Confidence Intervals for the Pythagorean Formula in Baseball

David D. Tung[1]

Abstract

In this paper, we will investigate the problem of obtaining confidence intervals for a baseball team's Pythagorean expectation, i.e. their expected winning percentage and expected games won. We study this problem from two different perspectives. First, in the framework of regression models, we obtain confidence intervals for prediction, i.e. more formally, prediction intervals for a new observation, on the basis of historical binomial data for Major League Baseball teams from the 1901 through 2009 seasons, and apply this to the 2009 MLB regular season. We also obtain a Scheffé-type simultaneous prediction band and use it to tabulate predicted winning percentages and their prediction intervals, corresponding to a range of values for $\log(RS/RA)$. Second, parametric bootstrap simulation is introduced as a data-driven, computer-intensive approach to numerically computing confidence intervals for a team's expected winning percentage. Under the assumption that runs scored per game and runs allowed per game are random variables following independent Weibull distributions, we numerically calculate confidence intervals for the Pythagorean expectation via parametric bootstrap simulation on the basis of each team's runs scored per game and runs allowed per game from the 2009 MLB regular season. The interval estimates, from either framework, allow us to infer with better certainty as to which teams are performing above or below expectations. It is seen that the bootstrap confidence intervals appear to be better at detecting which teams are performing above or below expectations than the prediction intervals obtained in the regression framework.

Keywords: Pythagorean expectation, baseball, sabermetrics, logistic regression, linear regression, confidence intervals, prediction intervals, Weibull distribution, likelihood inference, maximum likelihood estimation, parametric bootstrap simulation, bootstrap confidence intervals.

[1]E-mail: david.deming.tung@gmail.com

1 Introduction

A statistical model of a baseball team's expected winning percentage is given by the so-called "Pythagorean formula," which is given by

$$\pi = \frac{RS^{\lambda}}{RS^{\lambda} + RA^{\lambda}} = \frac{1}{1 + \left(\frac{RA}{RS}\right)^{\lambda}}, \quad \lambda > 0. \tag{1}$$

Here π is the expected winning percentage with RS and RA respectively denoting the observed runs scored and runs allowed totals, and λ is a constant parameter. The Pythagorean formula first appeared in Bill James' baseball abstracts of the early 1980's (cf. with James (1983)), and is generally used to determine if a baseball team is performing above or below expectations. An exponent of $\lambda = 2$ was originally used by James and because the denominator of that formula reminded him of the Pythagorean theorem in Euclidean geometry, the name, for better or worse, stuck. The best fitting exponent currently is about $\lambda = 1.86$.

There is a large body of literature where authors modify the Pythagorean formula. For instance, Vollmayr-Lee (2002) models expected winning percentage in terms of $u = \frac{RS}{RS+RA}$ by rewriting (1) as

$$\pi(u) = \frac{u^{\lambda}}{u^{\lambda} + (1-u)^{\lambda}} = \frac{\left(\frac{RS}{RS+RA}\right)^{\lambda}}{\left(\frac{RS}{RS+RA}\right)^{\lambda} + \left(1 - \frac{RS}{RS+RA}\right)^{\lambda}}, \quad \lambda > 0 \tag{2}$$

then considers higher-order Taylor approximations of $\pi(\cdot)$ about the point $u_0 = 1/2$. Miller (2006) provides a theoretical framework for the Pythagorean formula by assuming that runs scored per game and runs allowed per game are random variables following independent Weibull distributions. Davenport and Woolner (1999), Keri (2007) and Cochran (2008) each investigate the Pythagorean formula for specific circumstances and find optimal values for the exponent λ, which varies between 1.74 and 2.0 depending on the league, number of seasons, and time period under consideration. Braunstein (2010) demonstrates that there is a strong correlation between Pythagorean residuals and run distribution consistency, and from the latter, constructs a simple regression estimator that improves Pythagorean estimators in terms of root mean square error and the coefficient of determination.

The Pythagorean formula has become so popular that sports mediums, including ESPN, FOX Sports, Baseball-Reference.com, and MLB.com all make reference to the Pythagorean expectation. Note that we use the term Pythagorean expectation to refer to both the expected winning percentage, and the expected number of games won. It is rather surprising to see how little to nothing has been done to address the question of confidence intervals

for the Pythagorean expectation. If the purpose of the Pythagorean formula is to determine whether a baseball team is performing above or below expectations, then it is useful to have sensible and reliable confidence intervals for the Pythagorean expectation to complement any corresponding point estimate. Such interval estimates allow us to infer with better certainty as to which teams are performing above or below expectations, and infer with a measure of confidence that the Pythagorean expectation is within the bounds of its confidence interval. Moreover, such confidence intervals would certainly provide even more illumination to those who follow the sport on a regular basis.

In Section 2, we review the Pythagorean formula in relationship to both logistic and linear regression. In the regression setting, confidence intervals for predictions, i.e. prediction intervals more formally, of a team's winning percentage are obtained. In Section 3, parametric bootstrap simulation is introduced as a data-driven, computer-intensive approach to numerically computing confidence intervals for a team's expected winning percentage. For the reader's convenience, below are the final standings in both the American and National League from the 2009 regular season.

Team	Won	Lost	Win%	GB	RS	RA
New York Yankees	103	59	0.636	-	915	753
Boston Red Sox	95	67	0.586	8	872	736
Tampa Bay Rays	84	78	0.519	19	803	754
Toronto Blue Jays	75	87	0.463	28	798	771
Baltimore Orioles	64	98	0.395	39	741	876
Minnesota Twins	87	76	0.534	-	817	765
Detroit Tigers	86	77	0.528	1	743	745
Chicago White Sox	79	83	0.488	7.5	724	732
Cleveland Indians	65	97	0.401	21.5	773	865
Kansas City Royals	65	97	0.401	21.5	686	842
Anaheim Angels	97	65	0.599	-	883	761
Texas Rangers	87	75	0.537	10	784	740
Seattle Mariners	85	77	0.525	12	640	692
Oakland Athletics	75	87	0.463	22	759	761

Table 1. Final Standings for the 2009 American League Regular Season.

Team	Won	Lost	Win%	GB	RS	RA
Philadelphia Phillies	93	69	0.574	-	820	709
Florida Marlins	87	75	0.537	6	772	766
Atlanta Braves	86	76	0.531	7	735	641
New York Mets	70	92	0.432	23	671	757
Washington Nationals	59	103	0.364	34	710	874
St. Louis Cardinals	91	71	0.562	-	730	640
Chicago Cubs	83	78	0.516	7.5	707	672
Milwaukee Brewers	80	82	0.494	11	785	818
Cincinnati Reds	78	84	0.481	13	673	723
Houston Astros	74	88	0.457	17	643	770
Pittsburgh Pirates	62	99	0.385	28.5	636	768
Los Angeles Dodgers	95	67	0.586	-	780	611
Colorado Rockies	92	70	0.568	3	804	715
San Francisco Giants	88	74	0.543	7	657	611
San Diego Padres	75	87	0.463	20	638	769
Arizona Diamondbacks	70	92	0.432	25	720	782

Table 2. Final Standings for the 2009 National League Regular Season.

2 Pythagorean Expectation and Regression Models

2.1 Logistic Regression

From a statistical perspective, the Pythagorean formula is a logistic regression model. The Pythagorean exponent λ is an unknown parameter which can be estimated by fitting a logistic regression model to a large historical data set consisting of the seasonal won-lost records and corresponding runs scored and runs allowed totals, i.e. (W, L, RS, RA). A data set consisting of the 1871 through 2006 seasons can be found in Sean Lahman's baseball database at http://baseball1.com/statistics. We used a large part of this and data from recent seasons to form a historical data set consisting of the seasons 1901 through 2009.

Let N denote the number of teams contained in the historical data set. Let W_j denote the number of games won by team j in their season of n_j games. In the framework of logistic regression, the $\{W_j : j = 1, 2, \ldots N\}$ are independent Binomial(n_j, π_j) random variables and $p_j = W_j/n_j$ is the observed winning percentage of team j and their expected winning percentage

is the unknown Binomial success probability

$$\pi_j = \mathbb{E}(p_j \mid RS_j, RA_j) = \mathbb{E}\left(\frac{W_j}{n_j} \,\middle|\, RS_j, RA_j\right). \tag{3}$$

Logistic regression is used to model binomial data, which can come either in the form of observed successes and failures, or observed proportions. Such models belong to a class of linear statistical models known as generalized linear models (GLM) (cf. with Dobson (2002)). In the GLM framework, the goal is to model the unknown Binomial success probability π_j as a function of the covariates, i.e. we assume there is a function $g(\cdot)$ called a "link function," which simply describes how π_j depends on the linear predictor, e.g. $g(\pi_j) = \beta_0 + \beta_1(RS_j \times RA_j)$.

The Pythagorean expectation assumes a team's expected winning percentage or mean response has the form

$$\pi_j = \frac{RS_j^\lambda}{RS_j^\lambda + RA_j^\lambda} = \frac{\exp[\lambda\,(\log RS_j - \log RA_j)]}{1 + \exp[\lambda\,(\log RS_j - \log RA_j)]}. \tag{4}$$

Then the expected odds is simply the ratio between a team's expected winning and losing percentage, i.e.

$$\frac{\pi_j}{1-\pi_j} = \left(\frac{RS_j}{RA_j}\right)^\lambda. \tag{5}$$

Taking logarithms gives the log-expected odds or logit mean response

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \lambda\,\log\left(\frac{RS_j}{RA_j}\right) \tag{6}$$

which corresponds to the logit link function $g(\pi_j) = \log\left(\frac{\pi_j}{1-\pi_j}\right)$. The predictor variable of interest here is $\log(RS/RA)$. To be a bit more precise, we can include an intercept term β_0 in the linear predictor of the logit mean response, i.e.

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \lambda\,\log\left(\frac{RS_j}{RA_j}\right). \tag{7}$$

When we fit a logistic regression model to the historical binomial data, we obtain the fitted logit mean response model. In other words, given point estimates $\widehat{\beta}_0$ and $\widehat{\lambda}$, respectively for the intercept β_0 and the Pythagorean exponent λ, we have

$$\log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right) = \widehat{\beta}_0 + \widehat{\lambda}\,\log\left(\frac{RS}{RA}\right) \tag{8}$$

which is an estimate of the logit mean response. Note that the Pythagorean exponent λ can be interpreted as the change in the log-expected odds of the unknown expected winning percentage π corresponding to a unit increase in $\log(RS/RA)$. Moreover, using the inverse logit transformation gives us the fitted mean response

$$\widehat{\pi} = \frac{\exp\left[\widehat{\beta}_0 + \widehat{\lambda}\left(\log RS - \log RA\right)\right]}{1 + \exp\left[\widehat{\beta}_0 + \widehat{\lambda}\left(\log RS - \log RA\right)\right]} = \frac{\exp(\widehat{\beta}_0)\, RS^{\widehat{\lambda}}}{\exp(\widehat{\beta}_0)\, RS^{\widehat{\lambda}} + RA^{\widehat{\lambda}}} \tag{9}$$

which is our estimate for a team's expected winning percentage.

As an illustration, we fit the logistic regression model to the historical binomial data, which consists of $N = 2242$ teams. The statistical analysis is done in the `R` statistical environment (cf. R Development Core Team (2008)) From the logistic regression output, the intercept estimate is $\widehat{\beta}_0 = -0.0009753$, which is practically zero. The p-value for the intercept is 0.776, which indicates that the intercept term is not at all statistically significant. Thus, we are free to dispense with the intercept term. The point estimate for the Pythagorean exponent is roughly $\widehat{\lambda} = 1.86$. Overall, the fitted mean response model has the form

$$\widehat{\pi} = \frac{\exp\left[\widehat{\lambda}\left(\log RS - \log RA\right)\right]}{1 + \exp\left[\widehat{\lambda}\left(\log RS - \log RA\right)\right]} = \frac{RS^{1.86}}{RS^{1.86} + RA^{1.86}}. \tag{10}$$

The mean absolute difference between the observed and predicted games won is 3.231583 games, and the standard deviation of the absolute difference is 2.418614 games. The root mean square difference between the observed and predicted games won is 4.036113 games.

Coefficient	Estimate	Std.Error	z-value	p-value
Intercept	-0.0009753	0.0034263	-0.285	0.776
$\log(RS/RA)$	1.8603399	0.0203030	91.629	2e-16

Table 3. Logistic regression summary.

2.2 Linear Regression

In logistic regression, we assumed that the log-expected odds, i.e. the logit mean response, has the form

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \lambda \log\left(\frac{RS_j}{RA_j}\right). \tag{11}$$

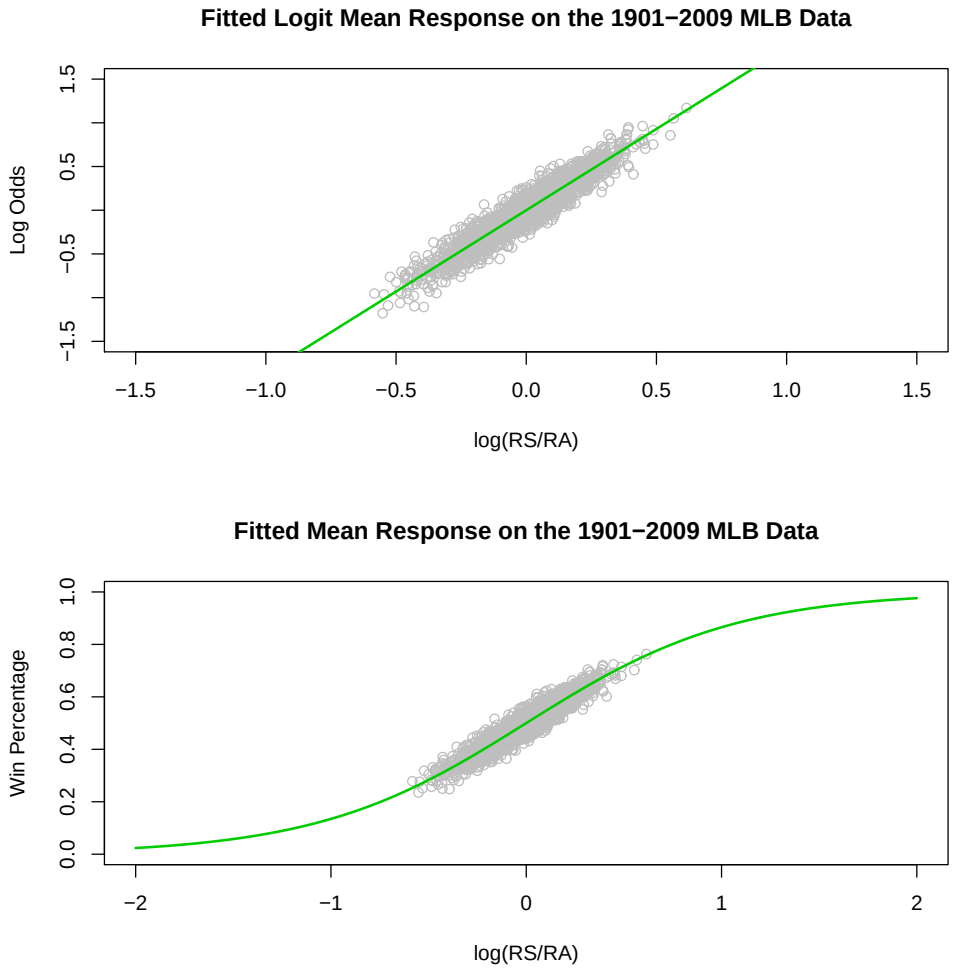

Figure 1: Logistic regression model fit for the 1901-2009 MLB data.

Now, let the random variable $Y_j = \log\left(\frac{p_j}{1-p_j}\right)$ denote the observed log-odds for team j in the historical binomial data. In linear regression, the conditional Normal model assumes that

$$Y_j = \log\left(\frac{p_j}{1-p_j}\right) = \delta_0 + \kappa \log\left(\frac{RS_j}{RA_j}\right) + \epsilon_j \tag{12}$$

where δ_0 and κ are respectively the intercept and slope of the linear regression model, and ϵ_j are independent $N_1(0, \sigma^2)$ random variables. The linear regression mean response has the form

$$\mathbb{E}(Y_j \,|\, RS_j, RA_j) = \mathbb{E}\left[\log\left(\frac{p_j}{1-p_j}\right) \,\middle|\, RS_j, RA_j\right] = \delta_0 + \kappa \log\left(\frac{RS_j}{RA_j}\right). \tag{13}$$

When we fit a linear regression model to the data, we obtain

$$\widehat{Y} = \widehat{\delta}_0 + \widehat{\kappa} \log\left(\frac{RS}{RA}\right) \tag{14}$$

which is an estimate of the linear regression mean response, given point estimates $\widehat{\delta}_0$ and $\widehat{\kappa}$, respectively for δ_0 and κ.

For the historical binomial data, it turns out that we can use linear regression as an approximation to the logistic regression model. We will see that a linear regression model fitted to the historical binomial data set should result as an approximation to the fitted logistic regression model, much in the same way that a Normal distribution can be used to approximate a Binomial distribution. We now attempt to rigorously justify the linear regression approximation to logistic regression.

From the DeMoivre-Laplace Central Limit Theorem, i.e. the Normal Approximation to the Binomial, we have in the limit as $n_j \to \infty$,

$$\sqrt{n_j}(p_j - \pi_j) \xrightarrow{D} N_1(0, \pi_j(1-\pi_j)). \tag{15}$$

Here, the symbol $\xrightarrow{D}$ indicates convergence in distribution (cf. with Resnick (2001) for a definition). Moreover, by the Delta Method, we have in the limit as $n_j \to \infty$,

$$\begin{aligned}\sqrt{n_j}\left[Y_j - \log\left(\frac{\pi_j}{1-\pi_j}\right)\right] &= \sqrt{n_j}\left[\log\left(\frac{p_j}{1-p_j}\right) - \log\left(\frac{\pi_j}{1-\pi_j}\right)\right] \\ &\xrightarrow{D} N_1\left(0, \frac{1}{\pi_j(1-\pi_j)}\right)\end{aligned} \tag{16}$$

i.e. the difference between the observed log-odds and the log-expected odds, when suitably normalized, converges in distribution to a limit random variable

having a Normal distribution with mean zero and variance $\frac{1}{\pi_j(1-\pi_j)}$. Therefore, when n_j is sufficiently large enough, we have

$$\mathbb{E}(Y_j \,|\, RS_j, RA_j) - \log\left(\frac{\pi_j}{1-\pi_j}\right) = \delta_0 - \beta_0 + (\kappa - \lambda)\,\log\left(\frac{RS_j}{RA_j}\right) \approx 0. \quad (17)$$

In other words, (17) says the difference between the linear regression mean response and the logit mean response is approximately zero, when n_j, the number of games played by team j, is large enough. This justifies using linear regression as an approximation to logistic regression. In terms of the fitted linear and logistic regression models, their difference should also be approximately zero, i.e.

$$\widehat{Y} - \log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right) = \widehat{\delta}_0 - \widehat{\beta}_0 + (\widehat{\kappa} - \widehat{\lambda})\,\log\left(\frac{RS}{RA}\right) \approx 0. \quad (18)$$

We fit the linear regression model to the historical binomial data. From the linear regression output, the intercept estimate is $\widehat{\delta}_0 = -0.001144$, which is practically zero. The p-value for the intercept is 0.614, thus indicating that the intercept term is not statistically significant. Thus, we are free to dispense with the intercept term. The point estimate for the Pythagorean exponent is about $\widehat{\kappa} = 1.86$. Overall, the fitted mean response model has the form

$$\widehat{\pi} = \frac{\exp\left[\widehat{\kappa}\,(\log RS - \log RA)\right]}{1 + \exp\left[\widehat{\kappa}\,(\log RS - \log RA)\right]} = \frac{RS^{1.86}}{RS^{1.86} + RA^{1.86}}. \quad (19)$$

The mean absolute difference between the observed and predicted games won is 3.231443 games and the standard deviation of the absolute difference is 2.418995 games. The root mean square difference between the observed and predicted games won is 4.036229 games. The sample correlation between the logarithm of the observed odds and the logarithm of runs scored totals over runs allowed totals is 0.95, which indicates that the Pythagorean expectation formula correlates very well with a baseball team's actual performance.

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	-0.001144	0.002267	-0.504	0.614
$\log(RS/RA)$	1.863569	0.013051	142.791	2e-16

Table 4. Linear regression summary.

2.3 Confidence Intervals for Prediction

Prediction is a type of statistical inference that is of interest in the regression framework. In particular, the goal is to make a prediction on the unobserved

response variable. A prediction interval is an interval on a random variable, not a parameter. Since random variables have more variation than parameters, which are fixed constants, one generally expects prediction intervals to be wider than confidence intervals of the same confidence level. In the logistic regression framework, there is no distinction possible between confidence intervals for a future observation and those for the mean response (cf. Faraway (2006), pg. 42). Therefore, in order to obtain useful confidence intervals for prediction, we must do so through the linear regression framework.

We assume that Y_0 is a new observation on the response variable $Y = \log\left(\frac{p}{1-p}\right)$ to be taken at $x_0 = \log\left(\frac{RS_0}{RA_0}\right)$. From linear regression theory, it is well-known that a $(1-\alpha)$ prediction interval for a new observation Y_0 (cf. with Casella and Berger (2002) and Kutner and Neter (2004)) is given by

$$(\widehat{\delta}_0 + \widehat{\kappa} x_0) \pm t_{N-2,\alpha/2} \cdot S \sqrt{1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{20}$$

where

$$S^2 = \frac{1}{N-2} \sum_{j=1}^{N} (y_j - \widehat{\delta}_0 - \widehat{\kappa} x_j)^2 \tag{21}$$

$$S_{xx} = \sum_{j=1}^{N} (x_j - \bar{x})^2. \tag{22}$$

To obtain the corresponding prediction interval for a new observation's winning percentage, the above prediction interval must be converted from the logit scale by the inverse logit transformation.

We can also obtain a prediction band to make inferences for all values of $x = \log(RS/RA)$. A $(1-\alpha)$ Scheffé-type simultaneous prediction band for $\widehat{Y} = \widehat{\delta}_0 + \widehat{\kappa} x$ has the form

$$(\widehat{\delta}_0 + \widehat{\kappa} x) \pm C(\alpha) \cdot S \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}}} \tag{23}$$

which holds simultaneously for all $x = \log(RS/RA)$, where $C(\alpha) = \sqrt{F_{v,N-2,\alpha}}$ and $v = \frac{(N+2)^2}{(N+1)^2+1}$. For completeness, we provide a derivation.

It is enough to find $C(\alpha) > 0$, for which

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \frac{[(\widehat{\delta}_0 + \widehat{\kappa} x) - (\delta_0 + \kappa x)]^2}{S^2 \left[1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}\right]} \leq C^2(\alpha) \right) = 1 - \alpha. \tag{24}$$

To make the above maximization easier, we can use a well-known reparameterization which results in independent estimators for δ_0 and κ. Put

$$\widehat{\delta}_0 + \widehat{\kappa}x = \bar{Y} + \widehat{\kappa}(x - \bar{x}) \tag{25}$$

$$\delta_0 + \kappa x = \delta_0 + \kappa\bar{x} + \kappa(x - \bar{x}). \tag{26}$$

and for notational convenience use $t = x - \bar{x}$. Then we obtain

$$\frac{[(\widehat{\delta}_0 + \widehat{\kappa}x) - (\delta_0 + \kappa x)]^2}{S^2\left[1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}\right]} = \frac{[(\bar{Y} - \delta_0 - \kappa\bar{x}) + (\widehat{\kappa} - \kappa)t]^2}{S^2\left[1 + \frac{1}{N} + \frac{t^2}{S_{xx}}\right]}. \tag{27}$$

The distribution of the maximum is not easy to write, but can be approximated. It can be shown using straightforward calculus that the maximum

$$\sup_{t\in\mathbb{R}} \frac{[(\bar{Y} - \delta_0 - \kappa\bar{x}) + (\widehat{\kappa} - \kappa)t]^2}{S^2\left[1 + \frac{1}{N} + \frac{t^2}{S_{xx}}\right]} = \frac{\frac{1}{N+1} \cdot \frac{(\bar{Y}-\delta_0-\kappa\bar{x})^2}{\sigma^2/N} + \frac{(\widehat{\kappa}-\kappa)^2}{\sigma^2/S_{xx}}}{S^2/\sigma^2}. \tag{28}$$

The numerator is a weighted sum of independent Chi-square random variables, i.e. $(\frac{1}{N+1})\chi_1^2$ and χ_1^2, and can be approximated by a χ_v^2/v distribution, while the denominator has a $\chi_{N-2}^2/(N-2)$ distribution. The degrees of freedom v can be approximated by the well-known Welch-Satterthwaite approximation (e.g. cf. with Casella and Berger (2002)), which gives

$$v = \frac{\left(\frac{1}{N+1}\chi_1^2 + \chi_1^2\right)^2}{\left(\frac{1}{N+1}\right)^2(\chi_1^2)^2 + (\chi_1^2)^2} = \frac{(N+2)^2}{(N+1)^2+1} \to 1, \text{ as } N \to \infty. \tag{29}$$

Therefore, we have

$$\sup_{t\in\mathbb{R}} \frac{[(\bar{Y} - \delta_0 - \kappa\bar{x}) + (\widehat{\kappa} - \kappa)t]^2}{S^2\left[1 + \frac{1}{N} + \frac{t^2}{S_{xx}}\right]} \asymp \frac{\chi_1^2}{\chi_{N-2}^2/(N-2)} \simeq F_{1,N-2}. \tag{30}$$

We have used the notation $\asymp$ to denote an approximate distribution and $\simeq$ to denote distributional equivalence. Recall that the Fisher-Snedecor F-distribution, with degrees of freedom 1 and q, is the square of Student's t-distribution with q degrees of freedom, i.e. $F_{1,q,\alpha} = t_{q,\alpha/2}^2$, and thus $C(\alpha) = \sqrt{F_{1,N-2,\alpha}} = t_{N-2,\alpha/2}$. Therefore, a $(1-\alpha)$ Scheffé-type simultaneous prediction band for $\widehat{Y} = \widehat{\delta}_0 + \widehat{\kappa}x$ has the form

$$(\widehat{\delta}_0 + \widehat{\kappa}x) \pm \sqrt{F_{1,N-2,\alpha}} \cdot S\sqrt{1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{S_{xx}}} \tag{31}$$

and may be rewritten as

$$(\widehat{\delta}_0 + \widehat{\kappa}x) \pm t_{N-2,\alpha/2} \cdot S\sqrt{1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}}}. \tag{32}$$

Again, to obtain the corresponding prediction band for predicted winning percentages, the above prediction band must be converted from the logit scale by the inverse logit transformation.

In a linear regression analysis, a prediction interval for a new observation can be obtained from most standard statistical packages, such as `R`. It is also fairly straightforward to numerically obtain the Scheffé-type simultaneous prediction band using the `R` function `predict` (cf. with Faraway (2005)). Figure 2 displays a 95% Scheffé-type simultaneous prediction band. From this simultaneous prediction band, we tabulate some predicted winning percentages, based on the Pythagorean formula, and their prediction intervals corresponding to a range of values for $\log(RS/RA)$. This "Pythagorean table" is very convenient and makes the 95% Scheffé-type simultaneous prediction band accessible for practical use.

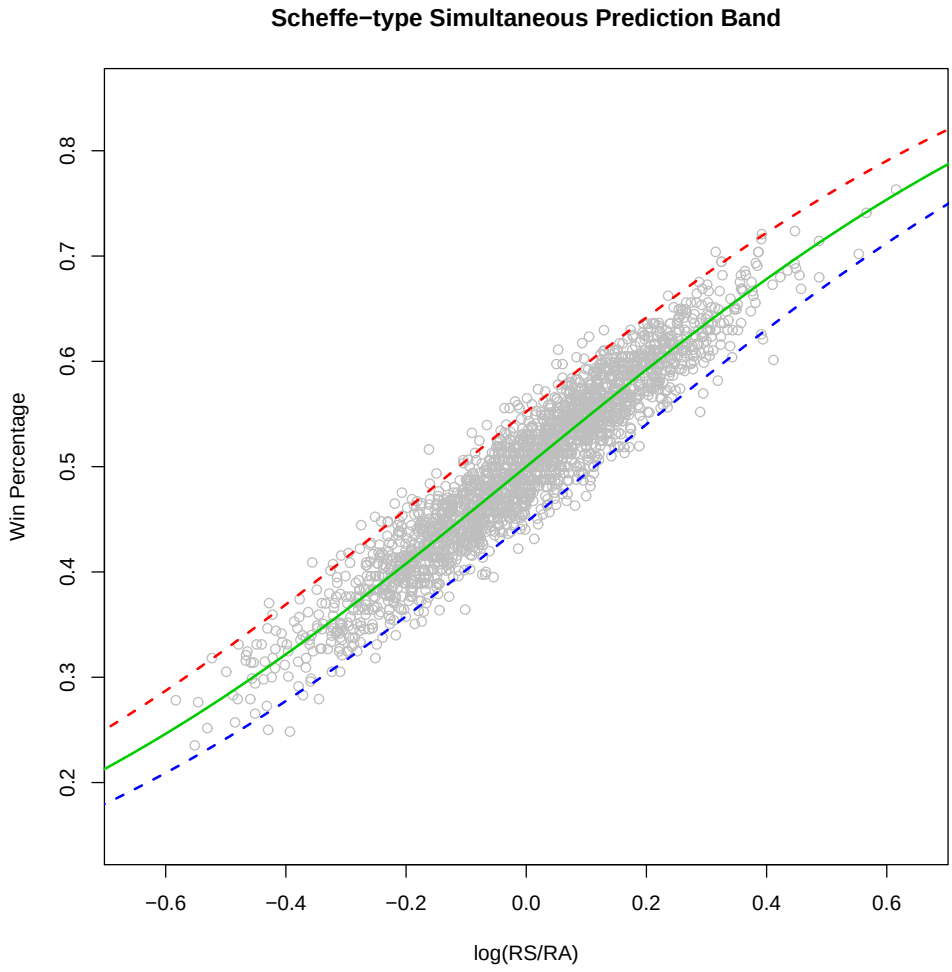

Figure 2: 95% Scheffé-type simultaneous prediction band.

Prd.Win%	PI for Win%	$\log\left(\frac{RS}{RA}\right)$	Prd.Win%	PI for Win%	$\log\left(\frac{RS}{RA}\right)$
0.670	(0.621, 0.715)	0.380	0.500	(0.447, 0.552)	0.000
0.665	(0.616, 0.710)	0.368	0.495	(0.442, 0.547)	-0.011
0.660	(0.611, 0.705)	0.355	0.490	(0.437, 0.542)	-0.022
0.655	(0.606, 0.701)	0.344	0.485	(0.433, 0.537)	-0.032
0.650	(0.600, 0.696)	0.332	0.480	(0.428, 0.532)	-0.043
0.645	(0.595, 0.691)	0.320	0.475	(0.423, 0.527)	-0.054
0.640	(0.590, 0.687)	0.308	0.470	(0.418, 0.522)	-0.065
0.635	(0.584, 0.682)	0.297	0.465	(0.413, 0.517)	-0.075
0.630	(0.579, 0.677)	0.285	0.460	(0.408, 0.512)	-0.086
0.625	(0.574, 0.672)	0.274	0.455	(0.403, 0.507)	-0.097
0.620	(0.569, 0.668)	0.263	0.450	(0.398, 0.502)	-0.108
0.615	(0.564, 0.663)	0.251	0.445	(0.393, 0.497)	-0.119
0.610	(0.558, 0.658)	0.240	0.440	(0.389, 0.492)	-0.130
0.605	(0.553, 0.654)	0.229	0.435	(0.384, 0.487)	-0.140
0.600	(0.548, 0.649)	0.217	0.430	(0.379, 0.482)	-0.151
0.595	(0.543, 0.644)	0.206	0.425	(0.374, 0.477)	-0.163
0.590	(0.538, 0.639)	0.195	0.420	(0.369, 0.472)	-0.174
0.585	(0.533, 0.635)	0.184	0.415	(0.365, 0.466)	-0.184
0.580	(0.528, 0.630)	0.173	0.410	(0.360, 0.461)	-0.195
0.575	(0.523, 0.625)	0.162	0.405	(0.355, 0.456)	-0.207
0.570	(0.517, 0.620)	0.151	0.400	(0.350, 0.451)	-0.218
0.565	(0.512, 0.615)	0.140	0.395	(0.345, 0.446)	-0.229
0.560	(0.507, 0.611)	0.129	0.390	(0.341, 0.441)	-0.240
0.555	(0.502, 0.606)	0.118	0.385	(0.336, 0.436)	-0.252
0.550	(0.497, 0.601)	0.107	0.380	(0.331, 0.430)	-0.263
0.545	(0.492, 0.596)	0.097	0.375	(0.327, 0.425)	-0.274
0.540	(0.487, 0.591)	0.086	0.370	(0.322, 0.420)	-0.286
0.535	(0.482, 0.586)	0.075	0.365	(0.317, 0.415)	-0.297
0.530	(0.477, 0.581)	0.064	0.360	(0.313, 0.410)	-0.309
0.525	(0.472, 0.577)	0.053	0.355	(0.308, 0.404)	-0.321
0.520	(0.467, 0.572)	0.043	0.350	(0.303, 0.399)	-0.333
0.515	(0.462, 0.567)	0.032	0.345	(0.299, 0.394)	-0.344
0.510	(0.457, 0.562)	0.021	0.340	(0.294, 0.388)	-0.356
0.505	(0.452, 0.557)	0.011	0.335	(0.289, 0.383)	-0.368
0.500	(0.447, 0.552)	0.000	0.330	(0.285, 0.378)	-0.380

Table 5. A Pythagorean Table based on the 95% Scheffé-type simultaneous prediction band.

2.4 Numerical Results from the Regression Framework

Using the fitted linear regression model, over the 30 teams from the 2009 MLB regular season, the mean absolute difference between observed and predicted games won is 3.94 games with a standard deviation of 2.73 games. The root mean square difference between observed and predicted games won is 4.77 games. These results are consistent with the observation that the Pythagorean Formula is usually accurate to about four games.

The difference between the predicted and observed games won is a measure of a team's performance relative to their predicted expectation. Large negative (positive) values of this difference seem to indicate a team is performing above (below) predicted expectations. We can arbitrarily or heuristically classify teams as "overachieving" or "underachieving" if they perform 5 games above or below predicted expectation, which sounds reasonable. In the American League, these so-called "overachieving" teams are the New York Yankees (-7.48), and the Seattle Mariners (-9.87), while the "underachieving" teams were the Toronto Blue Jays (8.59), Cleveland Indians (7.56), and Oakland Athletics (5.80). In the National League, these so-called "overachieving" teams were the Florida Marlins (-5.41), Houston Astros (-6.45), and San Diego Padres (-7.93), while the "underachieving" teams were the Atlanta Braves (5.25) and Washington Nationals (6.54).

Team	Won	Prd.Won	Prd.Win%	Win%	Diff.
New York Yankees	103	95.52	0.590	0.636	-7.48
Boston Red Sox	95	93.67	0.578	0.586	-1.33
Tampa Bay Rays	84	85.74	0.529	0.519	1.74
Toronto Blue Jays	75	83.59	0.516	0.463	8.59
Baltimore Orioles	64	68.49	0.423	0.395	4.49
Minnesota Twins	87	86.48	0.531	0.534	-0.52
Detroit Tigers	86	81.30	0.499	0.528	-4.70
Chicago White Sox	79	80.17	0.495	0.488	1.17
Cleveland Indians	65	72.56	0.448	0.401	7.56
Kansas City Royals	65	65.75	0.406	0.401	0.75
Anaheim Angels	97	92.13	0.569	0.599	-4.87
Texas Rangers	87	85.35	0.527	0.537	-1.65
Seattle Mariners	85	75.13	0.464	0.525	-9.87
Oakland Athletics	75	80.80	0.499	0.463	5.80

Table 6. Pythagorean Results for the 2009 American League.

Team	Won	Prd.Won	Prd.Win%	Win%	Diff.
Philadelphia Phillies	93	91.89	0.567	0.574	-1.11
Florida Marlins	87	81.59	0.504	0.537	-5.41
Atlanta Braves	86	91.25	0.563	0.531	5.25
New York Mets	70	71.95	0.444	0.432	1.95
Washington Nationals	59	65.54	0.405	0.364	6.54
St. Louis Cardinals	91	90.86	0.561	0.562	-0.14
Chicago Cubs	83	84.30	0.524	0.516	1.30
Milwaukee Brewers	80	77.90	0.481	0.494	-2.10
Cincinnati Reds	78	75.61	0.467	0.481	-2.39
Houston Astros	74	67.55	0.417	0.457	-6.45
Pittsburgh Pirates	62	66.52	0.413	0.385	4.52
Los Angeles Dodgers	95	99.09	0.612	0.586	4.09
Colorado Rockies	92	89.80	0.554	0.568	-2.20
San Francisco Giants	88	86.46	0.534	0.543	-1.54
San Diego Padres	75	67.07	0.414	0.463	-7.93
Arizona Diamondbacks	70	74.79	0.462	0.432	4.79

Table 7. Pythagorean Results for the 2009 National League.

To obtain better statistical inferences on teams performing above or below their expectations, we can use the prediction intervals from linear regression. Such inferences come with a measure of statistical reliability. For example, with some fixed level of confidence, e.g. say 95% confidence, we infer that a team's Pythagorean expectation falls somewhere within the bounds of its interval. Thus, if a team's observed winning percentage or observed games won exceeds (falls below) the upper bound (lower bound) of their respective intervals, then we are 95% confident that they are performing above (below) their Pythagorean expectation. Based on these prediction intervals, among the 2009 American League teams, it is seen that only the Seattle Mariners (85 wins; 0.525 win percentage) outperformed their expectations by exceeding the upper bound of their respective prediction intervals. The Toronto Blue Jays (75 wins; 0.463 win percentage) under-performed their expectations by falling below the lower bound of their respective prediction intervals, but only by a little. Among the 2009 National League teams, no team exceeded the bounds of their respective prediction intervals. In the 2009 American League, the largest upper estimates belong to the New York Yankees, while the smallest lower estimates belong to the Kansas City Royals. In the 2009 National League, the largest upper estimates belong to the Los Angeles Dodgers, while the smallest lower estimates belong to the Washington Nationals.

Team	Win%	PI for Win%	Won	PI for Won
New York Yankees	0.636	(0.538, 0.639)	103	(87.12, 103.58)
Boston Red Sox	0.586	(0.526, 0.628)	95	(85.22 101.80)
Tampa Bay Rays	0.519	(0.476, 0.581)	84	(77.18, 94.12)
Toronto Blue Jays	0.463	(0.463, 0.568)	75	(75.03, 92.01)
Baltimore Orioles	0.395	(0.372, 0.474)	64	(60.26, 76.86)
Minnesota Twins	0.534	(0.478, 0.582)	87	(77.39, 94.32)
Detroit Tigers	0.528	(0.446, 0.551)	86	(72.26, 89.25)
Chicago White Sox	0.488	(0.442, 0.547)	79	(71.64, 88.63)
Cleveland Indians	0.401	(0.396, 0.500)	65	(64.19, 81.00)
Kansas City Royals	0.401	(0.356, 0.457)	65	(57.64, 74.04)
Anaheim Angels	0.599	(0.516, 0.619)	97	(83.65, 100.33)
Texas Rangers	0.537	(0.474, 0.579)	87	(76.79, 93.74)
Seattle Mariners	0.525	(0.412, 0.516)	85	(66.68, 83.59)
Oakland Athletics	0.463	(0.446, 0.551)	75	(72.26, 89.25)

Table 8. 95% Prediction Intervals for the 2009 American League.

Team	Win%	PI for Win%	Won	PI for Won
Philadelphia Phillies	0.574	(0.515, 0.618)	93	(83.40, 100.10)
Florida Marlins	0.537	(0.451, 0.556)	87	(73.04, 90.03)
Atlanta Braves	0.531	(0.511, 0.614)	86	(82.75, 99.48)
New York Mets	0.432	(0.393, 0.496)	70	(63.60, 80.38)
Washington Nationals	0.364	(0.355, 0.456)	59	(57.44, 73.82)
St. Louis Cardinals	0.562	(0.508, 0.612)	91	(82.35, 99.10)
Chicago Cubs	0.516	(0.471, 0.575)	83	(76.26, 93.22)
Milwaukee Brewers	0.494	(0.428, 0.533)	80	(69.40, 86.37)
Cincinnati Reds	0.481	(0.415, 0.519)	78	(67.15, 84.07)
Houston Astros	0.457	(0.366, 0.468)	74	(59.36, 75.89)
Pittsburgh Pirates	0.385	(0.363, 0.465)	62	(58.77, 75.26)
Los Angeles Dodgers	0.586	(0.561, 0.660)	95	(90.80, 106.97)
Colorado Rockies	0.568	(0.502, 0.605)	92	(81.28, 98.08)
San Francisco Giants	0.543	(0.481, 0.585)	88	(77.90, 94.82)
San Diego Padres	0.463	(0.364, 0.465)	75	(58.90, 75.40)
Arizona Diamondbacks	0.432	(0.410, 0.514)	70	(66.35, 83.25)

Table 9. 95% Prediction Intervals for the 2009 National League.

3 Pythagorean Expectation and the Parametric Bootstrap

3.1 The Weibull Model and Maximum Likelihood Estimation

In a recent paper by Miller (2006), a baseball team's expected winning percentage is derived, under the assumptions that runs scored per game and runs allowed per game follow independent shifted Weibull distributions with different scale parameters, but sharing a common shape parameter and location parameter. Recall that the shifted Weibull distribution, with shape parameter γ, scale parameter α and location parameter θ, has a distribution function of the form

$$F(x \,|\, \gamma, \alpha, \theta) = 1 - \exp\left[-\left(\frac{x-\theta}{\alpha}\right)^{\gamma}\right] \cdot I(\theta \leq x < \infty) \tag{33}$$

with density function of the form

$$f(x \,|\, \gamma, \alpha, \theta) = \frac{\gamma}{\alpha^{\gamma}}(x-\theta)^{\gamma-1} \exp\left[-\frac{(x-\theta)^{\gamma}}{\alpha^{\gamma}}\right] \cdot I(\theta \leq x < \infty) \tag{34}$$

where $\gamma, \alpha, > 0$ and $\theta \in \mathbb{R}$. Here and throughout, we denote this by writing Weibull(γ, α, θ). Miller shows that if X and Y are independent random variables respectively following Weibull$(\gamma, \alpha_{RS}, \theta)$ and Weibull$(\gamma, \alpha_{RA}, \theta)$ distributions, then a team's expected winning percentage is

$$\mathbb{P}(X > Y) = \frac{(RS-\theta)^{\gamma}}{(RS-\theta)^{\gamma} + (RA-\theta)^{\gamma}} = \frac{\alpha_{RS}^{\gamma}}{\alpha_{RS}^{\gamma} + \alpha_{RA}^{\gamma}} \tag{35}$$

where $RS = \mathbb{E}(X) = \alpha_{RS}\,\Gamma(1+\gamma^{-1}) + \theta$ and $RA = \mathbb{E}(Y) = \alpha_{RA}\,\Gamma(1+\gamma^{-1}) + \theta$ are the expected runs scored per game and expected runs allowed per game, respectively. Miller takes $\theta = -1/2$. We shall make the same assumption here.

From the distributional assumptions on runs scored per game and runs allowed per game, we establish a statistical model appropriate for parametric bootstrap simulation. Let the random variables $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ denote independent random samples which are respectively drawn from Weibull$(\gamma, \alpha_{RS}, \theta = -1/2)$ and Weibull$(\gamma, \alpha_{RA}, \theta = -1/2)$ distributions, where γ, α_{RS} and α_{RA} are unknown parameters to be estimated on the basis of the runs scored per game and runs allowed per game. We use the method of maximum likelihood estimation (cf. with Casella and Berger (2002)).

For the Weibull model, the likelihood function has the form

$$L(\gamma, \alpha_{RS}, \alpha_{RA} \,|\, \mathbf{x}, \mathbf{y}) = \prod_{j=1}^{n} f_X(x_j \,|\, \gamma, \alpha_{RS}, \theta) \cdot f_Y(y_j \,|\, \gamma, \alpha_{RA}, \theta)$$
$$\frac{\gamma^{2n} \left(\prod_{j=1}^{n} (x_j - \theta)(y_j - \theta) \right)^{\gamma - 1}}{(\alpha_{RS} \cdot \alpha_{RA})^{n\gamma}} \cdot \exp\left[- \frac{\sum_{j=1}^{n} (x_j - \theta)^{\gamma}}{\alpha_{RS}^{\gamma}} - \frac{\sum_{j=1}^{n} (y_j - \theta)^{\gamma}}{\alpha_{RA}^{\gamma}} \right]. \tag{36}$$

Finding the maximum likelihood estimator (MLE) of the shape parameter γ, that is $\widehat{\gamma}$, requires extensive iterative numerical calculations, and can be obtained by solving the equation

$$\gamma^{-1} = \frac{\sum_{j=1}^{n} (x_j - \theta)^{\gamma} \log(x_j - \theta)}{2 \sum_{j=1}^{n} (x_j - \theta)^{\gamma}} + \frac{\sum_{j=1}^{n} (y_j - \theta)^{\gamma} \log(y_j - \theta)}{2 \sum_{j=1}^{n} (y_j - \theta)^{\gamma}}$$
$$- \frac{1}{2n} \sum_{j=1}^{n} [\log(x_j - \theta) + \log(y_j - \theta)]. \tag{37}$$

The MLE's of the scale parameters α_{RS} and α_{RA} are respectively given by the power means

$$\widehat{\alpha}_{RS} = \left(\frac{1}{n} \sum_{j=1}^{n} (x_j - \theta)^{\widehat{\gamma}} \right)^{1/\widehat{\gamma}} \tag{38}$$

$$\widehat{\alpha}_{RA} = \left(\frac{1}{n} \sum_{j=1}^{n} (y_j - \theta)^{\widehat{\gamma}} \right)^{1/\widehat{\gamma}}. \tag{39}$$

The expected winning percentage $\psi(\gamma, \alpha_{RS}, \alpha_{RA}) = \mathbb{P}(X > Y)$ is a functional parameter, and can be estimated by the plug-in principle, i.e.

$$\widehat{\psi} = \psi(\widehat{\gamma}, \widehat{\alpha}_{RS}, \widehat{\alpha}_{RA}) = \frac{\sum_{j=1}^{n} (x_j - \theta)^{\widehat{\gamma}}}{\sum_{j=1}^{n} (x_j - \theta)^{\widehat{\gamma}} + \sum_{j=1}^{n} (y_j - \theta)^{\widehat{\gamma}}}. \tag{40}$$

Using maximum likelihood estimation on the Weibull model, over the 30 teams from the 2009 MLB regular season, the mean of $\widehat{\gamma}$ over the 30 teams is 1.69 with a standard deviation of 0.08. Over the 30 teams, the mean absolute difference between observed and estimated games won is 5.28 games with a standard deviation of 3.38 games. The root mean square difference between observed and estimated games won is 6.24 games.

Team	Won	Est.Won	Est.Win%	Win%	Diff.	$\widehat{\gamma}$
New York Yankees	103	89.53	0.553	0.636	-13.47	1.70
Boston Red Sox	95	89.61	0.553	0.586	-5.39	1.63
Tampa Bay Rays	84	86.49	0.534	0.519	2.49	1.78
Toronto Blue Jays	75	84.59	0.522	0.463	9.59	1.77
Baltimore Orioles	64	71.75	0.443	0.395	7.75	1.75
Minnesota Twins	87	85.21	0.523	0.534	-1.79	1.71
Detroit Tigers	86	82.54	0.506	0.528	-3.46	1.71
Chicago White Sox	79	80.19	0.495	0.488	1.19	1.58
Cleveland Indians	65	76.31	0.471	0.401	11.31	1.67
Kansas City Royals	65	68.16	0.421	0.401	3.16	1.68
Anaheim Angels	97	88.04	0.543	0.599	-8.96	1.67
Texas Rangers	87	84.22	0.520	0.537	-2.78	1.61
Seattle Mariners	85	75.06	0.463	0.525	-9.94	1.74
Oakland Athletics	75	80.83	0.499	0.463	5.83	1.69

Table 10. Maximum Likelihood Results for the 2009 American League.

Team	Won	Est.Won	Est.Win%	Win%	Diff.	$\widehat{\gamma}$
Philadelphia Phillies	93	91.28	0.563	0.574	-1.72	1.74
Florida Marlins	87	81.46	0.503	0.537	-5.54	1.91
Atlanta Braves	86	89.66	0.553	0.531	3.66	1.65
New York Mets	70	72.72	0.449	0.432	2.72	1.70
Washington Nationals	59	69.59	0.430	0.364	10.59	1.85
St. Louis Cardinals	91	87.75	0.542	0.562	-3.25	1.61
Chicago Cubs	83	84.27	0.523	0.516	1.27	1.64
Milwaukee Brewers	80	77.77	0.480	0.494	-2.23	1.78
Cincinnati Reds	78	74.53	0.460	0.481	-3.47	1.67
Houston Astros	74	70.26	0.434	0.457	-3.74	1.62
Pittsburgh Pirates	62	71.91	0.447	0.385	9.91	1.61
Los Angeles Dodgers	95	97.51	0.602	0.586	2.51	1.73
Colorado Rockies	92	87.02	0.537	0.568	-4.98	1.69
San Francisco Giants	88	84.05	0.519	0.543	-3.95	1.53
San Diego Padres	75	69.37	0.428	0.463	-5.63	1.74
Arizona Diamondbacks	70	76.17	0.470	0.432	6.17	1.66

Table 11. Maximum Likelihood Results for the 2009 National League.

3.2 Parametric Bootstrap Simulation and Bootstrap Confidence Intervals

One approach to computing useful confidence intervals for the expected winning percentage and games won is to use bootstrap simulation methods. The bootstrap is a modern, computer-intensive, general purpose approach to statistical inference. The advantage of bootstrapping over any analytical method is its simplicity. As long as one has the data, it is relatively straightforward to apply the bootstrap to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of a distribution, such as percentile points, proportions, odds ratio, and correlation coefficients. A standard reference on bootstrap methods is Davison and Hinkley (1998).

A parametric bootstrap simulation would draw independent random samples

$$\begin{aligned} X_1^*, X_2^*, \ldots, X_n^* &\sim \text{Weibull}(\widehat{\gamma}, \widehat{\alpha}_{RS}, \theta = -1/2) \\ Y_1^*, Y_2^*, \ldots, Y_n^* &\sim \text{Weibull}(\widehat{\gamma}, \widehat{\alpha}_{RA}, \theta = -1/2) \end{aligned}$$

where $\widehat{\alpha}_{RS}$ and $\widehat{\alpha}_{RA}$ and $\widehat{\gamma}$ are the MLEs. These are random samples simulated from independent Weibull$(\widehat{\gamma}, \widehat{\alpha}_{RS}, \theta = -1/2)$ and Weibull$(\widehat{\gamma}, \widehat{\alpha}_{RA}, \theta = -1/2)$ distributions. These are the so-called "plug-in distributions" or "fitted parametric models" (cf. with Davison and Hinkley (1998) and Casella and Berger (2002)). We want a large number, say B, of such independent samples simulated from the fitted parametric models:

$$\begin{gathered} (X_1^*, X_2^*, \ldots, X_n^*)^{(1)} \text{ and } (Y_1^*, Y_2^*, \ldots, Y_n^*)^{(1)} \\ (X_1^*, X_2^*, \ldots, X_n^*)^{(2)} \text{ and } (Y_1^*, Y_2^*, \ldots, Y_n^*)^{(2)} \\ \vdots \\ (X_1^*, X_2^*, \ldots, X_n^*)^{(B)} \text{ and } (Y_1^*, Y_2^*, \ldots, Y_n^*)^{(B)}. \end{gathered}$$

We will use the formula

$$\widehat{\psi} = \frac{\sum_{j=1}^n (x_j - \theta)^{\widehat{\gamma}}}{\sum_{j=1}^n (x_j - \theta)^{\widehat{\gamma}} + \sum_{j=1}^n (x_j - \theta)^{\widehat{\gamma}}} = t(\mathbf{x}, \mathbf{y}) \tag{41}$$

to compute an estimate of a team's expected winning percentage based on the original data set, and each of the B independent samples, i.e.

$$\begin{gathered} (X_1, X_2, \ldots, X_n) \text{ and } (Y_1, Y_2, \ldots, Y_n) \mapsto t \\ (X_1^*, X_2^*, \ldots, X_n^*)^{(1)} \text{ and } (Y_1^*, Y_2^*, \ldots, Y_n^*)^{(1)} \mapsto t_1^* \\ (X_1^*, X_2^*, \ldots, X_n^*)^{(2)} \text{ and } (Y_1^*, Y_2^*, \ldots, Y_n^*)^{(2)} \mapsto t_2^* \\ \vdots \\ (X_1^*, X_2^*, \ldots, X_n^*)^{(B)} \text{ and } (Y_1^*, Y_2^*, \ldots, Y_n^*)^{(B)} \mapsto t_B^* \end{gathered}$$

so that we have t and $t_1^*, t_2^*, \ldots, t_B^*$.

By the strong law of large numbers, with probability 1, in the limit as $B \to \infty$,

$$\bar{T^*} = \frac{1}{B}\sum_{i=1}^{B} T_i^* \to t. \tag{42}$$

In other words, if B is sufficiently large, we have

$$\bar{t^*} = \left(\frac{1}{B}\sum_{i=1}^{B} t_i^*\right) \approx t. \tag{43}$$

Respectively, estimates for the bias and variance of T are

$$\text{Bias}(T) \approx \bar{t^*} - t = \left(\frac{1}{B}\sum_{i=1}^{B} t_i^*\right) - t, \tag{44}$$

$$\text{Var}(T) \approx \text{Var}_B^*(t) = \frac{1}{B-1}\sum_{i=1}^{B}(t_i^* - \bar{t^*})^2. \tag{45}$$

For a large enough B, an approximate 95% confidence interval for winning percentage is

$$t - \text{Bias}(t) \pm 1.96\sqrt{\text{Var}(t)} \approx (2t - \bar{t^*}) \pm 1.96\sqrt{\text{Var}_B^*(t)}. \tag{46}$$

and an approximate 95% confidence interval for games won is

$$nt - \text{Bias}(nt) \pm 1.96\sqrt{\text{Var}(nt)} \approx (2nt - n\bar{t^*}) \pm 1.96\sqrt{\text{Var}_B^*(nt)}, \tag{47}$$

When the number of games played is not large enough, the distribution of T may not follow an approximate Normal distribution, so the approximate confidence intervals may not be reliable. In this case, we can use an equal-tailed 95% confidence interval for winning percentage which is

$$\left(2t - t^*_{((B+1)(0.975))}, 2t - t^*_{((B+1)(0.025))}\right) \tag{48}$$

and an equal-tailed 95% confidence interval for games won is

$$\left(2nt - nt^*_{((B+1)(0.975))}, 2nt - nt^*_{((B+1)(0.025))}\right). \tag{49}$$

The accuracy of the estimates for bias, variance, and quantiles depends on the value of B. To be safe, B will need to be at least 1000, but really good results usually require $B \geq 5000$.

3.3 Numerical Results from the Parametric Bootstrap Framework

Using parametric bootstrap simulation, we computed both approximate and equal-tailed 95% confidence intervals for the Pythagorean expectation for all 30 MLB teams on the basis of their runs scored per game and runs allowed per game data from the 2009 regular season. The numerical results are based on $B = 5000$ bootstrap samples.

Like we did with the prediction intervals based on linear regression, we can also use these bootstrap confidence intervals to obtain better statistical inferences on teams performing above or below their expectations. Again, such inferences come with a measure of statistical reliability. For example, with 95% confidence, we may infer that a team's Pythagorean expectation falls somewhere within the bounds of its confidence interval. Thus, if a team's observed winning percentage or observed games won exceeds (falls below) the upper bound (lower bound) of their respective confidence intervals, then we are 95% confident that they are performing above (below) their Pythagorean expectation.

Team	Win%	CI for Win%	Won	CI for Won
New York Yankees	0.636	(0.499, 0.607)	103	(80.85, 98.27)
Boston Red Sox	0.586	(0.499, 0.608)	95	(80.85, 98.45)
Tampa Bay Rays	0.519	(0.478, 0.588)	84	(77.05, 95.25)
Toronto Blue Jays	0.463	(0.470, 0.577)	75	(76.09, 93.51)
Baltimore Orioles	0.395	(0.390, 0.496)	64	(63.17, 80.38)
Minnesota Twins	0.534	(0.469, 0.576)	87	(76.40, 93.86)
Detroit Tigers	0.528	(0.453, 0.560)	86	(73.78, 91.25)
Chicago White Sox	0.488	(0.440, 0.550)	79	(71.29, 89.10)
Cleveland Indians	0.401	(0.416, 0.525)	65	(67.41, 85.04)
Kansas City Royals	0.401	(0.367, 0.474)	65	(59.53, 76.75)
Anaheim Angels	0.599	(0.491, 0.597)	97	(79.48, 96.70)
Texas Rangers	0.537	(0.466, 0.573)	87	(75.47, 92.84)
Seattle Mariners	0.525	(0.409, 0.518)	85	(66.26, 83.84)
Oakland Athletics	0.463	(0.444, 0.554)	75	(71.90, 89.69)

Table 12. Approximate 95% Confidence Intervals for the 2009 American League.

Team	Win%	CI for Win%	Won	CI for Won
Philadelphia Phillies	0.574	(0.509, 0.618)	93	(82.47, 100.04)
Florida Marlins	0.537	(0.448, 0.557)	87	(72.53, 90.29)
Atlanta Braves	0.531	(0.499, 0.608)	86	(80.90, 98.39)
New York Mets	0.432	(0.396, 0.502)	70	(64.08, 81.32)
Washington Nationals	0.364	(0.376, 0.483)	59	(60.91, 78.31)
St. Louis Cardinals	0.562	(0.489, 0.596)	91	(79.23, 96.50)
Chicago Cubs	0.516	(0.469, 0.578)	83	(75.45, 93.04)
Milwaukee Brewers	0.494	(0.425, 0.534)	80	(68.90, 86.54)
Cincinnati Reds	0.481	(0.406, 0.514)	78	(65.81, 83.31)
Houston Astros	0.457	(0.381, 0.487)	74	(61.73, 78.90)
Pittsburgh Pirates	0.385	(0.392, 0.500)	62	(63.12, 80.58)
Los Angeles Dodgers	0.586	(0.550, 0.654)	95	(89.04, 106.01)
Colorado Rockies	0.568	(0.483, 0.592)	92	(78.20, 95.88)
San Francisco Giants	0.543	(0.466, 0.572)	88	(75.42, 92.67)
San Diego Padres	0.463	(0.375, 0.482)	75	(60.69, 78.07)
Arizona Diamondbacks	0.432	(0.416, 0.525)	70	(67.33, 85.06)

Table 13. Approximate 95% Confidence Intervals for 2009 National League.

Team	Win%	CI for Win%	Won	CI for Won
New York Yankees	0.636	(0.499, 0.607)	103	(80.83, 98.30)
Boston Red Sox	0.586	(0.501, 0.609)	95	(81.13, 98.60)
Tampa Bay Rays	0.519	(0.480, 0.588)	84	(77.76, 95.22)
Toronto Blue Jays	0.463	(0.471, 0.578)	75	(76.24, 93.57)
Baltimore Orioles	0.395	(0.390, 0.497)	64	(63.13, 80.56)
Minnesota Twins	0.534	(0.470, 0.575)	87	(76.59, 93.75)
Detroit Tigers	0.528	(0.453, 0.560)	86	(73.90, 91.24)
Chicago White Sox	0.488	(0.441, 0.550)	79	(71.45, 89.07)
Cleveland Indians	0.401	(0.416, 0.526)	65	(67.38, 85.16)
Kansas City Royals	0.401	(0.367, 0.474)	65	(59.48, 76.84)
Anaheim Angels	0.599	(0.490, 0.598)	97	(79.41, 96.90)
Texas Rangers	0.537	(0.467, 0.576)	87	(75.57, 93.24)
Seattle Mariners	0.525	(0.409, 0.516)	85	(66.27, 83.57)
Oakland Athletics	0.463	(0.444, 0.553)	75	(71.92, 89.51)

Table 14. Equal-tailed 95% Confidence Intervals for the 2009 American League.

Team	Win%	CI for Win%	Won	CI for Won
Philadelphia Phillies	0.574	(0.509, 0.618)	93	(82.48, 100.06)
Florida Marlins	0.537	(0.448, 0.557)	87	(72.51, 90.26)
Atlanta Braves	0.531	(0.500, 0.608)	86	(81.03, 98.47)
New York Mets	0.432	(0.395, 0.501)	70	(63.99, 81.22)
Washington Nationals	0.364	(0.374, 0.481)	59	(60.65, 77.97)
St. Louis Cardinals	0.562	(0.488, 0.595)	91	(79.12, 96.45)
Chicago Cubs	0.516	(0.469, 0.578)	83	(75.58, 93.13)
Milwaukee Brewers	0.494	(0.426, 0.535)	80	(68.96, 86.60)
Cincinnati Reds	0.481	(0.407, 0.513)	78	(65.89, 83.04)
Houston Astros	0.457	(0.380, 0.487)	74	(61.54, 78.90)
Pittsburgh Pirates	0.385	(0.390, 0.500)	62	(62.72, 80.44)
Los Angeles Dodgers	0.586	(0.550, 0.656)	95	(89.13, 106.26)
Colorado Rockies	0.568	(0.484, 0.593)	92	(78.38, 96.02)
San Francisco Giants	0.543	(0.465, 0.572)	88	(75.40, 92.61)
San Diego Padres	0.463	(0.374, 0.482)	75	(60.52, 78.02)
Arizona Diamondbacks	0.432	(0.414, 0.523)	70	(67.11, 84.80)

Table 15. Equal-tailed 95% Confidence Intervals for 2009 National League.

Based on these bootstrap confidence intervals, among the 2009 American League teams, it is seen that the New York Yankees (103 wins; 0.636 win percentage), the Anaheim Angels (97 wins; 0.599 win percentage), and the Seattle Mariners (85 wins; 0.525 win percentage) outperformed their expectations by exceeding the upper bound of their respective confidence intervals. It is also seen that the Toronto Blue Jays (75 wins; 0.463 win percentage) and Cleveland Indians (65 wins; 0.401 win percentage) under-performed their expectations by falling below the lower bound of their respective confidence intervals. Among the 2009 National League teams, no team outperformed their expectations by exceeding the upper bound of their respective confidence interval. It is also seen that the Washington Nationals (59 wins; 0.364 win percentage), Pittsburgh Pirates (62 wins; 0.385 win percentage) under-performed their expectations by falling below the lower bound of their respective confidence intervals. In the 2009 American League, the largest upper estimates belong to the Boston Red Sox, while the smallest lower estimates belong to the Kansas City Royals. In the 2009 National League, the largest upper estimates belong to the Los Angeles Dodgers, while the smallest lower estimates belong to the San Diego Padres.

4 Conclusions

We have seen that interval estimates for the Pythagorean expectation are useful in determining, with some measure of statistical reliability, whether a

team is playing above or below expectations. Based on the results obtained for the 2009 MLB regular season, the bootstrap confidence intervals, from the Weibull model, appear to be better at inferring or detecting which teams are performing above or below expectations than the prediction intervals obtained in the regression framework. This may be due to the fact that the Weibull model produces conservative point estimates compared to the Pythagorean formula. On the scale of winning percentage, the length of the prediction intervals are only slightly shorter than the length of the bootstrap intervals. As for future research, it would be of interest to study confidence interval estimation in the framework of other Pythagorean-type methods.

References

Braunstein, A. (2010). Consistency and Pythagoras, *Journal of Quantitative Analysis in Sports* **6, No. 1**.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, second edn, Duxbury.

Cochran, J. J. (2008). The optimal value and potential alternatives of Bill James' pythagorean method of baseball, *STAtOR* **2**.

Davenport, C. and Woolner, K. (1999). Revisiting the pythagorean theorem, *http://www.baseballprospectus.com/article.php?articleid=342* .

Davison, A. C. and Hinkley, D. V. (1998). *Bootstrap Methods and their Application*, Cambridge University Press.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, second edn, Chapman and Hall.

Faraway, J. J. (2005). *Linear Models with R*, first edn, Chapman and Hall.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, first edn, Chapman and Hall.

James, B. (1983). *The Bill James Baseball Abstract 1983*, Ballantine.

Keri, J. (2007). *Baseball Between the Numbers: Why Everything You Know about the Game Is Wrong*, Perseus Publishing.

Kutner, M. H., N. C. J. and Neter, J. (2004). *Applied Linear Regression Models*, fourth edn, McGraw-Hill.

Miller, S. J. (2006). A derivation of the pythagorean won-loss formula in baseball, *By the Numbers* **16, No. 1**: 1–41.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Resnick, S. I. (2001). *A Probability Path*, Birkhäuser.

Vollmayr-Lee, B. (2002). More than you probably ever wanted to know about the 'pythagorean' method, *http://www.eg.bucknell.edu/bvollmay/baseball/pythagoras.html* .

E-mail Address: `david.deming.tung@gmail.com`