins by a baseball team from the number of runs scored and allowed during the season. The formula is the win percentage of a baseball team during a season:

$$\frac{RS_{obs}^{\gamma}}{RS_{obs}^{\gamma} + RA_{obs}^{\gamma}} \tag{1.1}$$

where RS_{obs} and RA_{obs} represent the observed average number of runs scored or allowed during a season, and γ represents the Pythagorean exponent that is typically estimated by fitting score data over many seasons. The formula is typically calculated near the midpoint of the season for each team to estimate its end-of-season record and assess whether it is performing above or below expectations. Miller [2] derived from first principles the Pythagorean formula and matched James' result by making some reasonable assumptions of the run-scoring probability distributions.

The Pythagorean has been applied to baseball, basketball and other sports leagues with varying degrees of success. However, its applications to soccer have not been as successful and in general have resulted in an underprediction of points won over a season. One reason for this outcome is that the Pythagorean formula does not allow for the possibility of a tied result, which happens in a nontrivial percentage of soccer matches during a season. Another reason is that the Pythagorean exponent as derived for baseball may not be suitable for soccer because of the differences in the

scoring distributions between the two sports. While the scoring distribution in soccer has been shown to be similar to one observed in baseball [1], the shape and skew of the distribution are different. There has not yet appeared a suitable estimation of the Pythagorean exponent based on observed match results in professional soccer leagues.

In this publication the Pythagorean formula is extended for soccer by including a term that accounts for the probability of a tied result, using statistical assumptions that are identical to those used by Miller in his derivation of the baseball Pythagorean. The derivation of this term is contained in Section 2. The extended formula is then applied to a number of soccer leagues from around the world and the results are presented in Section 3.

2. Derivation of Extended Pythagorean

There are some initial assumptions about the statistics. First, the goals scored and allowed are statistically independent random variables. This assumption is a fair one to make in soccer because of the possibility of a draw. The second assumption is that the goals scored and allowed during a season follow a three-parameter Weibull distribution:

$$f(x;\alpha,\beta,\gamma) = \frac{\gamma}{\beta}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}},\ x \geq \beta \tag{2.1}$$

$$= 0,\ x < \beta \tag{2.2}$$

This is the same distribution that was used in [2] and it is a flexible statistical distribution that can take on a wide variety of shapes. The distribution parameter is defined as β, the scale parameter is α, and the shape parameter is the exponent γ, which is defined as the Pythagorean exponent. The third assumption is more of a mathematical construct in that the data are placed into N bins:

$$[-.5,.5] \cup [.5,1.5] \cup \cdots \cup [6.5,7.5] \cup [7.5,8.5] \cup \cdots \cup [N-.5, N+.5] \tag{2.3}$$

This construction ensures that the means of the bins correspond to its center, which is where all of the data in the bins will be located since there are only integer goals in soccer. Therefore the translation parameter $\beta = -0.5$. This mathematical construction also permits calculation of draw probabilities calculating the integral of the probability distribution function within an interval on the real axis.

To wrap up the preliminaries, we define expressions for the scale parameters of the two Weibull distributions, α_{GS} and α_{GA}. These parameters appear in the expressions for the means of the two distributions, which are equal to the average goals scored and allowed - G_S and G_A, respectively - and are defined as the following:

$$\begin{aligned} \alpha_{GS} &= \frac{G_S - \beta}{\Gamma(1+\gamma^{-1})} = \frac{\widehat{G}_S}{\Gamma(1+\gamma^{-1})} \\ \alpha_{GA} &= \frac{G_A - \beta}{\Gamma(1+\gamma^{-1})} = \frac{\widehat{G}_A}{\Gamma(1+\gamma^{-1})} \end{aligned}$$

The $\widehat{G}_S$ and $\widehat{G}_A$ terms are the average goals scored/allowed adjusted by the translation parameter, and the Γ term is the Gamma function:

$$\Gamma(s) = \int_0^\infty e^{-u} u^{s-1} du$$

With these assumptions in place, we begin the derivation for the draw probability $P(X = Y = c)$, where c is the number of goals scored.

$$P(X = Y = c) = \int_{c+\beta}^{c-\beta} \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}} \left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma-1} e^{-((x-\beta)/\alpha_{GS})^\gamma} \frac{\gamma}{\alpha_{GA}} \left(\frac{y-\beta}{\alpha_{GA}}\right)^{\gamma-1} e^{-((y-\beta)/\alpha_{GA})^\gamma} dy\, dx$$

After working through the integral and setting $\beta = -.5$, we arrive at the resulting expression for the probability of a drawn match at c goals:

$$P(X = Y = c) = \left[e^{-((c+1)/\alpha_{GS})^\gamma} - e^{-(c/\alpha_{GS})^\gamma}\right] \left[e^{-((c+1)/\alpha_{GA})^\gamma} - e^{-(c/\alpha_{GA})^\gamma}\right]$$

Then sum over N , which represents the highest number of goals scored by one side in a score draw, to obtain the total probability for a draw between teams X and Y:

$$P(X = Y) = \sum_{c=0}^{N} \left[e^{-((c+1)/\alpha_{GS})^\gamma} - e^{-(c/\alpha_{GS})^\gamma}\right] \left[e^{-((c+1)/\alpha_{GA})^\gamma} - e^{-(c/\alpha_{GA})^\gamma}\right] \tag{2.4}$$

Equations (1.1) and (2.4) are combined to form the extended Pythagorean formula, which will give the expected points won per game:

$$3 \cdot \frac{\alpha_{GS}^\gamma}{\alpha_{GS}^\gamma + \alpha_{GA}^\gamma} + \sum_{c=0}^{N} \left[\exp\left\{-\left(\frac{c+1}{\alpha_{GS}}\right)^\gamma\right\} - \exp\left\{-\left(\frac{c}{\alpha_{GS}}\right)^\gamma\right\}\right] \left[\exp\left\{-\left(\frac{c+1}{\alpha_{GA}}\right)^\gamma\right\} - \exp\left\{-\left(\frac{c}{\alpha_{GA}}\right)^\gamma\right\}\right]$$

A practitioner using this formula will have the goals scored and goals allowed data, which will be converted into average goals scored/allowed per game, and the league Pythagorean exponent. It is possible to rewrite the above equation in terms of statistics that the practitioner will have more readily. Let $\kappa = \Gamma\left(1 + {}^{1}/_{\gamma}\right)$ and then substitute for α_{GS} and α_{GA}in both terms to obtain the extended Pythagorean in terms of the goal averages and the Pythagorean exponent:

(2.5)

$$3 \cdot \frac{\hat{G}_S^{\gamma}}{\hat{G}_S^{\gamma} + \hat{G}_A^{\gamma}} + \sum_{c=0}^{N} \left[\exp\left\{ -\left(\frac{\kappa\,(c+1)}{\hat{G}_S} \right)^{\gamma} \right\} - \exp\left\{ -\left(\frac{\kappa c}{\hat{G}_S} \right)^{\gamma} \right\} \right] \left[\exp\left\{ -\left(\frac{\kappa\,(c+1)}{\hat{G}_A} \right)^{\gamma} \right\} - \exp\left\{ -\left(\frac{\kappa c}{\hat{G}_A} \right)^{\gamma} \right\} \right]$$

3. Application to League Competitions

The application of the extended Pythagorean to league data is a two-part procedure. The first part develops distributions of the goals scored and allowed and calculates the α_{GF}, α_{GA}, and γ parameters that fit the data to the Weibull distribution, and the second part uses the extended Pythagorean to calculate the estimated points won per game. The curve fit parameters minimize the sum of squares of the error between the goal scoring data and the probability distributions for goals scored and allowed

$$\min_{\{\alpha_{GF}, \alpha_{GA}, \gamma\}} || \, p_i - f(x_i; \alpha_{GF}, -0.5, \gamma) \, ||^2 + || \, q_i - f(x_i; \alpha_{GA}, -0.5, \gamma) \, ||^2$$

where $p_i$and q_i represent the histograms of the goals scored and conceded, respectively. The Pythagorean exponents for all the teams in the league are averaged to obtain the league Pythagorean exponent, and the κ term is calculated from the league exponent. Finally these terms and the goal scored and allowed averages are substituted in the extended Pythagorean formula in order to calculate the estimated point average for each league side.

An example of the curve fit of the goal distribution is shown in Figure 3.1, in which the best fit Weibull distribution is overlaid with the two goal scoring distributions. The fit does not capture heavily skewed distributions of the goal distributions very well because it considers the offensive and defensive distributions simultaneously, which is required for the Pythagorean estimation.

To examine the ability of the extended Pythagorean to predict the end-of-season results, the formula is applied to goal scoring data from various league competitions in North America, Europe, and Asia, as well as the World Cup qualifying tournaments in North and South America. Table 1 displays a summary of the Pythagorean exponents derived for these competitions. While one might have expected that the Pythagorean exponents would differ across various domestic leagues

Country	Season Year	League Pythagorean Exponent
Belgium	2008-09	1.71 ± 0.14
Germany	2008-09	1.71 ± 0.23
Italy	2008-09	1.70 ± 0.15
Japan	2008	1.73 ± 0.22
Portugal	2008-09	1.67 ± 0.20
USA	2008	1.73 ± 0.27
CONCACAF Hexagonal	2009	1.88 ± 0.55
CONMEBOL WCQ 2010	2007-09	1.42 ± 0.39

TABLE 1. Summary of league Pythagorean exponents for various league competitions.

according to different emphases on offensive or defensive play, the data for this small sample of league competitions indicate otherwise. The data show that the league Pythagorean exponents might reside within a narrow band; if verified with more data from other leagues and other seasons, such a result would greatly simplify the use of the extended Pythagorean for domestic league competitions. The exponents for the World Cup qualifying competitions in North and South America (which follow a league competition) fall outside the narrow band observed for the domestic league competitions. These results appear to be a consequence of the smaller number of matches played and the heavy number of either goalless draws, 1-0 results, or lopsided results (3-0 or more) in either competition.

In general the extended Pythagorean tends to overestimate the final points total by as much as 12-15 points. This result corresponds well to that found by Miller and James for which the baseball Pythagorean is accurate to within three or four games. Smaller differences in point totals have been observed for the teams at the top of the standings and larger variations observed for bottom sides. It is interesting that even though the estimated point totals do not correspond well with reality, the relative positions of the teams in the league are predicted very well. The predicted and actual positions of the teams in the league standings could give an even better indication of the team's performance with respect to their expectations.

4. Conclusions

The Pythagorean formula has been extended for soccer by the inclusion of a term that accounts for the probability of a tied result under similar statistical assumptions made in the derivation of the baseball Pythagorean. The application of the formula to domestic league competitions around the world has indicated that there might be a universal Pythagorean exponent that could be applied over leagues and seasons. The final paper will include a study of the variation of the league Pythagorean exponent over a number of league seasons and an attempt to derive

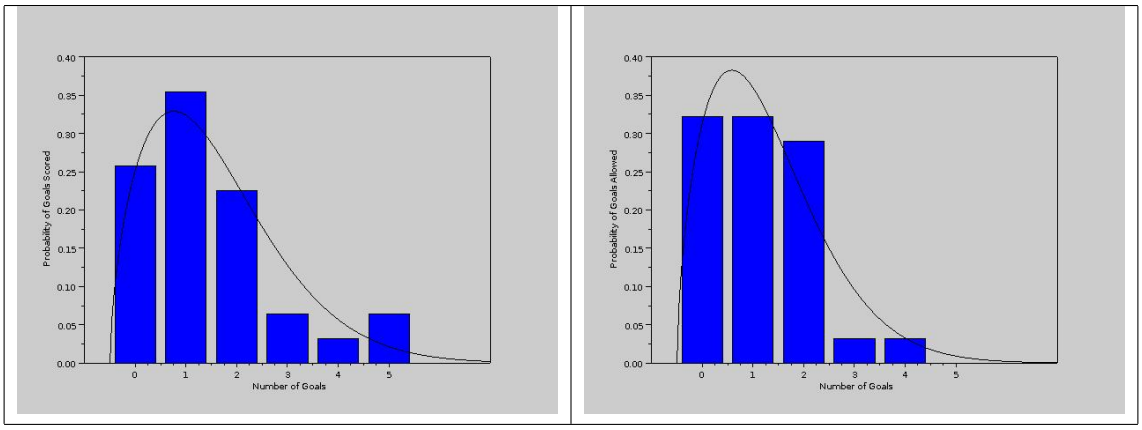

FIGURE 3.1. Best fit Weibull distribution overlaid with observed histograms of goals scored and allowed by Chicago Fire during the 2008 Major League Soccer regular season.

"second-order points", which would consider the points won by a side normalized by the strength of their offensive performance and their opponents' defensive performance.

5. Acknowledgements

I gratefully acknowledge the assistance of Dave Clark, Axel Becker, and Aditya Arpoorva who helped me compile the match result data for the leagues studied in this publication.

References

1. E. Bittner, A. Nussbaumer, W. Janke, and M. Weigel, *Self-affirmation model for football goal distributions*, Europhysics Letters **78** (2007), 58002.
2. Steven J. Miller, *A Derivation of the Pythagorean Won-Loss Formula in Baseball*, Chance Magazine **20** (2007), no. 1, 40–48.

Principal, Soccermetrics Research LLC, Tucson, AZ USA

E-mail address: howard.h.hamilton@gmail.com

URL: www.soccermetricsblog.com