Predict the Winning Ratio of Major League Baseball Teams: Generalized Pythagorean Formula Approach

Hsiao Chiu-Ming
Department of Finance, National Yunlin University of Science and Technology

Abstract

This study investigates the winning ratio of Major League Baseball (MLB) season games by using a generalized Pythagorean formula. Based on the method of regression analysis, this study verifies the ability of the generalized Pythagorean formula to predict MLB team winning rates. Empirical results show that the multivariate regression model has a highest adjusted R2 (0.9126) than other regression models. In addition, either MSE or MAE of in-sample and out-of-sample, the generalized Pythagorean formula derived from Cobb-Douglas production function has the lowest values than other versions of Pythagorean formula. Based on this generalized Pythagorean formula, team's managers can acquire potential rookies through drafts or player trades to recruit the players they need to achieve the team's goals.

Keywords: Pythagorean formula for baseball, Cobb-Douglas function, win-loss ratio

Corresponding author: Hsiao Chiu-Ming
E-mail: shiaucm@yuntech.edu.tw
DOI：10.53106/2226535X2024061302002

1. Introduction

To estimate the expected winning percentage of baseball team, Bill James, the pioneer sabermetrician, has developed a basic formula based on the runs scored and runs allowed (Yoon and Choi, 2022). This is a more statistically sound way to evaluate a team's performance than just looking at their actual wins and losses. Over the long run, he has found that a team's runs scored and runs allowed are strong indicators of their performance.

In addition, the Pythagorean expectation formula is also used in other sports and fields to assess team performance (Boudreaux, Ehrlich, Ghimire, and Sanders, 2021; Chen and Li, 2016; Cochran and Blackstock, 2009; Dayaratna and Miller, 2013; Sarlis and Tjortjis, 2020; Senevirathne and Manage, 2021). In essence, it's a simple yet powerful way to quantify the relationship between scoring runs and preventing runs in sports where the objective is to outscore the opponent.

Beside the Pythagorean formula for baseball, analysts and fans use other formulae to evaluate teams' performances and to identify those that might be over-performing or underperforming relative to their underlying statistics, such as, Hsiao, Zhang, Chen, and Chung, (2022) employs the probit model to estimate the win-loss ratio of NBA teams, meanwhile, Hsiao, Chou, Lin, Lin, Tsao, and Wang (2023) employs multivariate logit regression model to estimate the game's results of football teams in English Premier League. The formulae can help assess whether a team is likely to maintain its success (Sarlis and Tjortjis, 2020).

Although the Pythagorean expectation formula is useful and easily to understand, however, it also has some drawbacks. Since it is too simple to estimate the expected win-loss ratio such that it does not account for other factors, such as, luck, injuries, weather conditions, and other external variables which may affect a team's performance. Therefore, this study tries to develop a generalized Pythagorean formula to estimate the win-loss ratio.

The generalized Pythagorean formula in this study has the following improvements. First, the exponents of runs scored and runs allowed are not necessarily equal, which is different from previous versions of Pythagorean formula. Second, due to some non-quantitative factors affecting the team's winning rate, such as the aforementioned player suspension due to injury, players transfer team mid-season, mistakes, etc. Therefore, in the setting of the generalized Pythagorean formula, there is a constant K to capture the impact of this non-quantitative factor. Third, by setting a Cobb-Douglas production function to model the win-loss ratio, a generalized Pythagorean formula can be deduced. Therefore, the estimates of exponents for runs scored and runs allowed can be estimated through econometric models in Economics. Such that, either the original version or modified version of Pythagorean formula, the drawback of equal exponent assumption can be corrected. Finally, according to the derivation process of the generalized Pythagorean formula, the factors that affect the win-loss ratio can be extended. In other words, beside the original runs scored and runs allowed, some other factors can be added into the model if the factors have impact on the team's win-loss ratio (Mizels, Erickson, and Chalmers, 2022; Valero, 2016). That is to say, the generalized Pythagorean formula is extensible (Zech, 1981).

The structure of this study is as follows: Chapter 2 is the methodology that introduces the model setting and mathematical derivation of the generalized Pythagorean formula. Chapter 3 shows the empirical results of both regression and Pythagorean formulae. According to Yoon and Choi (2022), the assessments of several versions of Pythagorean formula are the mean squared errors (MSE) and mean absolute errors (MAE) and the assessment results are shown own in Chapter 4. The discussion and implications part is in the Chapter 5. In this chapter, this study will explore the theoretical implications of the generalized Pythagorean formula and try to put forward some management implications based on the empirical results. Final, Chapter 6 is the conclusion of this study.

2. Literature Review

In Jang, Lee, and Fort (2019), it explores the statistical properties of the winning ratios for nine major sports leagues around the world. Some are distributed as normal with higher kurtosis and some are skew which is abnormal. Furthermore, in accordance to Cefis (2023), data science is applied in several areas of daily life and there have been many applications to sports. Such that, the big data analysis can also be used in the professional sports management. Moreover, as shown in Baumer, Matthews, and Nguyen (2023), sports analytics has broadly defined as the pursuit of improvement in athletic performance through the analysis of data and has expanded its footprint both in the professional sports industry and in academia since 1980s.

From the movie “Moneyball”, the Oakland Athletics, under the leadership of manager Billy Beane, won a record of 102 wins and 60 losses, in the 2001-02 season (Elitzur, 2020). And the team owner spent only $34 million to pay the players’ salary, which was lower than either the New York Yankees or Texas Rangers that year, which salary are above of $100 millions of dollars. Such that, it can be regarded as a sign of high capital efficiency. In Lewis (2003), a baseball team’s analytical and statistical performance to the cost of capital are discussed first. Hereinafter, Hakes and Sauer (2006) extended Lewis’ approach by providing more empirical evidence that supports the claim that a baseball team’s performance goes beyond the box score output. Rosenfeld, Fisher, Adler, and Morris (2010) estimated the winning ratio of NBA, NFL and MLB teams by using the Pythagorean formula for the overtime games. They found that if a team has a 75% chance of winning a full-length game, its chances of winning an overtime game is 63% for the MLB. In addition, Joseph (2019) implements statistical examinations of winnings for the North American professional sports.

Beside the statistical approach, Yoon and Choi (2022) also implements mathematical approach to derive the Pythagorean formula for MLB. Traditionally, there are two approaches to estimate the Pythagorean exponent: (1) statistical approach

by assuming runs scored and runs allowed to be randomized (Chen and Li, 2016; Dayaratna and Miller, 2012; Ehrlich, Boudreaux, Boudreau, and Sanders, 2020; Heumann, 2016; McGoldrick and Voeks, 2005; Miller, 2007; Miller, Corcoran, Gossels, Luo, and Porfilio, 2014; Rothman, 2014); and (2) mathematical approach for calculating the Pythagorean exponent (Lee, 2014; Yoon and Choi, 2022). In this study, runs scored and runs allowed can be regarded as economic inputs, and their output is the team's odds ratio of win-loss. Such that, the exponents of these two inputs can be estimated using the Cobb-Douglas production function.

Furthermore, in order to verify the validation of the generalized Pythagorean formula, this study also introduces the model evaluation method of Yoon and Choi (2022) to compare various versions of the Pythagorean formula with the estimation results of regression models, including ordinary least squares (OLS), two-stage regression, and Poisson regression. OLS is an estimation of multivariate regression models that can be used to estimate which factors are significant to a team's winning ratio and their loadings. However, when there is a high degree of correlation among many influencing factors, there may be multicollinearity among these variables. This can even cause the original estimated values of individual parameters to be unstable and the standard errors to be too large (Greene, 2018; Wen and Chiou, 2009). Therefore, this study also introduces the two-stage regression method to investigate the nexus between winning ratio and the explanatory variables. Firstly, regress the home winning ratio on some highly correlated variables. Secondly, after calculating the residuals of this part, and then add it with other variables to the explanatory variables to regress to the winning ratio.

Since the number of wins in the entire season is a positive integer, such that, the winning ratio of the season can be estimated by estimating the number of wins of the team. For this reason, this study further introduces the Poisson regression model which is used to explain the dependent variable when it is a counting number to investigate the nexus between wins and other explanatory variables.

2.1 Basic Pythagorean formula

First, consider the original Pythagorean formula for baseball proposed first by James (1980). The basic Pythagorean formula for the win-loss ratio (WCPT) is given by

$$WCPT_{James2} \equiv \frac{\left(S/_{GP}\right)^2}{\left(S/_{GP}\right)^2+\left(A/_{GP}\right)^2}, \tag{1}$$

where, S and A represent the total runs scored and total runs allowed of the team, respectively; GP is the games played in the season. A corrected version in Valero (2016) is given as follows:

$$WCPT_{James183} \equiv \frac{\left(S/_{GP}\right)^{1.83}}{\left(S/_{GP}\right)^{1.83}+\left(A/_{GP}\right)^{1.83}}. \tag{2}$$

In addition, Ehrlich et al. (2020) used simulated MLB game results and estimated the exponent to be 1.722. However, Boudreaux et al. (2021) used MLB data from the 2003 season to the 2015 season to estimate the exponent at 1.859. This study will compare the results predicted by these methods with that of the generalized Pythagorean formula empirically by using the MLB data from season 2013 to season 2022.

2.2 Miller's Pythagorean formula

Next, Miller (2007) developed a modified version of the James' Pythagorean formula using a Weibull distribution specification of the win-loss ratio. The Miller's Pythagorean formula for the win-loss ratio is given as follows:

$$WCPT_{Miller} \equiv \frac{\left(S/_{GP}-\theta\right)^{\gamma}}{\left(S/_{GP}-\theta\right)^{\gamma}+\left(A/_{GP}-\theta\right)^{\gamma}}, \tag{3}$$

for some parameters θ and γ.

The Miller's version of Pythagorean formula is based on the assumption that runs scored and runs allowed are random variables distributed as Weibull distribution (Miller et al., 2014). However, it is not easy to estimate the parameters of such an

exquisite probability distribution. Therefore, it is not easy to estimate the team's win-loss ratio.

In order to simplify the complex calculations of the Miller's version, some modified versions set the parameter θ in the Miller's version of the Pythagorean formula to 0. For example, David Smyth, who invented the base runs (BsR), uses $\theta = 0$ and $\gamma = \left(\frac{S+A}{GP}\right)^{0.287}$; Clay Davenport, a co-founder of Baseball Prospectus (BP), employed $\theta = 0$ and $\gamma = 0.45 + 1.5 \times \ln\left(\frac{S+A}{GP}\right)$.

Furthermore, Lee (2015) used Korean Professional Baseball data from the 2004 season to the 2013 season to estimate an exponent higher than MLB. Moreover, in the Lee's empirical result, the exponents of runs scored and runs allowed are significantly unequal. This result implies that the exponents of runs scored and runs allowed may not be the same value.

2.3 Generalized Pythagorean formula

Aforementioned, the exponents of runs scored and runs allowed in the modified versions of the Pythagorean formula are set to be equal. It can be thought that the runs scored and runs allowed are equally weighted to the winning ratio. As Sun Bin, the greatest Chinese military strategist, said that team is skillful in attack whose opponent does not know what to defend; and it is skillful in defense whose opponent does not know what to attack. In the sports, every score can be regarded as a display of the overall offensive results of an individual or a team (Jones and Tappin, 2005); and every point lost can also be thought as the price paid for a defensive failure (Jones and Tappin, 2005; Heumann, 2016).

Furthermore, the integration of offense and defense is an important factor for a team to win. Attack and defense cannot be partial to either side. As indicated in Porter and Scully (1982), the baseball team's winning ratio is related to team's hitting performance and team's pitching performance. Based on the model of Porter

and Scully (1982), the performance of team's managers can be evaluated, although Ruggiero, Hadley, Ruggiero and Knowles (1998) does not think so. In this study, some correction should be made including the performance of team, the factors which affect the performance and their corresponding exponents.

First, according to Scully (1974), the team performance is an important measure for the managers who take some strategies to manage the team. In the past, most of the research on winning ratio focused on the econometric models, with less discussion of its theoretical basis (Dayaratna and Miller, 2016). However, recently, win-loss ratio has been the research topic not only because it can represent the performance of the team, but also has some economic implications of the model by setting the Cobb-Douglas production function of the win-loss ratio.

Second, Cochran and Blackstock (2009) investigates the winning ratio of National Hockey League (NHL) by using modified Pythagorean formula which allows for different values of the exponents in the three positions of the formula. Such that, as indicated in Zech (1981), consider the following Cobb-Douglass production function (Cobb and Douglas, 1928) for the win-loss ratio:

$$\frac{WCPT_G}{1-WCPT_G} = f\left(\frac{S}{GP}, \frac{A}{GP}\right) \equiv K \cdot \left({}^{S}/_{GP}\right)^{\alpha} \cdot \left({}^{A}/_{GP}\right)^{\beta}, \tag{4}$$

for some constants α, β, and K is positive. Then eq. (4) can be transformed into

$$K \cdot \left({}^{S}/_{GP}\right)^{\alpha} \cdot \left({}^{A}/_{GP}\right)^{\beta} - K \cdot \left({}^{S}/_{GP}\right)^{\alpha} \cdot \left({}^{A}/_{GP}\right)^{\beta} \cdot WCPT_G = WCPT_G. \tag{5}$$

Such that,

$$WCPT_G \equiv \frac{K \cdot \left({}^{S}/_{GP}\right)^{\alpha} \cdot \left({}^{A}/_{GP}\right)^{\beta}}{1 + K \cdot \left({}^{S}/_{GP}\right)^{\alpha} \cdot \left({}^{A}/_{GP}\right)^{\beta}}. \tag{6}$$

As a result, it can be rewritten into

$$WCPT_G = \frac{K \cdot \left({}^{S}/_{GP}\right)^{\alpha}}{K \cdot \left({}^{S}/_{GP}\right)^{\alpha} + \left({}^{A}/_{GP}\right)^{-\beta}}. \tag{7}$$

Hence, the generalized Pythagorean formula for the expected winning ratio can be found in the above equation.

Moreover, according to the Cobb-Douglas form of win-loss, then the logarithmic win-loss ratio is given as follows:

$$\ln\left(\frac{WCPT_G}{1-WCPT_G}\right) = \ln K + \alpha \cdot \ln\left({}^{S}/_{GP}\right) + \beta \cdot \ln\left({}^{A}/_{GP}\right), \tag{8}$$

such that, the constants in the model can be estimated by using a double-log linear regression and historical data.

Therefore, as Yoon and Choi (2022), based on the mean square errors (MSE) and mean absolute errors (MAE), this study compares the results of various predictions of MLB team winning rates, including three regression models and four main versions of the Pythagorean formula.

3. Empirical Results

3.1 Data collection and descriptive statistics

Since the MLB regular season games is the main object to study, such that the empirical data is downloaded from the MLB official website (https://www.mlb.com/stats/team). The full data period is from 2013 to 2022 for the 30 teams in MLB. Based on the data, this study establishes six models to fit the data and then uses to make in-sample prediction of winning ratio and out-of-sample forecasting of 2023 game results.

In this study, *WinRatio* and *HomeWR* represent the total winning ratio and winning ratio at home court, respectively. *WinRatio* is the dependent variable and defined as the ratio of wins and games played in the season. *HomeWR* is defined as the ratio of wins and games played at home curt. In addition, RS and RA are the runs scored and allowed, respectively. *ERA* is the average number of earned runs a pitcher allows per 9 innings. *AVG* is the rate of hits per bat. *GO2AO* is the ratio of groundout to air out. *SO9* is the strikeout rate per 9 innings and *K2BB* is the strikeout to walk rate. The following table shows the descriptive statistics of the variables.

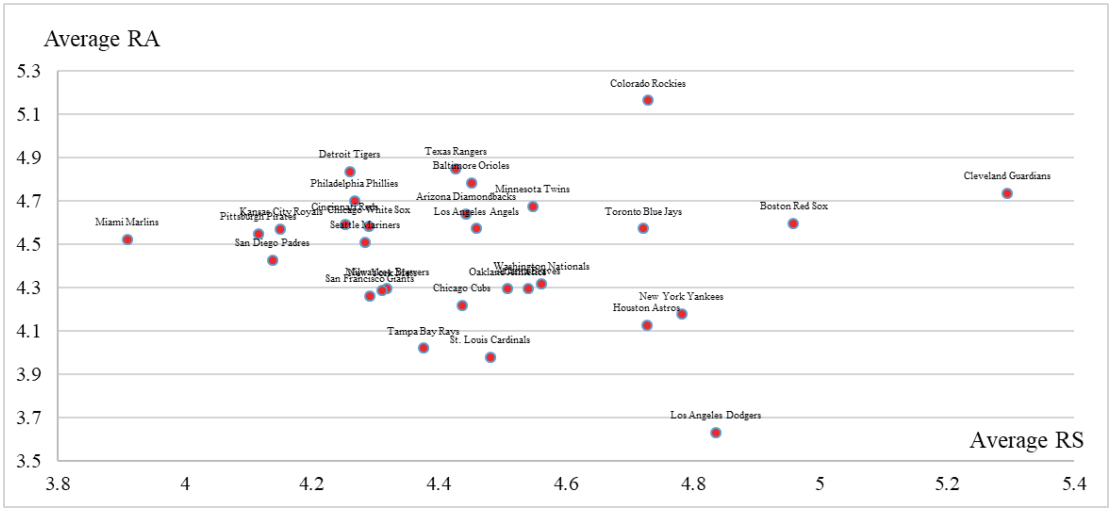
Table 1
The descriptive statistics of variables (Data period: 2013 to 2022)

Variable	Observations	Mean	Std. Dev.	Median	Min	Max
WinRatio	300	0.5003	0.0798	0.50	0.2901	0.7167
HomeW	300	0.5393	0.1022	0.54	0.2733	1.3333
RS	300	4.4596	0.6554	4.41	3.1667	11.9500
RA	300	4.4609	0.7212	4.38	3.1667	12.1667
ERA	300	4.1416	0.5745	4.07	2.8000	5.8400
AVG	300	0.2499	0.0147	0.25	0.2070	0.2830
GO2AO	300	1.0350	0.1267	1.02	0.7800	1.5100
SO9	300	8.3383	0.8550	8.29	6.1100	10.9800
K2BB	300	2.6654	0.4432	2.60	1.7600	3.9800

Data source: MLB official website (https://www.mlb.com/stats/team/).

Table 1 shows that the average value (median) of *WinRatio* is 0.5003 (0.50), half-and half winning ratio MLB teams. In addition, the average value (median) of *HomeWR* is 0.5393 (0.54), slightly higher than 0.5, which means that the home-court advantage is valid but insignificant statistically. Moreover, the average values of runs scored *(RS)* and runs allowed *(RA)* are almost equal. In other words, the average offensive and defensive performance are almost the same.

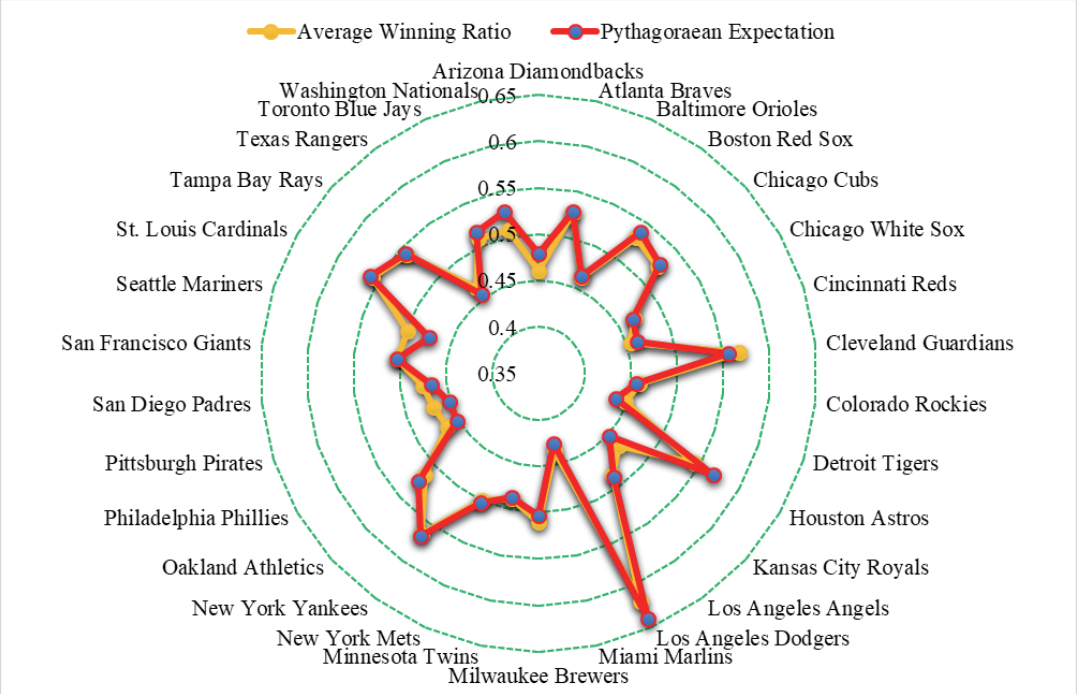
Figurer 1
Distribution of average runs scored and average runs allowed (2013-2022)



On average, Figure 1 shows that Los Angeles Dodgers has the lowest runs allowed while Colorado Rockies has the highest runs allowed. In addition, Cleveland Guardians has the highest runs scored, however, Miami Marlins has the lowest runs scored.

Figure 2

Average winning ratio and average Pythagorean expectation (2013-2022)



In Figure 2, the Pythagorean expectations (in red) of each team follow the same pattern as their average winning ratios (in yellow). Moreover, the two values for each team are also almost the same. It indicates that the Pythagorean expectation may be an indicator for the team's winning ratio. This result is coincide with James (1980).

Table 2
The pairwise correlation coefficients of variables

	WinRatio	*Home*	*RS*	*RA*	*ERA*	*AVG*	*GO2AO*	*SO9*
HomeW	0.8220***							
RS	0.4984***	0.6587***						
RA	-0.5573***	-0.1953***	0.3697***					
ERA	-0.0243	0.0205	-0.0351	-0.0066				
AVG	-0.0037	0.0696	0.0999*	0.0716	0.7450***			
GO2AO	-0.1113*	-0.0660	-0.0103	0.0840	-0.1651***	0.0944		
SO9	0.0253	-0.0423	-0.1297**	-0.114*	-0.2360***	-0.6279***	-0.1189**	
K2BB	0.0554	0.0130	0.0121	-0.0244	-0.6770***	-0.6443***	-0.0056	0.6183***

Note: *, ** and *** are represented the significant level of 10%, 5% and 1%, respectively. Data period is from 2003 to 2022 downloaded from MLB official website.

Table 2 shows that the *WinRatio* is significantly positively correlated to runs scored *(RS)* and negatively correlated to runs allowed *(RA)*. This result supports to the definition Pythagorean formula proposed by James (1980). Such a phenomenon also appears in *HomeWR*.

3.2 The multivariate regression results

For the regression models, this study will employ the backward reduction method to eliminating the variables from the full regression model when the parameter estimate is statistically insignificant (*p*-value is larger than 10%).

3.2.1 Linear regression

According to Rothman (2014), a linear regression model is considered as follows:

$$WinRatio = \alpha + \gamma_1 \cdot HomeWR + \beta_1 \cdot RS + \beta_2 \cdot RA + \gamma_2 \cdot SO9 + \varepsilon, \quad (9)$$

where, ε is the random error. The estimating results are shown in the following table.

Table 3
The estimation result of multivariate regression model

Variable	*WinRatio*
HomeW	0.2153***
	(9.38)
RS	0.0713***
	(18.84)
RA	-0.0792***
	(-30.07)
SO9	0.0031*
Constant	0.3938***
	(20.11)
Observations	300
Adj.	0.9126
Mean VIF	2.28

Note: The value in parentheses is the t-value of the parameter estimate. *, ** and *** are represented the significant level of 10%, 5% and 1%, respectively.

In Table 3, *RS* and *RA* have significantly effect on the winning ratio, however, they are in different direction. Moreover, the null hypothesis: $|\beta_1| = |\beta_2|$ is not rejected with *p*-value 0.23, that is, the absolute value of loadings of *RS* and *RA* are statistically insignificantly different. This result is coincide with the result of Rothman (2014).

3.2.2 Two-stage regression

As shown in Table 2, there are significantly correlated relationship between *HomeWR* and RS, RA. In accordance with Greene (2018), when there is a high degree of correlation between variables, the accuracy of the estimation of the regression coefficients may be reduced, making the variation larger and the estimation less stable. To reduce the effect of high correlation between variables in the regression model, Yang and Swartz (2004) proposed a two-stage regression method to predict the winning ratio. Such that, the first stage is to regress *HomeWR* by using the following equation:

$$HomeWR = c_0 + c_1 \cdot RS + c_2 \cdot RA + \epsilon_H, \tag{10}$$

where, ϵ_H is random error.

Using the result in the first stage, the estimated *HomeWR* is given in the following

$$\widehat{HomeWR} = 0.2719851 + 0.1319986 \times RS - 0.0720253 \times RA, \quad (11)$$

and the residuals can be found as follows:

$$HomeWRErr = HomeWR - \widehat{HomeWR}. \quad (12)$$

Next, the second stage regression model is given as follows:

$$WinRatio = \alpha + \beta_1 \cdot RS + \beta_2 \cdot RA + \delta \cdot HomeWRErr + \varepsilon. \quad (13)$$

The following table shows the estimating results of the second stage regression.

Table 4
The estimation result of two-stage regression model

Variable	***WinRatio***
RS	0.0993***
	(43.85)
RA	-0.0950***
	(-46.16)
Home	0.2162***
	(9.38)
Constant	0.4812***
	(43.99)
Observations	300
Adj.	0.9109
Mean VIF	1.11

Note: The value in parentheses is the t-value of the parameter estimate. *, ** and *** are represented the significant level of 10%, 5% and 1%, respectively.

As shown in Table 4, the estimates of RS and RA are 0.0993 and -0.0950, respectively. Such that, the null hypothesis: β_1=-β_2 is not rejected with the *p*-value 0.46. This result supports the assumption of all versions of Pythagorean formula, except the generalized one.

3.2.3 Poisson regression result

Since the wins of a regular season are always counting numbers, such that, a

Poisson regression model can also be employed to estimate the team's wins (Hsiao, 2022). As described in Greene (2018), the Poisson regression model is given as follows:

$$\ln E[Wins|\boldsymbol{X}] = \beta_0 + \boldsymbol{X} \cdot \boldsymbol{\beta}, \tag{14}$$

with the likelihood function of Poisson distribution:

$$\Pr[Wins = n|\boldsymbol{X}] = \frac{e^{-\lambda(\boldsymbol{X})} \cdot [\lambda(\boldsymbol{X})]^n}{n!}, n \in \mathbb{N} \cup \{0\}. \tag{15}$$

Hence, the estimating results of Poisson regression are shown in the following table.

Table 5
The estimation result of Poisson regression model

Variable	***WinRatio***
WinsHome	0.0272***
	(28.53)
RS	-0.0386***
	(-2.96)
RA	-0.0146
	(-1.11)
Constant	3.4235***
	(44.80)
Observations	300
Pseudo	0.4238

Note: The value in parentheses is the z-value of the parameter estimate. *, ** and *** are represented the significant level of 10%, 5% and 1%, respectively.

Hence, the expected wins of the team can be estimated by the following equation:

$$\widehat{Wins} = e^{3.4235 + 0.0272 \times HomeRatio - 0.0386 \times RS - 0.0146 \times RA}, \tag{16}$$

and the expected winning ratio is given as follows:

$$\widehat{WinRatio} = \widehat{Wins} \div GP, \tag{17}$$

where, *GP* represents the total games played in the regular season.

3.3 Pythagorean formulae

3.3.1 Miller's Pythagorean formula

As in the definition of James' Pythagorean formula (Winston, Nestler, and Pelechrinis, 2022), the expected win-loss ratio is given as follows:

$$\frac{WCPT}{1-WCPT} = \left(S/A\right)^2. \tag{18}$$

Such that, the logarithmic value of expected win-loss ratio can be found by

$$\ln\left(\frac{WCPT}{1-WCPT}\right) = 2 \cdot \ln\left(S/A\right), \tag{19}$$

Furthermore, the following regression model can be used to estimate the exponent of the runs scored:

$$\ln\left(\frac{WinRatio}{1-WinRatio}\right) = \beta \cdot \ln\left(S/A\right) + \epsilon. \tag{20}$$

The estimate of the above regression equation is $\hat{\beta} \fallingdotseq 1.766423$ with standardized error 0.0367 for the MLB regular games from 2013 to 2022. It is different to the value of 1.83 which was estimated in the previous researches (Kaplan and Rich, 2017). Moreover, it is also different to the result in Valero (2016). Hence, the Miller's version of Pythagorean formula is given as follows:

$$\widehat{WCPT}_{Miller} \equiv \frac{\left(S/GP\right)^{1.766423}}{\left(S/GP\right)^{1.766423} + \left(A/GP\right)^{1.766423}}. \tag{21}$$

3.3.2 Generalized Pythagorean formula

As mentioned above, the following double-log regression model can be used to find the estimates of the parameters:

$$\ln(Odds) = \ln(K) + \beta_1 \cdot \ln(RS) + \beta_2 \cdot \ln(RA) + \epsilon, \tag{22}$$

where, *Odds* stands for the odds ratio which is defined as the ratio of wins and losses. And ϵ is the random error. The following table shows the regression results.

Table 6
The estimating results for the generalized Pythagorean formula

Variable	***LnOdds***
ln(RS)	1.7773***
	(34.13)
ln(RA)	-1.7585***
	(-38.22)
ln(K)	-0.03
	(-0.32)
Obs.	300
Pseudo	0.8850

Note: The value in parentheses is the t-value of the parameter estimate. *, ** and *** are represented the significant level of 10%, 5% and 1%, respectively.

According to Table 6, it can be found that the estimates of ln(*RS*) and ln(*RA*) are different. As the test statistic for testing the null hypothesis: $\beta_1=-\beta_2$ is 2.15 with *p*-value 0.04, then the hypothesis is rejected. In other words, the exponents of *RS* and *RA* in the Pythagorean formula for baseball are significantly different. This result is contradict to all versions of Pythagorean formula for baseball.

4. Models Assessment

In this section, there are two assessment methods can be used, i.e., mean squared errors (MSE) and mean absolute errors (MAE). The MSE and MAE are defined as

$$\mathrm{MSE} \equiv \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\mathrm{WinRatio}}_i - \mathrm{WinRatio}_i\right)^2, \tag{23}$$

and

$$\mathrm{MAE} \equiv \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{\mathrm{WinRatio}}_i - \mathrm{WinRatio}_i\right|, \tag{24}$$

respectively. In which, the $\widehat{\mathrm{WinRatio}}$ is the estimated WinRatio.

Moreover, according to Chen and Li (2016), the accuracy rate of a model is defined as

$$\mathrm{Accuracy}(\lambda) \equiv 100 \times \frac{1}{N}\sum_{i=1}^{N} I_{\left\{\frac{\left|\widehat{\mathrm{WinRatio}}_i - \mathrm{WinRatio}_i\right|}{\mathrm{WinRatio}_i} \le \lambda\right\}}, \tag{25}$$

where, I_A is an indicator function that gives value of 1 if $x \in A$ and 0 otherwise. In addition, λ is a threshold value, in this study, the values of λ are given three threshold values of 3%, 8%, and 10%, respectively.

4.1 Overall assessment of models

Therefore, the MSE, MAE and Accuracy of in-sample prediction and out-of-sample forecasting are shown in the following table.

Table 7
MSE and MAE for the in-sample prediction and out-of-sample forecasting

	OLS	2-Stage	Poisson	James	Miller	Smyth	Generalized
Panel A: In-sample prediction							
MSE	0.0025	0.0031	0.0195	0.0034	0.0032	0.0032	0.0032
MAE	0.0372	0.0415	0.0962	0.0449	0.0433	0.0433	0.0433
Accuracy (3%)	51.67	47.67	19.67	44.67	44.67	45.00	44.33
Accuracy (8%)	89.33	87.67	56.00	85.33	86.67	86.67	86.67
Accuracy (10%)	95.67	92.67	66.67	90.33	92.67	91.67	92.67
Panel B: Out-of-sample forecasting							
MSE	0.0027	0.0027	0.0085	0.0039	0.0033	0.0034	0.0033
MAE	0.0416	0.0424	0.0703	0.0512	0.0449	0.0469	0.0450
Accuracy (3%)	46.67	46.67	36.67	36.67	33.33	36.67	33.33
Accuracy (8%)	86.67	86.67	56.67	80.00	86.67	86.67	86.67
Accuracy (10%)	96.67	96.67	73.33	90.00	90.00	90.00	90.00

As shown in Panel A of Table 7, the MSE and MAE using OLS is the lowest than that of other models. Indeed, since the accuracy rate of the multivariate regression model is as high as 95.67%, it is the highest among all models, however, the accuracy rate of the generalized Pythagorean formula is 92.67% which is the highest among all version of the Pythagorean formula. Therefore, as for the assessment of in-sample prediction, the OLS and generalized Pythagorean formula are the best model in their individual class, respectively.

On the other hand, in Panel B, the MSE, MAE and accuracy rate of the

Pythagorean formula are as good as that of the regression models, except the James' version of Pythagorean formula. Furthermore, the MSE and MAE of the generalized Pythagorean formula proposed in this study are lower than that of other versions of Pythagorean formula. Hence, the generalized Pythagorean formula is a better model in forecasting the winning ratio of the MLB team.

4.2 Prediction accuracy by leagues

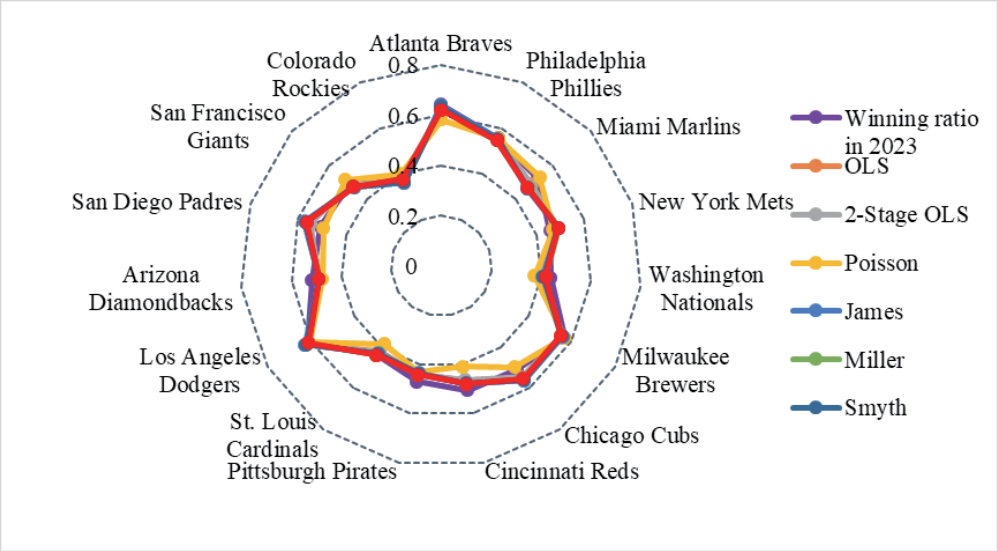
Totally, MLB has a total of 30 teams and is evenly divided into two leagues: the American League and the National League. Each league is divided into three divisions: East, West and Central. Therefore, each division has 5 teams.

4.2.1 Prediction accuracy for National League

There are 15 teams in the National League. The team with the highest rank in each division of the Eastern, Western and Central divisions will qualify for the playoffs. In addition, the other 12 teams will be ranked according to their records, and the top three will qualify for the wild card. Based on the aforementioned models, this study estimates the 2022-2023 winning ratio for each team. Estimating results are shown in the following diagram. The true winning ratio of 16 American League teams is in purple; forecasting results from OLS, 2-Stage OLS and Poison regression are in orange, grey and yellow, respective. Forecasting results by James' version, Miller's version, Smyth's version and generalized version of Pythagorean formula are in light blue, green, deep blue and red, respectively.

Figure 3

The estimates of winning ratios for teams in National League



The results in Figure 3 show that, beside the Poisson regression model, both the OLS and 2-Stage regression models can correctly estimate the winning ratio of the eight National League teams entering the 2022-23 playoffs. However, each version of the Pythagorean formula underestimates the winning ratios of these eight teams. Moreover, when it comes to estimating the winning ratios of the top two teams in each division of the National League, the generalized Pythagorean formula has good estimation results. Furthermore, the assessment results are shown in the table below.

Table 8

Forecasting winning ratio of 2023 regular games: National League

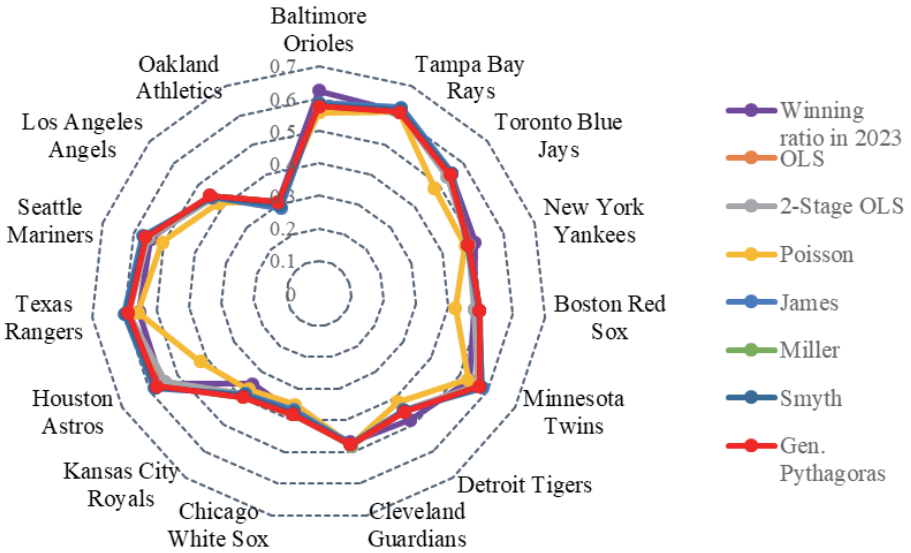
Assessment	OLS	2-Stage OLS	Poisson	James	Miller	Smyth	Generalized Pythagoras
MSE	0.0030	0.0029	0.0074	0.0043	0.0035	0.0038	0.0035
MAE	0.0446	0.0453	0.0660	0.0525	0.0497	0.0495	0.0500
Accuracy (3%)	40.00	40.00	40.00	40.00	26.67	40.00	26.67
Accuracy (8%)	80.00	86.67	53.33	80.00	80.00	80.00	80.00
Accuracy (10%)	100.00	100.00	80.00	86.67	86.67	86.67	86.67

In Table 8, except for the Poisson regression model, the estimation results of the two regression models are better than those of each version of the Pythagorean formulae. Moreover, in terms of accuracy, even though the threshold is released, the Pythagorean formulae is lower than those of the regression models. However, the generalized Pythagorean formula for baseball proposed in this study has more theoretical significance since it is derived from a Cobb-Douglas production function.

4.2.2 Prediction accuracy for American League

Next, this study also uses the aforementioned models to forecast the winning ratio of the 15 American League teams in the 2022-23 season. The team with the highest rank in each division of the Eastern, Western and Central divisions will qualify for the playoffs. In addition, the other 12 teams will be ranked according to their records, and the top three will qualify for the wild card. Such that, the estimating results are shown in the figure below.

Figure 4



The estimates of winning ratios for teams in American League.

As shown in the Figure 4, the eight teams that entered the 2022-2023 playoffs, except for the underrated Toronto Blue Jays and the overrated Seattle Mariners, seven were correctly predicted to qualify. In addition, with the exception of the Tampa Bay Rays, the no other team's record predictions in the American League is higher than the Texas Rangers. Furthermore, the assessment results are shown in the table below.

Table 9
Forecasting winning ratio of 2023 regular games: American League

Assessment	OLS	2-Stage OLS	Poisson	James	Miller	Smyth	Generalized Pythagoras
MSE	0.0024	0.0025	0.0096	0.0034	0.0030	0.0030	0.0030
MAE	0.0387	0.0394	0.0745	0.0499	0.0401	0.0442	0.0401
Accuracy (3%)	53.33	53.33	33.33	33.33	40.00	33.33	40.00
Accuracy (8%)	93.33	86.67	60.00	80.00	93.33	93.33	93.33
Accuracy (10%)	93.33	93.33	66.67	93.33	93.33	93.33	93.33

In Table 9, except for the Poisson regression model, the estimation results of the two regression models are better than those of each version of the Pythagorean formulae. However, in terms of accuracy, when the threshold is released, the Pythagorean formulae is consistent with both the regression models. Therefore, the generalized Pythagorean formula proposed in this study has more theoretical significance since the performance of the generalized Pythagorean formula for baseball can also be improved by introducing other possible factors, such as superstar players, last season's record, main players changing teams or entering the disabled list, etc.

5. Discussion and Implications

5.1 Discussion

As the generalized Pythagorean formula proposed in this study has pointed

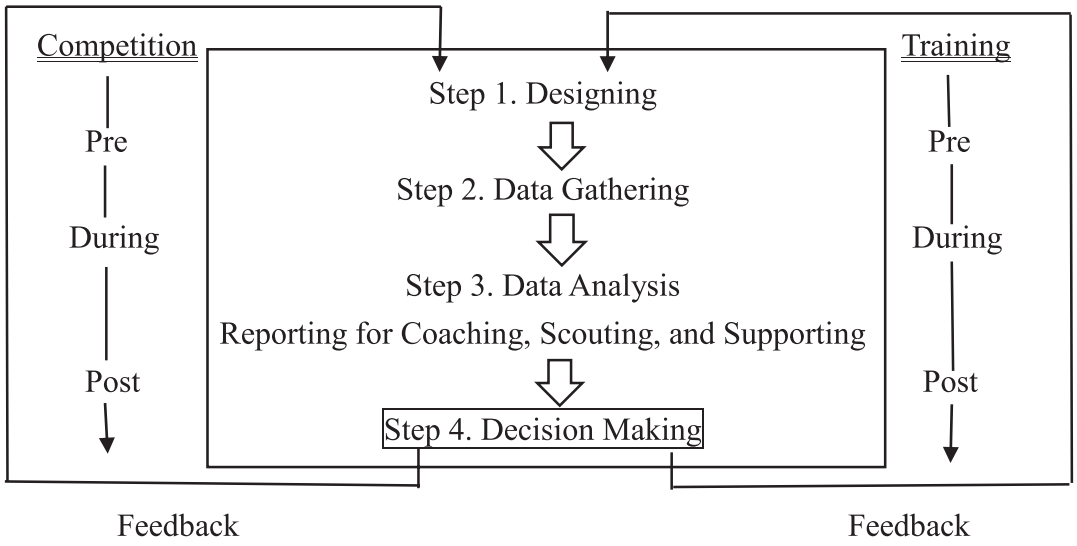
out that the effects of offensive performance and defensive performance are not equivalent. This result is obviously different from various versions of the Pythagorean formula in the past. However, it has more economic significance and extensibility. For economic significance, through the Cobb-Douglas production function in macroeconomics, incur two main inputs: offense and defense, and then estimate the team's win-loss ratio (odds ratio), which is used as the team's entire season performance.

Furthermore, the extensibility is that the Cobb-Douglas production function is not limited to only two inputs, so other factors that may affect the output can be included, too. The factors are then brought into the function in order to be more consistent with the actual situation of the output model. For instance, the total salary of the players which represents the amount of investment the team has made in the players. It is hoped that the addition of players will bring positive effects to the team and improve the team's performance. Such that, the regression model in eq. (22) can then add the factor "the average of the total salaries of the players". Another factor that may affect the team's winning ratio is the fans, because their strong support will be one of the reasons for the players to win. Therefore, another variable that can be added to the model is the "average number of attendance".

5.2 Management implications

According to the empirical results of this study, the winning ratio of a professional team is significantly correlated to its overall offensive performance and overall defensive performance (Martin, 2016; Pavitt, 2011; Yoon and Choi, 2022). The overall offensive performance is significantly positively correlated to team's winning ratio, however, the overall defensive performance is negatively significantly. In competition or training in sports, the evaluation of a team's performance is generally completed by the coaches, agents and the public media, etc. (Park and Kang, 2014). According to the opinion in Park and Kang (2014), a process in evaluating sports performance is shown in the following figure.

Figure 5
Process for evaluating sports performance



In the above-mentioned sports performance evaluation process, a very important point is “Feedback” (Park and Kang, 2014). Whether it is the collection of training data or the analysis of competition results, it is necessary to use a feedback mechanism to allow the coaching team, training team and management to understand the team status and then to facilitate the activation of subsequent support systems (Martin, 2016).

5.2.1 Strategies for offense promotion

First, how does a baseball team create runs scored? Naturally, it is to compete with opponents, improve offensive performance, and create a niche. In the generalized Pythagorean formula, a professional baseball is concerned, finding/cultivating excellent hitters, and promoting players with high on-base rates, etc. These are ways to increase offensive performance. For instance, the Los Angeles Dodgers recruited Shohei Ohtani in 2023, focusing on his hitting performance, which can bring more offensive efficiency to the team. Therefore, the Dodgers arranged Ohtani in the first

three at-bats to facilitate the start being able to attack opponents smoothly and did not allow Ohtani to step onto the mound as a pitcher.

5.2.2 Strategies for defense promotion

In addition, as for how a team reduce losses? For professional baseball teams, good defensive performance is to reduce the opponent's score. Therefore, to find/train pitchers with low (ERA), and to promote players with high blocking rates, etc., can be regarded as a means to improve defensive performance. For example, before the opening game of the 2022-23 season, the Twins and the Padres reached a deal, which will have a significant impact on the strength of the two teams. The transaction package is as follows: the Twins released the 31-year-old left-hand closer Taylor Rogers, the 27-year-old former hitter Brent Rooker was traded to the Padres for 26-year-old right-hander Chris Paddack and right-hander Emilio Pagán, who will turn 31 in May of that year.

Moreover, a good field commander can direct the pitcher's distribution of the ball to deal with the opponent's batters, and guide the defender's defensive zone adjustment to reduce the chance of missing the ball and increase the blocking rate. Such a commander is a catcher. To find a good catcher is also an important defense promotion strategy for the team. For example, on December 13, 2022, a three-party trade occurred in Major League Baseball. The Atlanta Braves, Oakland Athletics and Milwaukee Brewers completed a large-scale transaction involving 9 players. The Braves acquired 28-year-old catcher Sean Murphy, who is 2021 American League Gold Glove players, the Warriors spent a lot of money and sent out 7 players in one go.

5.2.3 Overall strategies

Overall, a sound coaching team is also one of the key factors in the team's success (Soebbing and Washington, 2011). Before the season, the coaching team

formulates a training plan, uses spring training and warm-up games to adjust the players' physical and mental condition and order, and after the start of the season tactical changes. Before the game, the coaching team draws up a combat plan. Players on the field accept the instructions from the coaching team to implement tactics. Offensive tactics and defensive strategies must be balanced. Otherwise, they will lose sight of one and lose their footing. Therefore, hiring an excellent coaching team is also an improvement strategy for the overall strength of the team. For instance, in 2021, Atlanta Braves head coach Brian Snitker who has been consistent for forty-five years, dedicating his life to Braves. Until this year, he led Braves to play all the way to the World Championship. The Braves also defeated the Astros 7-0 in Game 6 of the World Series on November 3 of that season, winning the fourth Gold Cup.

Furthermore, managers are also a very important part in team management (Schyvinck, Babiak, Constandt, and Willem, 2021). In addition to hiring an excellent coaching team, recruiting players, and mediating disputes, managers must also organize a medical team to take care of the players so that they can maintain their best condition and compete. Moreover, due to the emphasis on fans, managers must work harder to manage consumer loyalty to the team, build the team's brand value (Lee, Bang, and Shonk, 2020), and promote players to participate in social responsibility activities to enhance their personal value (Inoue, Kent, & Lee, 2011; Walzel, Robertson, and Anagnostopoulos, 2018). Therefore, good managers can be one of the factors that improves the overall performance of the team.

6. Conclusions

6.1 Limitations

The purpose of this study is to investigate the winning ratio of Major League Baseball teams by proposing the generalized Pythagorean formula. Therefore, the

empirical research data of this study are the season results of the 30 Major League Baseball teams. However, for professional baseball in other regions, such as Chinese Professional Baseball League (CPBL), Korean Professional Baseball League (KBO), Japan Professional Baseball League (NPB), etc., it may not necessarily have the same results. Moreover, for the other professional sports, such as football, hockey, soccer, volleyball, tennis, golf, etc., due to the large difference in properties, the research method proposed in this study may not be applicable, and other models may need to be developed to study it.

6.2 Conclusions

This study proposes a generalized Pythagorean formula in estimating in estimating the winning ratio of MLB teams. By setting a Cobb-Douglas function for the win-loss ratio, the expected winning ratio can be derived by this generalized Pythagorean formula as like as the form of the Pythagorean formula which is first proposed by James (1980). Different to the original form, the generalized Pythagorean formula has different exponents for the factors, a constant to capture the impact of some non-quantitative factors. More important, the generalized Pythagorean formula is extensible by employing more factors that may have effects on the winning ratio.

Moreover, the empirical results show that that each version of Pythagorean formula has a higher accurate rate of predicting the expected winning ratio. The generalized Pythagorean formula does as well as all version of Pythagorean formulae do. In addition, the parameters in the generalized Pythagorean formula can be estimated by using a double log-linear regression model. Such that, it is testable by the historical data.

Furthermore, based on the definition of the Cobb-Douglas function, many other factors that may affect the win-loss ratio can be added to the model, such as ERA, *AVG*, or other factors. And since stochastic efficiency analysis is also based on

Cobb-Douglas function. Therefore, team managers can use the results of this study to explore the team's efficiency performance during the season (Lee, 2011). Moreover, as indicated in Gordon (2020), this method can be used as a reference for team formation in the coming year, such as selecting potential players through the draft, recruiting new players who are low ERA pitchers, high run-batted-in (RBI) hitters or high blocking rate fielders, or player trades to improve the team's performance.

6.3 Future research directions

Future research directions can be divided into two aspects: First, methodological improvements. By introducing artificial intelligence and machine learning, we can find data-driven models and further extend them to the study of winning ratios in other professional sports. The second is practical applications. Based on the results of this study, some data mining methods can be used to find players who are actually in line with the team's plan and recruit them to further achieve the team's goals.

References

Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). Big Ideas in Sports Analytics and Statistical Tools for their Investigation. *WIREs Computational Statistics, 15*(6), e1612.

Boudreaux, C., Ehrlich, J., Ghimire, S., & Sanders, S. (2021). Application of the Pythagorean Expected Wins Percentage and Cross-Validation Methods in Estimating Team Quality. *Mathematics and Sports, 2*(2), 1-8.

Cefis, M. (2023). Observed Heterogeneity in Players' Football Performance Analysis using PLS-PM. *Journal of Applied Statistics, 50*(15), 3088-3107.

Chen, J., & Li, T. F. (2016). The Shrinkage of the Pythagorean Exponents. *Journal of Sports Analytics, 2*(1), 37-48.

Cobb, C. W., & Douglas, P. H. (1928). A Theory of Production. *American Economic Review, 18*(1), 139-165.

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

Dayaratna, K. D., & Miller, S. J. (2012). First Order Approximations of the Pythagorean Win-Loss Formula for Predicting MLB Teams' Winning Percentage. Available from https://doi.org/10.48550/arXiv.1205.4750.

Dayaratna, K. D., & Miller, S. J. (2013). The Pythagorean Win-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey. *The Hockey Research Journal, 16*(1), 193-209.

Dayaratna, K. D., & Miller, S. J. (2016). *Multiple Regression Analysis: Understanding the Impact of Offensive and Defensive Contributions to Team Performance*. Unpublished Lecture Note.

Ehrlich, J., Boudreaux, C., Boudreau, J., & Sanders, S. (2020). An Analysis of an Alternative Pythagorean Expected Win Percentage Model: Applications Using Major League Baseball Team Quality Simulations. *Mathematics and Sports, 1*(1), 5-9.

Elitzur, R. (2020). Data Analytics Effects in Major League Baseball. *Omega: The International Journal of Management Science, 90*(1), 1-13.

Gordon, G. (2020). Pitching or Hitting: Using Baseball's Pythagorean Theorem to Improve

Your Team. *Math Horizons, 28*(2), 12-14.

Greene, W. H. (2018). Econometric Analysis, 8th Edition. Pearson Education Inc.

Hakes, J. K., & Sauer, R. D. (2006). An Economic Evaluation of the Moneyball Hypothesis. *Journal of Economic Perspectives, 20*(3), 173-185.

Heumann, J. (2016). An Improvement to the Baseball Statistic “Pythagorean Wins”. *Journal of Sports Analytics, 2*(1), 49-59.

Hsiao, C. M. (2022). Effects on Second Waves of COVID-19 Epidemics: Social Stringency, Economic Forces, and Public Health. *Theoretical Economics Letters, 12*(1), 287-320.

Hsiao, C. M., Chou, M. S., Lin, C. Y., Lin, K. H., Tsao, Y. W., & Wang, B. Q. (2023). Investigate the Factors for Predicting the Winning Probability? Evidence from English Premier League. 2023 Asian Association of Sport Management (AASM) Conference, 2023/08/19-20, Imperial Hotel Kuching, Sarawak, Malaysia.

Hsiao, C. M., Zhang, H. K., Chen, G. Y., & Chung, T. H. (2022). Estimation of the Probability of Winning for NBA Home Court Teams: A Probit Model Approach. *Journal of Taiwan Society for Sports Management, 22*(1), 73-95.

Inoue, Y., Kent, A., & Lee, S. (2011). CSR and the Bottom Line: Analyzing the Link between CSR and Financial Performance for Professional Teams. *Journal of Sport Management, 25*(6), 531-549.

James, B. (1980). *Baseball Abstract*. Manuscript: Lawrence, KS.

Jang, H., Lee, Y. H., & Fort, R. (2019). Winning in Professional Team Sports: Historical Moments. *Economic Inquiry, 57*(1), 103-120.

Jones, M. A., & Tappin, L. A. (2005). The Pythagorean Theorem of Baseball and Alternative Models. *The UMAP Journal, 26*(2), 23-34.

Joseph, E. (2019). The Statistical Examination of Winning and Succeeding in Sports. *Journal of Mathematics and Statistics, 15*(1), 70-78.

Kaplan, E. H., & Rich, C. (2017). Decomposing Pythagoras. *Journal of Quantitative Analysis in Sports, 13*(4), 141-149.

Lee, J. T. (2014). Estimation of Exponent Value for Pythagorean Method in Korean Pro-Baseball. *Journal of the Korean Data and Information Science Society, 25*(3), 493-499.

Lee, Y. H. (2011). Is the Small-Ball Strategy Effective in Winning Games? A Stochastic

Frontier Production Approach. *Journal of Productivity Analysis, 35*(1), 51-59.

Lee, C., Bang, H., & Shonk, D. J. (2020). Professional Team Sports Organizations' Corporate Social Responsibility Activities: Corporate Image and Chosen Communication Outlets' Influence on Consumers' Reactions. *International Journal of Sport Communication, 14*(2), 280-297.

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton and Company, New York.

Martin, L. (2016). Sports Performance Measurement and Analytics. Pearson Education, Inc.

McGoldrick, K., & Voeks, L. (2005). We Got Game: An Analysis of Win-Loss Probability and Efficiency Differences between the NBA and WNBA. *Journal of Sports Economics, 6*(1), 5-23.

Miller, S. J. (2007). A Derivation of the Pythagorean Win-Loss Formula in Baseball. *Chance Magazine, 20*(1), 40-48.

Miller, S. J., Corcoran, T., Gossels, J., Luo, V., & Porfilio, J. (2014). Pythagoras at the Bat. *In Social Networks and the Economics of Sports* edited by Pardalos P. M. & Zamaraev V., 89-114.

Mizels, J., Erickson, B., & Chalmers, P. (2022). Current State of Data and Analytics Research in Baseball. *Current Reviews in Musculoskeletal Medicine, 15*(4), 283-290.

Park, J. H., & Kang, M. S. (2014). Evaluation in Sports Performance. *In Social Networks and the Economics of Sports*, edited by Pardalos P. M. & Zamaraev V., 75-88.

Pavitt, C. (2011). An Estimate of How Hitting, Pitching, Fielding, and Basestealing Impact Team Winning Percentage in Baseball. J*ournal of Quantitative Analysis in Sports, 7*(4), 1-20.

Porter, P. K., & Scully, G. W. (1982). Measuring Managerial Efficiency: The Case of Baseball. *Southern Economic Journal, 48*(3), 642-650.

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

Rothman, S. (2014). A New Formula to Predict a Team's Winning Percentage. *The Baseball Research Journal, 43*(2), 97-105.

Ruggiero, J., Hadley, L., Ruggiero, G., & Knowles, S. (1998). A Note on the Pythagorean

Theorem of Baseball Production. *Management and Decision Economics, 18*(4), 335-342.

Sarlis, V., & Tjortjis, C. (2020). Sports Analytics - Evaluation of Basketball Players and Team Performance. *Information Systems, 93*, 101562.

Schyvinck, C., Babiak, K., Constandt, B., & Willem, A. (2021). What Does Entrepreneurship Add to the Understanding of Corporate Social Responsibility Management in Sport? *Journal of Sport Management, 35*(5), 452-464.

Scully, G. W. (1974). Pay and Performance in Major League Baseball. *American Economic Review, 64*(6), 915-930.

Senevirathne, H. K. W., & Manage, A. B. W. (2021). Predicting the Winning Percentage of Limited-Overs Cricket using the Pythagorean Formula. *Journal of Sports Analytics, 7*(3), 169-183.

Soebbing, B. P., & Washington, M. (2011). Leadership Succession and Organizational Performance: Football Coaches and Organizational Issues. *Journal of Sport Management, 25*(6), 550-561.

Valero, S. C. (2016). Predicting Win-Loss Outcomes in MLB Regular Season Games – A Comparative Study using Data Mining Methods. *International Journal of Computer Science in Sport, 15*(2), 91-112.

Walzel, S., Robertson, J., & Anagnostopoulos, C. (2018). Corporate Social Responsibility in Professional Team Sports Organizations: An Integrative Review. *Journal of Sport Management, 32*(6), 511-530.

Wen, F. H., & Chiou, H. J. (2009). Multilevel Moderated Mediation of Organizational Study: An Empirical Analysis of Organizational Innovation Climate, Organizational Commitment and Job Satisfaction. *Journal of Management, 26*(2), 189-211.

Winston, W. L., Nestler, S., & Pelechrinis, K. (2022). *Mathletics: How Gamblers, Managers, and Fans Use Mathematics in Sports, 2nd edition. Princeton University Press*. Chapter 1, 3-11.

Yang, T. Y., & Swartz, T. (2004). A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball. *Journal of Data Science, 2*(1), 61-73.

Yoon, J. H., & Choi, S. H. (2022). Pythagorean Exponents Induced by Mathematical and

Statistical Methods in the Major League Baseball. *International Journal of Fuzzy Logic and Intelligent Systems, 22*(3), 245-251.

Zech, C. E. (1981). An Empirical Estimation of a Production Function: The Case of Major League Baseball. *The American Economist, 25*(2), 19-23.

國立臺灣體育運動大學學報
第 13 卷第 2 期（113.06），21-54 頁

運用廣義畢氏公式估計美國職棒大聯盟球隊的勝率

蕭秋銘
國立雲林科技大學財務金融系

摘要

本文欲藉由廣義畢氏公式法來估計美國職棒大聯盟常規賽的勝率。經由過去 10 年賽季 30 支美國職棒大聯盟球隊的攻守紀錄，本文驗證了廣義畢氏公式對於球隊的勝率具有極高的預測能力。實證研究的結果亦顯示，本文的多元迴歸模型相或是二階段迴歸對於 Poisson 迴歸模型都有較高的預估能力。而且不管是樣本內估計，還是樣本外預測的 MSE 與 MAE，本文所提出的廣義畢氏公式對於勝率的估計都不比其他版本的畢氏公式差。然而，由於本文的廣義畢氏公式是經由 Cobb-Douglas 產出函數推導而得，因此不僅具備經濟意涵，可以透過計量模型檢驗之，而且此模型更具有彈性 (flexibility) 與擴充性 (extensibility) 。再者，球團管理階層亦可透過此公式招募所需的球員，以便提升戰力，提高勝率。

關鍵詞：棒球畢氏公式、Cobb-Douglas 產出函數、勝敗比