RELIEVING AND READJUSTING PYTHAGORAS

VICTOR LUO AND STEVEN J. MILLER

ABSTRACT. Bill James invented the Pythagorean expectation in the late 70's to predict a baseball team's winning percentage knowing just their runs scored and allowed. His original formula estimates a winning percentage of $\mathrm{RS}^2/(\mathrm{RS}^2+\mathrm{RA}^2)$, where RS stands for runs scored and RA for runs allowed; later versions found better agreement with data by replacing the exponent 2 with numbers near 1.83. Miller and his colleagues provided a theoretical justification by modeling runs scored and allowed by independent Weibull distributions. They showed that a single Weibull distribution did a very good job of describing runs scored and allowed, and led to a predicted won-loss percentage of $(\mathrm{RS}_{\mathrm{obs}}-1/2)^\gamma/((\mathrm{RS}_{\mathrm{obs}}-1/2)^\gamma+(\mathrm{RA}_{\mathrm{obs}}-1/2)^\gamma)$, where $\mathrm{RS}_{\mathrm{obs}}$ and $\mathrm{RA}_{\mathrm{obs}}$ are the observed runs scored and allowed and γ is the shape parameter of the Weibull (typically close to 1.8). We show a linear combination of Weibulls more accurately determines a team's run production and increases the prediction accuracy of a team's winning percentage by an average of about 25% (thus while the currently used variants of the original predictor are accurate to about four games a season, the new combination is accurate to about three). The new formula is more involved computationally; however, it can be easily computed on a laptop in a matter of minutes from publicly available season data. It performs as well (or slightly better) than the related Pythagorean formulas in use, and has the additional advantage of having a theoretical justification for its parameter values (and not just an optimization of parameters to minimize prediction error).

CONTENTS

1. INTRODUCTION

The Pythagorean Win/Loss Formula, also known as the Pythagorean formula or Pythagorean expectation, was invented by Bill James in the late 1970s to use a team's observed runs scored and

1

allowed to predict their winning percentage. Originally given by

$$\text{Won} - \text{Loss Percentage} \; = \; \frac{\text{RS}^2}{\text{RS}^2 + \text{RA}^2}, \tag{1.1}$$

with RS the runs scored and RA the runs allowed, it earned its name from the similarity of the denominator to the sums of squares in the Pythagorean formula from geometry.[1] Later versions found better agreement by replacing the exponent 2 with numbers near 1.83, leading to an average error of about three to four games per season.

The formula is remarkably simple, requiring only the runs scored and allowed by a team in a season, and the calculation (even with the improved exponent) is easily done on any calculator or phone. It is one of the most commonly listed expanded statistics on websites. One reason for its prominence is its accuracy in predicting a team's future performance through a simple calculation and not through computationally intense simulations. Additionally, it allows sabermetricians and fans to assess a manager's impact on a team, and estimate the value of new signings by seeing how their presence would change the predictions.

Because of its widespread use and utility, it is very desirable to have improvements. In his senior thesis, the first named author, supervised by the second named author, explored various attempted improvements to the Pythagorean formula. These included replacing the observed runs scored and allowed each game with adjusted numbers, with the adjustments coming from a variety of sources (such as ballpark effects, game state[2], WHIP, ERA+, and WAR of the pitcher, ...).

As these led to only minor improvements[3] (see [Luo] for a detailed analysis of these and other adjustments), we turned our attention to the successful theoretical model used by Miller and his colleagues [DaMil1, DaMil2, Mil, MCGLP], where it was assumed runs scored and allowed were independently drawn from Weibull distributions with the same shape parameter. Recall the three parameter Weibull density is given by

$$f(x; \alpha, \beta, \gamma) \; = \; \begin{cases} \frac{\gamma}{\alpha} \left((x - \beta)/\alpha \right)^{\gamma - 1} e^{-((x - \beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

The effect of α is to control the spread of the output, while β translates the distribution. The most important parameter is γ, which controls the shape. See Figure 1 for some plots.

Their success is due to the fact that the three parameter Weibull is a very flexible family of distributions, capable of fitting many one hump distributions, including to a statistically significant degree the observed runs scored and allowed data. Miller chose to use Weibulls for two reasons. First, they lead to double integrals for the probabilities that can be evaluated in closed form. This is extremely important if we desire a simple expression such as the one posited by James (see [HJM] for alternative simple formulas). Second, in addition to being flexible, special values of

---

[1]Though of course the more natural shape in baseball is the diamond, save for some interesting stadium features, such as the triangle in Fenway Park.

[2]For example, if a team is up by a large amount late in a game, they frequently use weaker relief pitchers and rest some starters, while the trailing team makes similar moves; thus the offensive productions from this point onward may not be indicative of the team's true abilities and a case can be made to ignore such data.

[3]The only adjusted formula that was at least on par or very near the accuracy of the original Pythagorean W/L Formula was that of ballpark factor.
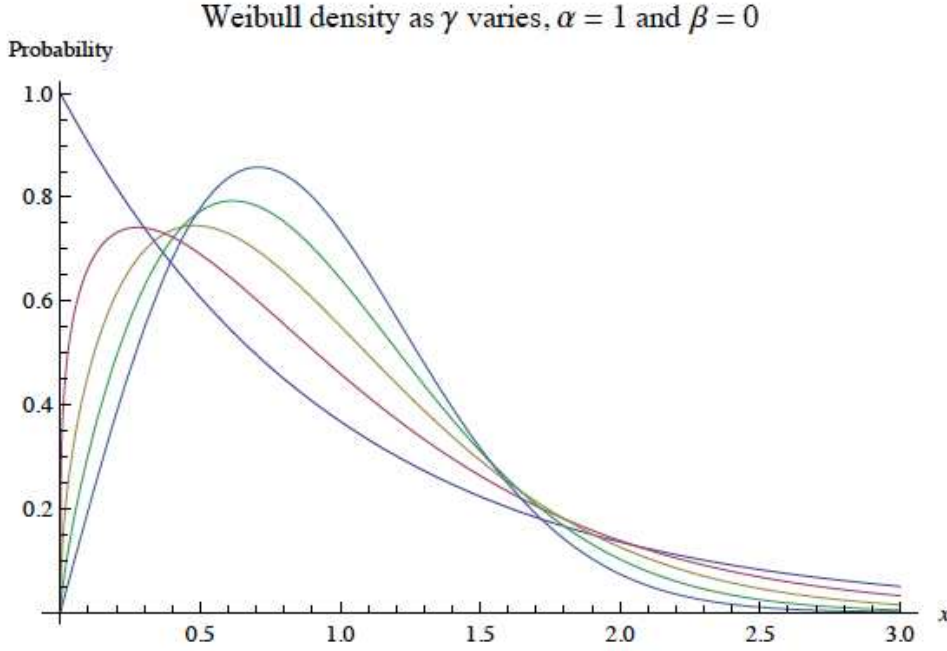
FIGURE 1. The varying distributions of the Weibull family with $\alpha = 1$ and $\beta = 0$.

the Weibulls correspond to well-known distributions ($\gamma = 1$ is an exponential, while $\gamma = 2$ is the Rayleigh distribution).

The goal of this paper is to show that one can significantly improve the predictive power if instead of modeling runs scored and allowed as being drawn from independent Weibulls, we instead model them as being drawn from linear combinations of independent Weibulls. The advantage of this approach is that we are still able to obtain tractable double integrals which can be done in closed form. There is a cost, however, as now more analysis is needed to find the parameters and the correct linear combinations. While this results in a more complicated formula than the standard variant of James' formula, it is well worth the cost as on average it is better by one game per season (thus a typical error is 3 games per team per year, as opposed to 4 which is the typical result in the current formula). Comparing it to `baseball-reference.com`'s calculated Expected WL from 1979 to 2013, we find that the linear combination of Weibulls is approximately .06 of a game better, a difference is not statistically significant. However, there are noticeable trends that appear in certain eras.

2. THEORETICAL CALCULATIONS

2.1. **Preliminaries.** It is important to note that we assume that runs scored and allowed are taken from continuous, not discrete, distributions. This allows us to deal with continuous integrals rather than discrete sums, which most of the time leads to easier calculations. While a discrete distribution would probably more effectively map runs in baseball, the assumption of drawing runs from a continuous distribution allows for more manageable calculations, and is a very sensible estimate of those runs observed. It also will lead to closed form expressions, which are much easier to work with and allow us to avoid having to resort to simulations.

The Weibulls lead to significantly easier calculations because if we have a random variable X chosen from a Weibull distribution with parameters α, β, and γ, then $X^{1/\gamma}$ is exponentially distributed with parameter α^γ; thus, a change of variables yields a simpler integral of exponentials, which can be done in closed form (see Appendix 9.1 in [MCGLP] for details).

In all arguments below we always take $\beta = -1/2$, though we often write β to keep the discussion more general for applications to other sports. The reason we do this is that we use procedures such as the Method of Least Squares to find the best fit parameters, and this requires binning the observed runs scored and allowed data. As baseball scores are discrete, there are issues if these values occur at the boundary of bins; it is much better if they are at the center. By taking $\beta = -1/2$ we break the data into bins

$$\left[-\frac{1}{2}, \frac{1}{2}\right), \quad \left[\frac{1}{2}, \frac{3}{2}\right), \quad \left[\frac{3}{2}, \frac{5}{2}\right), \quad \cdots . \tag{2.1}$$

Our final assumption is that runs scored and runs allowed are independent. This obviously cannot be true, as a baseball game never ends in a tie. For example, if the Orioles and the Red Sox are playing and the O's score 5 runs, then the Sox cannot score 5. Statistical analyses support this hypothesis; see the independence tests with structural zeros in [Mil] or Appendix 9.2 in [MCGLP] for details.

We end this subsection with the mean and the variance of the Weibull. The calculation follows from standard integration (see the expanded version of [Mil] or Appendix 9.1 in [MCGLP] for a proof of the formula for the mean; the derivation of the variance follows in a similar fashion).

Lemma 2.1. *Consider a Weibull with parameters α, β, γ. The mean, $\mu_{\alpha,\beta,\gamma}$, equals*

$$\mu_{\alpha,\beta,\gamma} = \alpha\Gamma(1 + \gamma^{-1}) + \beta \tag{2.2}$$

while the variance, $\sigma^2_{\alpha,\beta,\gamma}$, is

$$\sigma^2_{\alpha,\beta,\gamma} = \alpha^2\Gamma(1 + 2\gamma^{-1}) - \alpha^2\Gamma(1 + \gamma^{-1})^2, \tag{2.3}$$

where $\Gamma(x)$ is the Gamma function, defined by $\Gamma(x) = \int_0^\infty e^{-u}u^{x-1}du$.

2.2. **Linear Combination of Weibulls.** We now state and prove our main result for a linear combination of two Weibulls, and leave the straightforward generalization to combinations of more Weibulls to the reader. The reason such an expansion is advantageous and natural is that, following [Mil], we can integrate pairs of Weibulls in the regions needed and obtain simple closed form expressions. The theorem below also holds if $\gamma < 0$; however, in that situation the more your runs scored exceeds your runs allowed, the worse your predicted record due to the different shape of the Weibull (in all applications of Weibulls in survival analysis, the shape parameter γ must be positive).

Theorem 2.2. *Let the runs scored and allowed per game be two independent random variables drawn from linear combinations of independent Weibull distributions with the same β's and γ's. Specifically, if $W(t; \alpha, \beta, \gamma)$ represents a Weibull distribution with parameters (α, β, γ), and we choose non-negative weights[4] $0 \leq c_i, c'_j \leq 1$ (so $c_1 + c_2 = 1$ and $c'_1 + c'_2 = 1$), then the density of runs scored, X is*

$$f(x; \alpha_{\mathrm{RS}_1}, \alpha_{\mathrm{RS}_2}, \beta, \gamma, c_1, c_2) = c_1 W(x; \alpha_{\mathrm{RS}_1}, \beta, \gamma) + c_2 W(\alpha_{\mathrm{RS}_2}, \beta, \gamma) \tag{2.4}$$

---

[4]If we had more terms in the linear combination, we would simply choose non-negative weights summing to 1.

and runs allowed, Y, is

$$f(y; \alpha_{\mathrm{RA}_1}, \alpha_{\mathrm{RA}_2}, \beta, \gamma, c_1', c_2') = c_1' W(y; \alpha_{\mathrm{RA}_1}, \beta, \gamma) + c_2' W(\alpha_{\mathrm{RA}_2}, \beta, \gamma). \qquad (2.5)$$

In addition, we choose α_{RS_1} and α_{RS_2} so that the mean of X is $\mathrm{RS}_{\mathrm{obs}}$ and choose α_{RA_1} and α_{RA_2} such that the mean of Y is $\mathrm{RA}_{\mathrm{obs}}$. For $\gamma > 0$, we have

$$\mathrm{Won} - \mathrm{Loss\ Percentage}(\alpha_{\mathrm{RS}_1}, \alpha_{\mathrm{RS}_2}, \alpha_{\mathrm{RA}_1}, \alpha_{\mathrm{RA}_2}, \beta, \gamma, c_1, c_2, c_1', c_2')$$

$$= c_1 c_1' \frac{\alpha_{\mathrm{RS}_1}^{\gamma}}{\alpha_{\mathrm{RS}_1}^{\gamma} + \alpha_{\mathrm{RA}_1}^{\gamma}} + c_1 c_2' \frac{\alpha_{\mathrm{RS}_1}^{\gamma}}{\alpha_{\mathrm{RS}_1}^{\gamma} + \alpha_{\mathrm{RA}_2}^{\gamma}}$$

$$+ c_2 c_1' \frac{\alpha_{\mathrm{RS}_2}^{\gamma}}{\alpha_{\mathrm{RS}_2}^{\gamma} + \alpha_{\mathrm{RA}_1}^{\gamma}} + c_2 c_2' \frac{\alpha_{\mathrm{RS}_2}^{\gamma}}{\alpha_{\mathrm{RS}_2}^{\gamma} + \alpha_{\mathrm{RA}_2}^{\gamma}}$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \frac{\alpha_{\mathrm{RS}_i}^{\gamma}}{\alpha_{\mathrm{RS}_i}^{\gamma} + \alpha_{\mathrm{RA}_j}^{\gamma}}. \qquad (2.6)$$

Proof. As the means of X (runs scored) and Y (runs allowed) are $\mathrm{RS}_{\mathrm{obs}}$ and $\mathrm{RA}_{\mathrm{obs}}$, respectively, and the random variables are drawn from linear combinations of independent Weibulls, by Lemma 2.1

$$\mathrm{RS}_{\mathrm{obs}} = c_1(\alpha_{\mathrm{RS}_1}\Gamma(1 + \gamma^{-1}) + \beta) + (1 - c_1)(\alpha_{\mathrm{RS}_2}\Gamma(1 + \gamma^{-1}) + \beta)$$
$$\mathrm{RA}_{\mathrm{obs}} = c_1'(\alpha_{\mathrm{RA}_1}\Gamma(1 + \gamma^{-1}) + \beta) + (1 - c_1')(\alpha_{\mathrm{RA}_2}\Gamma(1 + \gamma^{-1}) + \beta). \qquad (2.7)$$

We now calculate the probability that X exceeds Y. We constantly use the fact that the integral of a probability density is 1. We need the two β and the two γ's to be equal in order to obtain closed form expressions.[5] We find

$$\mathrm{Prob}(X > Y) = \int_{x=\beta}^{\infty} \int_{y=\beta}^{x} f(x; \alpha_{\mathrm{RS}_1}, \alpha_{\mathrm{RS}_2}, \beta, \gamma, c_1, c_2) f(y; \alpha_{\mathrm{RA}_1}, \alpha_{\mathrm{RA}_2}, \beta, \gamma, c_1', c_2') dy dx$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} \int_{x=0}^{\infty} \int_{y=0}^{x} c_i c_j' \frac{\gamma}{\alpha_{\mathrm{RS}_i}} \left(\frac{x}{\alpha_{\mathrm{RS}_i}}\right)^{\gamma-1} e^{-(\frac{x}{\alpha_{\mathrm{RS}_i}})^{\gamma}} \frac{\gamma}{\alpha_{\mathrm{RA}_j}} \left(\frac{x}{\alpha_{\mathrm{RA}_j}}\right)^{\gamma-1} e^{-(\frac{x}{\alpha_{\mathrm{RA}_j}})^{\gamma}} dy dx$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\mathrm{RS}_i}} \left(\frac{x}{\alpha_{\mathrm{RS}_i}}\right)^{\gamma-1} e^{-(\frac{x}{\alpha_{\mathrm{RS}_i}})^{\gamma}} \left[\int_{y=0}^{x} \frac{\gamma}{\alpha_{\mathrm{RA}_j}} \left(\frac{y}{\alpha_{\mathrm{RA}_j}}\right)^{\gamma-1} e^{-(\frac{y}{\alpha_{\mathrm{RA}_j}})^{\gamma}} dy\right] dx$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\mathrm{RS}_i}} \left(\frac{x}{\alpha_{\mathrm{RS}_i}}\right)^{\gamma-1} e^{-(\frac{x}{\alpha_{\mathrm{RS}_i}})^{\gamma}} * \left[1 - e^{-(\frac{x}{\alpha_{\mathrm{RA}_j}})^{\gamma}}\right] dx$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \left[1 - \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\mathrm{RS}_i}} \left(\frac{x}{\alpha_{\mathrm{RS}_i}}\right)^{\gamma-1} e^{-(\frac{x}{\alpha_{\mathrm{RS}_i}})^{\gamma} - (\frac{x}{\alpha_{\mathrm{RA}_j}})^{\gamma}} dx\right]. \qquad (2.8)$$

---

[5]If the β's are differen then in the integration below we might have issues with the bounds of integration, while if the γ's are unequal we get incomplete Gamma functions, though for certain rational ratios of the γ's these can be done in closed form.

We set

$$\frac{1}{\alpha_{ij}^{\gamma}} = \frac{1}{\alpha_{\mathrm{RS}_i}^{\gamma}} + \frac{1}{\alpha_{\mathrm{RA}_j}^{\gamma}} = \frac{\alpha_{\mathrm{RS}_i}^{\gamma} + \alpha_{\mathrm{RA}_j}^{\gamma}}{\alpha_{\mathrm{RS}_i}^{\gamma} \alpha_{\mathrm{RA}_j}^{\gamma}}$$

for $1 \le i, j \le 2$, so that (2.8) becomes

$$\sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \left[1 - \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\mathrm{RS}_i}} \left(\frac{x}{\alpha_{\mathrm{RS}_i}} \right)^{\gamma-1} e^{-\left(\frac{x}{\alpha_{ij}} \right)^{\gamma}} dx \right]$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \left[1 - \frac{\alpha_{ij}^{\gamma}}{\alpha_{\mathrm{RS}_i}^{\gamma}} \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{ij}} \left(\frac{x}{\alpha_{ij}} \right)^{\gamma-1} e^{-\left(\frac{x}{\alpha_{ij}} \right)^{\gamma}} dx \right]$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \left[1 - \frac{\alpha_{ij}^{\gamma}}{\alpha_{\mathrm{RS}_i}^{\gamma}} \right]$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \left[1 - \frac{1}{\alpha_{\mathrm{RS}_i}^{\gamma}} * \frac{\alpha_{\mathrm{RS}_i}^{\gamma} \alpha_{\mathrm{RA}_j}^{\gamma}}{\alpha_{\mathrm{RS}_i}^{\gamma} + \alpha_{\mathrm{RA}_j}^{\gamma}} \right]$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} c_i c_j' \left[\frac{\alpha_{\mathrm{RS}_i}^{\gamma}}{\alpha_{\mathrm{RS}_i}^{\gamma} + \alpha_{\mathrm{RA}_j}^{\gamma}} \right]$$

$$= c_1 c_1' \frac{\alpha_{\mathrm{RS}_1}^{\gamma}}{\alpha_{\mathrm{RS}_1}^{\gamma} + \alpha_{\mathrm{RA}_1}^{\gamma}} + c_1 c_2' \frac{\alpha_{\mathrm{RS}_1}^{\gamma}}{\alpha_{\mathrm{RS}_1}^{\gamma} + \alpha_{\mathrm{RA}_2}^{\gamma}}$$

$$+ c_2 c_1' \frac{\alpha_{\mathrm{RS}_2}^{\gamma}}{\alpha_{\mathrm{RS}_2}^{\gamma} + \alpha_{\mathrm{RA}_1}^{\gamma}} + c_2 c_2' \frac{\alpha_{\mathrm{RS}_2}^{\gamma}}{\alpha_{\mathrm{RS}_2}^{\gamma} + \alpha_{\mathrm{RA}_2}^{\gamma}}, \tag{2.9}$$

completing the proof of Theorem 2.2. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

3. CURVE FITTING

3.1. **Theory.** We now turn to finding the values of the parameters leading to the best fit. We require $\beta = -1/2$ (for binning purposes), but otherwise the parameters (α_{RS_1}, α_{RS_2}, α_{RA_1}, α_{RA_2}, γ, c_1, c_2, c_1', c_2') are free.[6] Our first approach was to use the Method of Moments, where we compute the number of moments equal to the number of parameters. Unfortunately the resulting equations were too involved to permit simple solutions for them in terms of the observed data; for completeness they are given in Appendix A (or see [Luo]). We thus turned to the Method of Least Squares (though one could also do an analysis through the Method of Maximum Likelihood).

We looked at the 30 teams of the entire league from the 2004 to 2012 season. We display results from the 2011, but the results from any other season are similar readily available (see [Luo]). We implemented the Method of Least Squares using the bins in (2.1), which involved minimizing the sum of squares of the error of the runs scored data plus the sum of squares of the error of the runs allowed data. There were seven free parameters: α_{RS_1}, α_{RS_2}, α_{RA_1}, α_{RA_2}, γ, c_1, and c_1'. Letting $\mathrm{Bin}(k)$ be the kth bin of (2.1), $\mathrm{RS}_{\mathrm{obs}}(k)$ and $\mathrm{RA}_{\mathrm{obs}}(k)$ represent the observed number of games

---

[6]Subject to, of course, $0 \le c_i, c_j' \le 1$ and $c_1 + c_1 = c_1' + c_2' = 1$.

with number of runs scored and allowed in $\mathrm{Bin}(k)$, and $A(\alpha_1, \alpha_2, \beta, \gamma, c_1, k)$ denote the area under the linear combination of two Weibulls with parameters $(\alpha_1, \alpha_2, \beta, \gamma, c_1)$ in $\mathrm{Bin}(k)$, then for each team we found the values of $(\alpha_{\mathrm{RS}_1}, \alpha_{\mathrm{RS}_2}, \alpha_{\mathrm{RA}_1}, \alpha_{\mathrm{RA}_2}, \gamma, c_1, c_1')$ that minimized

$$\sum_{k=1}^{\text{Num. Bins}} (\mathrm{RS}_{\mathrm{obs}}(k) - \#\text{Games} * A(\alpha_{\mathrm{RS}_1}, \alpha_{\mathrm{RS}_2}, -.5, \gamma, c_1, k))^2$$

$$+ \sum_{k=1}^{\text{Num. Bins}} (\mathrm{RA}_{\mathrm{obs}}(k) - \#\text{Games} * A(\alpha_{\mathrm{RA}_1}, \alpha_{\mathrm{RA}_2}, -.5, \gamma, c_1', k))^2. \qquad (3.1)$$

3.2. **Results.** For each team, we found the best fit linear combination of Weibulls. In Figure 2, we compared the predicted wins, losses, and won-loss percentage with the observed ones.

Team	Obs Wins	Pred Wins	ObsPerc	PredPerc	GamesDiff	γ
Boston Red Sox	90	89.6	0.556	0.553	0.38	1.65
New York Yankees	97	100.3	0.599	0.619	-3.26	2.07
Baltimore Orioles	69	69.1	0.426	0.426	0.	2.02
Tampa Bay Devil Rays	91	87.8	0.562	0.542	3.20	1.65
Toronto Blue Jays	81	79.7	0.500	0.492	1.31	1.99
Minnesota Twins	63	63.5	0.389	0.392	-0.48	1.82
Chicago White Sox	79	77.6	0.488	0.479	1.42	1.72
Cleveland Indians	80	77.4	0.494	0.478	2.57	1.76
Detroit Tigers	95	89.2	0.586	0.551	5.80	1.79
Kansas City Royals	71	75.3	0.438	0.465	-4.33	2.05
Los Angeles Angels	86	84.9	0.531	0.524	1.14	1.66
Oakland Athletics	74	78.4	0.457	0.484	-4.41	1.63
Texas Rangers	96	95.6	0.593	0.590	0.42	1.64
Seattle Mariners	67	69.8	0.414	0.431	-2.78	1.66
Atlanta Braves	89	85.3	0.549	0.526	3.73	1.74
Philadelphia Phillies	102	98.2	0.630	0.606	3.77	1.66
Florida Marlins	72	74.5	0.444	0.460	-2.52	2.02
New York Mets	77	78.8	0.475	0.486	-1.81	1.82
Washington Nationals	80	79.5	0.497	0.494	0.45	1.75
St. Louis Cardinals	90	88.5	0.556	0.547	1.45	1.84
Houston Astros	56	65.0	0.346	0.401	-8.96	1.82
Chicago Cubs	71	70.7	0.438	0.436	0.34	2.00
Cincinnati Reds	79	82.9	0.488	0.512	-3.93	2.25
Pittsburgh Pirates	72	72.3	0.444	0.446	-0.29	1.80
Milwaukee Brewers	96	90.8	0.593	0.560	5.23	1.80
Los Angeles Dodgers	82	81.6	0.509	0.507	0.41	1.62
San Francisco Giants	86	80.0	0.531	0.494	6.03	1.71
San Diego Padres	71	77.1	0.438	0.476	-6.06	1.73
Colorado Rockies	73	75.8	0.451	0.468	-2.83	1.85
Arizona Diamondbacks	94	86.8	0.580	0.536	7.17	2.23

FIGURE 2. Results for the 2011 season using Method of Least Squares.

The code used is available in [Luo]. Using the Method of Least Squares, the mean γ over all 30 teams is 1.83 with a standard deviation of 0.18 (the median is 1.79). We can see that the exponent 1.83, considered as the best exponent, is clearly within the region of one standard deviation from the mean γ. Considering the absolute value of the difference between observed and predicted wins, we have a mean of 2.89 with a standard deviation of 2.34 (median is 2.68). Without considering the absolute value, the mean is 0.104 with a standard deviation of 3.75 (and a median of 0.39). We only concern ourselves with the absolute value of the difference, as this really tells how accurate our predicted values are.

These values are significant improvements on those obtained when using a single Weibull distribution to predict runs (which essentially reproduces James' original formula, though with a slightly different exponent), which produces a mean number of games off of 4.43 with standard deviation

3.23 (and median 3.54) in the absolute value case. We display the results over seasons from 2004 to 2012 in Figure 3. It is apparent that the linear combination of Weibulls better estimates teams' win/loss percentage; in fact, it is over one game better at estimating than the single Weibull! The mean number of games off for a single Weibull from 2004 to 2012 was 4.22 (with a standard deviation of 3.03), while that of the linear combination of Weibulls was 3.11 (with a standard deviation of 2.33). In addition, there is less standard deviation in the estimates. Thus, it appears that the linear combination of Weibulls provides a much tighter, better estimate than the single Weibull does.
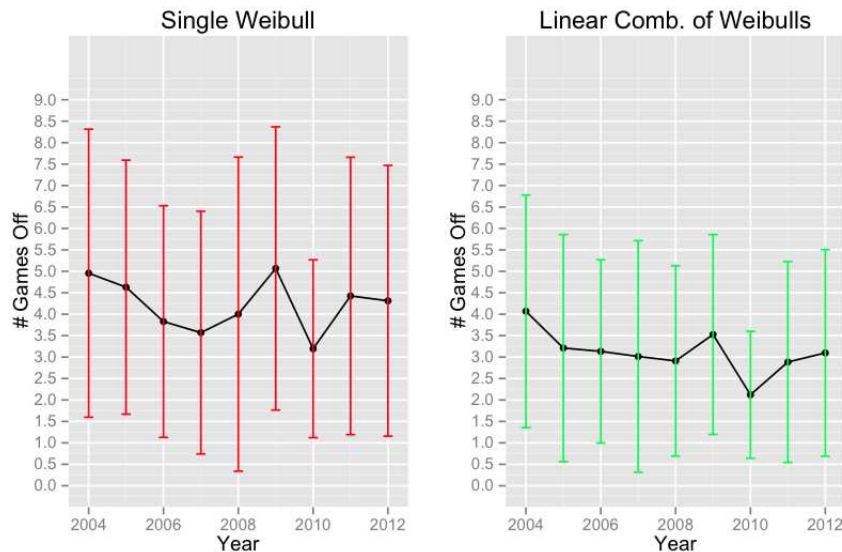

FIGURE 3. Mean number of games off (with standard deviation) for single Weibull and linear combination of Weibulls from 2004-2012.

To further demonstrate how accurate the quality of the fit is, we compare the best fit linear combination of Weibulls of runs scored and allowed with those observed of the 2011 Seattle Mariners in Figure 4; we can see that the fit is visually very good. Of course, the fit *cannot* be worse, as we can always set $c_1 = 0 = c'_1$; however, we can see the linear combination of Weibulls does a better job tracking the shape of the runs scored.

We then performed an independent two-sample t-test with unequal variances in R using the t.test command to see if the difference between the games off determined by the single Weibull and those by linear combinations of Weibulls is statistically significant in Figure 5. With a p-value less than 0.01 and a 95% confidence interval that does not contain 0, we can see that the difference is in fact statistically significant.

In addition, we compared the mean number of games off of `baseball-reference.com`'s Pythagorean Win-Loss statistic (pythWL) and those of the linear combination of Weibulls from 1979 to 2013. Originally, we used ESPN's ExWL statistic[7] which used an exponent of 2; however, ESPN only went down to the year 2002, and it has been shown that using the exponent 1.83 is more accurate than using ESPN's exponent of 2. Using `baseball-reference.com`'s Pythagorean

---

[7]See the bottom of the page `http://espn.go.com/mlb/stats/rpi/_/year/2011`.
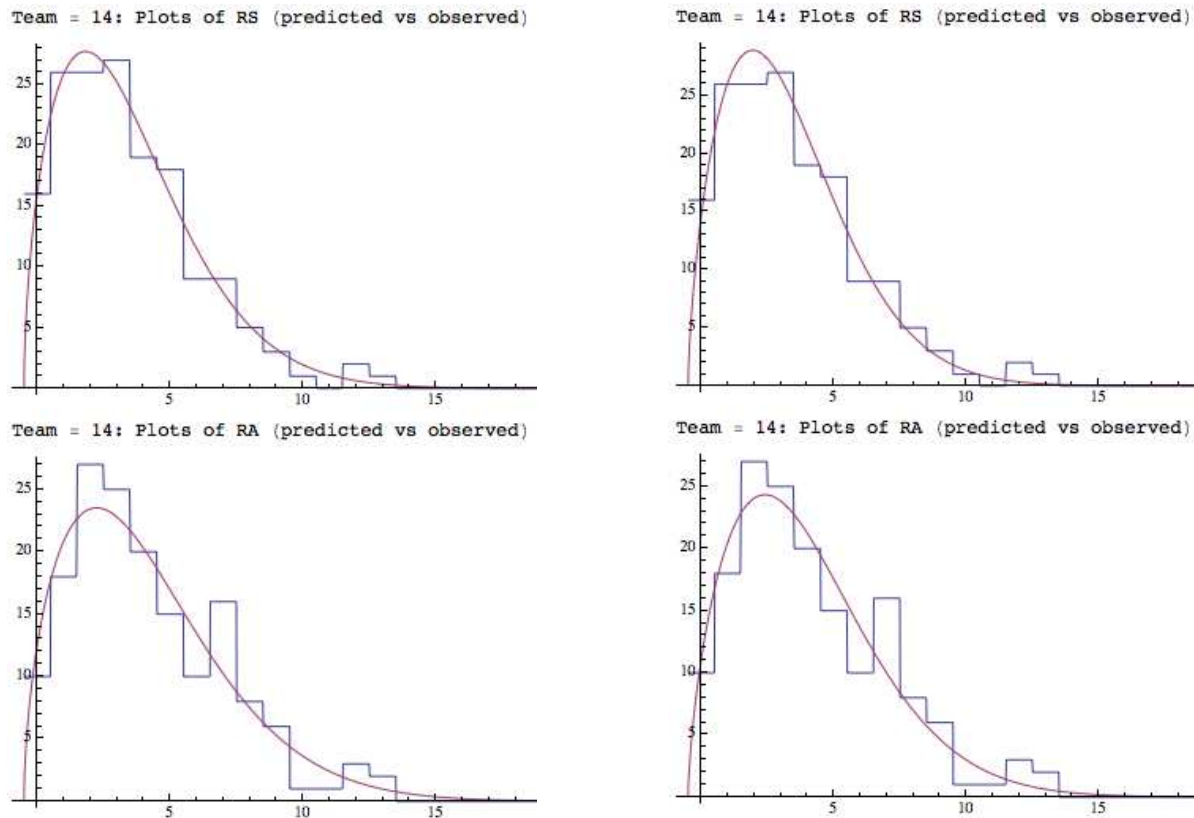
(A) Single Weibull mapping runs scored and allowed.

(B) Linear Combination of Weibulls mapping runs scored and allowed.

FIGURE 4. Comparison of best fit linear combination of Weibulls versus single Weibull for runs scored (top) and allowed (bottom) for the 2011 Seattle Mariners against the observed distribution of scores.

```
> t.test(twoweight,oneweibull)

    Welch Two Sample t-test

data:  twoweight and oneweibull
t = -4.0899, df = 15.483, p-value = 0.0009091
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.691890 -0.534674
sample estimates:
mean of x mean of y
 3.107333  4.220616
```

FIGURE 5. t-test to determine whether the difference between the games off determined by the single Weibull and those by linear combinations of Weibulls is statistically significant.

Win-Loss statistic (pythWL) to obtain more data (`baseball-reference.com` allowed us to go all the way down to 1979, rather than just 2002, which ESPN gives), the pythWL statistic[8] is calculated as

$$(\text{runs scored}^{1.83})/(\text{runs scored}^{1.83} + \text{runs allowed}^{1.83}).$$

We display the results of our comparisons in Figure 6. The mean number of games off for the pythWL statistic was 3.09 with a standard deviation of 2.26, numbers only slightly worse than those of the linear combination of Weibulls (mean of 3.03 with standard deviation of 2.21). So, we can see that the linear combination of Weibulls is doing, on average, about .06 of a game better than the pythWL statistic. We performed an independent two-sample t-test with unequal variances in R using the t.test command to see if the difference between the games off determined by the pythWL statistic and the linear combinations of Weibulls is statistically significant in Figure 7. With a very large p-value, we fail to reject the null hypothesis, suggesting that the difference in mean number of games off is not in fact statistically significant. We also display a plot (Figure 8) that models the difference in the number of games between the pythWL statistic and the linear combination of Weibulls; it seems to be a constant positive value for the most part, suggesting that the linear combination of Weibulls is doing slightly better than the pythWL statistic.
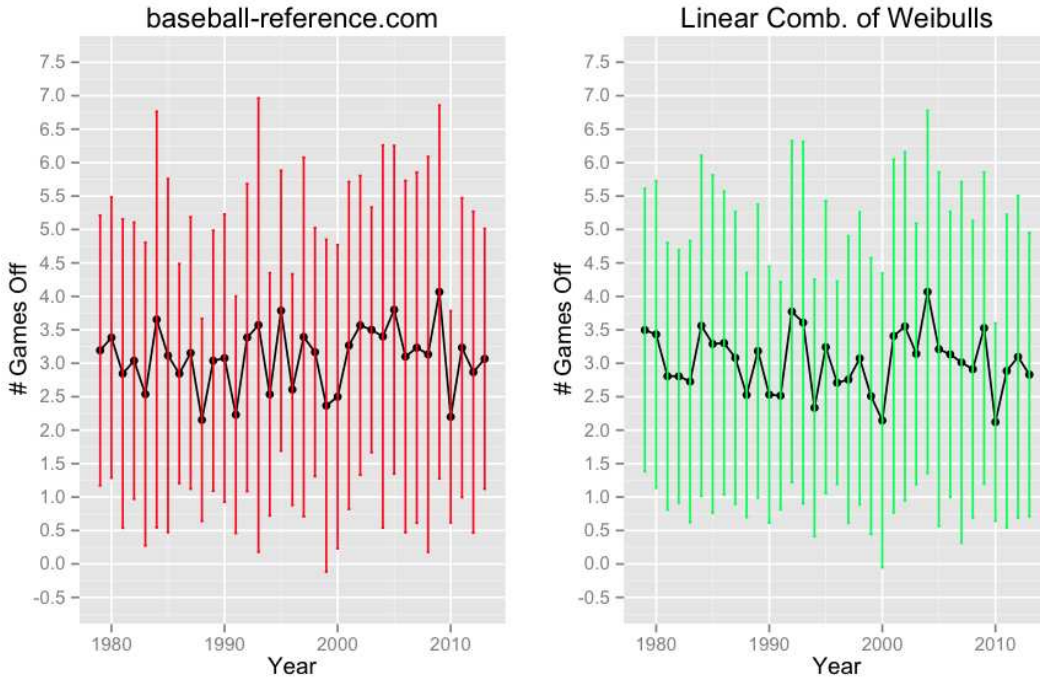

FIGURE 6. Mean number of games off (with standard deviation) for `baseball-reference.com`'s pythWL statistic and linear combination of Weibulls from 1979-2013.

Looking at Figure 8 more closely, we can see that there are parts/eras of the graph in which the pythWL statistic does better, and parts where the linear combination of Weibulls does better. In

[8]At `http://www.sports-reference.com/blog/baseball-reference-faqs/` see the section "What is Pythagorean Winning Percentage?".

the era from 1979-1989, the pythWL statistic is more accurate, beating the linear combination of Weibulls in 7 out of the 11 years. However, from 1990 to 2013, the linear combination of Weibulls wins in 15 out of the 24 years, and does so by around 0.3 games in those years. Furthermore, when the pythWL statistic does beat the linear combination of Weibulls in the years from 1990 to 2013, it does so by around 0.25 games, including the point at 2004, which seems very out of the ordinary; without this point, the pythWL statistic wins by about .2 games in the years between 1990 and 2013 that it does beat the linear combination of Weibulls. Thus, in more recent years, it may make more sense to use the linear combination of Weibulls. In addition, with respect to the standard deviation of number of games off of the pythWL statistic (2.26) and the linear combination of Weibulls (2.21), we can see that the linear combination of Weibulls provides on average a tighter fit, i.e., there is less fluctuation in the mean number of games off for each team in each year (from 1990 to 2013, the pythWL statistic standard deviation in games off is 2.34 while that of the linear combination of Weibulls is 2.22, so we again see that the linear combination does noticeably better in recent years).

It is important to note that the pythWL statistic just takes the functional form of the Pythagorean Win/Loss Formula with an exponent (γ) of 1.83, while we give theoretical justification for our formula.

```
> t.test(baseballref,twoweight)

    Welch Two Sample t-test

data:  baseballref and twoweight
t = 0.4386, df = 67.92, p-value = 0.6623
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1734837  0.2712292
sample estimates:
mean of x mean of y
 3.086094  3.037221
```

FIGURE 7. t-test to determine whether the difference between the games off determined by ESPN ExWL and those by linear combinations of Weibulls is statistically significant.

We also performed χ^2 tests to determine the goodness of fit to see how well the linear combination of Weibulls maps the observed data, and whether runs scored and allowed are independent. We used the bins as in (2.1) and test statistic

$$\sum_{k=1}^{\# \text{Bins}} \frac{(\text{RS}_{\text{obs}}(k) - \#\text{Games} * A(\alpha_{\text{RS}_1}, \alpha_{\text{RS}_2}, -.5, \gamma, c_1, k))^2}{\#\text{Games} * A(\alpha_{\text{RS}_1}, \alpha_{\text{RS}_2}, -.5, \gamma, c_1, k)}$$

$$+ \sum_{k=1}^{\# \text{Bins}} \frac{(\text{RA}_{\text{obs}}(k) - \#\text{Games} * A(\alpha_{\text{RA}_1}, \alpha_{\text{RA}_2}, -.5, \gamma, c'_1, k))^2}{\#\text{Games} * A(\alpha_{\text{RA}_1}, \alpha_{RS_A}, -.5, \gamma, c'_1, k)} \tag{3.2}$$

for the goodness of fit tests, with $2 * (\#\text{Bins} - 1) - 1 - 7 = 16$ degrees of freedom, the factor of 7 coming from estimating 7 parameters, namely α_{RS_1}, α_{RS_2}, α_{RA_1}, α_{RA_2}, γ, c_1, and c'_1. We did not estimate β, as we took it to be -.5. Having 16 degrees of freedom gives critical threshold values of 26.3 (at the 95% level) and 32.0 (at the 99% level). However, since there are multiple
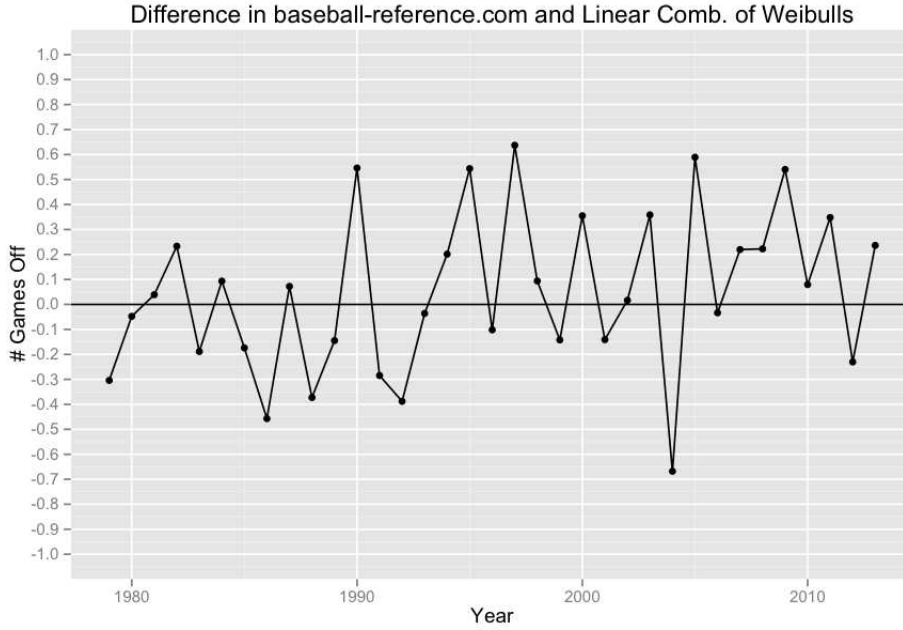
FIGURE 8. Difference in mean number of games off for `baseball-reference.com`'s pythWL statistic and linear combination of Weibulls from 1979-2013.

comparisons being done (namely 30 for the different teams), we use a Bonferroni adjustment and obtain critical thresholds of 37.7 (95%) and 42.5 (99%). From the first column of Figure 9, all the teams fall within the unadjusted 99% threshold, with the exception of the Texas Rangers (just barely!), who easily fall into the Bonferroni adjusted 95% threshold. Therefore, the observed data closely follows a linear combination of Weibulls with the proper estimated parameters.

Since the test for independence of runs scored and allowed requires that the row and column of the contingency table have at least one non-zero entry, the bins used to bin the runs score and allowed were

$$[0, 1) \cup [1, 2) \cup \cdots \cup [9, 10) \cup [11, \infty). \tag{3.3}$$

We use integer endpoints because we are using the observed runs from games. We have a 12 by 12 contingency table with zeroes along the diagonal, since runs scored and allowed can never be equal. This leads to an incomplete 12 by 12 contingency table with $(12-1)^2 - 12 = 109$ degrees of freedom; constructing a test requires the use of structural zeroes. The theory behind tests using structural zeroes can be seen in [Mil] or Appendix 9.2 of [MCGLP]. We observe that 109 degrees of freedom give critical threshold values of 134.37 (at the 95% level) and 146.26 (at the 99% level). Again, since we are doing multiple comparisons, we use a Bonferroni adjustment, obtaining critical thresholds of 157.68 (95%) and 166.45 (99%). From the second column of Figure 9, all the teams fall within the 99% threshold, with the exception of the Los Angeles Angels (just barely!), who easily fall into the Bonferroni adjusted 95% threshold. Thus, runs scored and allowed are acting as though they are statistically independent.

A more in depth discussion of the justification behind the tests can be found in [Mil].

Team	RS+RA $\chi 2$: 16 d.f.	Indep $\chi 2$: 109 d.f
Boston Red Sox	32.19	110.43
New York Yankees	22.40	94.77
Baltimore Orioles	20.64	113.00
Tampa Bay Devil Rays	26.12	130.73
Toronto Blue Jays	21.36	144.77
Minnesota Twins	19.90	115.55
Chicago White Sox	24.53	123.83
Cleveland Indians	19.72	127.71
Detroit Tigers	20.85	117.01
Kansas City Royals	13.15	130.11
Los Angeles Angels	13.43	147.98
Oakland Athletics	21.87	96.17
Texas Rangers	34.85	128.06
Seattle Mariners	11.08	114.28
Atlanta Braves	18.69	116.30
Philadelphia Phillies	26.05	112.60
Florida Marlins	30.56	126.76
New York Mets	17.43	127.31
Washington Nationals	25.93	113.19
St. Louis Cardinals	10.71	108.63
Houston Astros	12.03	131.24
Chicago Cubs	22.38	121.58
Cincinnati Reds	12.59	125.93
Pittsburgh Pirates	15.30	93.10
Milwaukee Brewers	26.30	101.20
Los Angeles Dodgers	14.88	112.87
San Francisco Giants	21.90	119.33
San Diego Padres	22.50	115.07
Colorado Rockies	22.36	109.59
Arizona Diamondbacks	13.96	114.54

FIGURE 9. χ^2 test results of the 2011 season from least squares of goodness of fit and independence of runs score and allowed.

4. FUTURE WORK AND CONCLUSIONS

While a one game improvement in prediction is very promising, as our formula requires us to fit the runs scored and allowed distributions we explored simplifications. We tried to simplify the formula, even giving up some accuracy, in order to devise a formula that could be easily implemented using just a team's runs scored and allowed (and the variance of each of these) in order to determine the team's winning percentage. Unfortunately, the weight parameters c_1 and c_1' plays too much of a factor; in 2011, the mean of the parameter c_1 is 0.21 with a standard deviation of 0.39 (and a median of 0.21). With such large fluctuations in the weight parameters from team to team, the task of finding a simpler formula was almost impossible, as creating a uniform formula that every team could use was not feasible when two of the key parameters were so volatile. Taking this into account, we tried fixing the γ, c_1, and c_1' parameters, allowing for us to just solve a quartic involving the first and second moments to find the other parameters (α_{RS_1}, α_{RS_2}, α_{RA_1}, and α_{RA_2}). However, while we were able to solve for the other parameters, plugging these values of the parameters gave us a significantly worse prediction of teams' win-loss percentage compared to linear combination of Weibulls and baseball-reference.com's Pythagorean Win-Loss statistic. One of the great attractions of James' Pythagorean formula is its ease of use; we hope to return to other simplifications and approximations in a later paper. Our hope is to find a linearization or approximation of our main result, similar to how Dayaratna and Miller [DaMil1] showed the linear predictor of Jones and Tappin [JT] follows from a linearization of the Pythagorean formula.

To summarize our results, using a linear combination of Weibulls rather than a single Weibull increases the prediction accuracy of a team's W/L percentage. More specifically, we saw that the single Weibull's predictions for a team's wins were on average 4.22 games off (with a standard deviation of 3.03), while the linear combination of Weibull's predictions for a team's wins were from 2004-2012 were on average 3.11 games off (with a standard deviation of 2.33), producing about a 25% increase in prediction accuracy. We also performed χ^2 goodness of fit tests for the linear combination of Weibulls and tested the statistical independence of runs scored and allowed (a necessary assumption), and see that in fact the linear combination of Weibulls with properly estimated parameters obtained from least squares analysis closely maps the observed runs and that runs scored and allowed are in fact statistically independent. In addition, when compared against `baseball-reference.com`'s Pythagorean Win-Loss statistic, the linear combination of Weibulls does .06 of a game better in the years from 1979 to 2013, but this improvement cannot be considered statistically significant. However, in more recent years, it is worth noting that it does appear that the linear combination of Weibulls is doing better than `baseball-reference.com`'s Pythagorean Win-Loss statistic.

APPENDIX A. MOMENTS OF WEIBULLS

For the runs scored data, we have $c_2 = 1 - c_1$ and the density equals

$$c_1 \frac{\gamma}{\alpha_{\text{RS}_1}} \left(\frac{x - \beta}{\alpha_{\text{RS}_1}} \right)^{\gamma - 1} e^{-(\frac{x-\beta}{\alpha_{\text{RS}_1}})^\gamma} + (1 - c_1) \frac{\gamma}{\alpha_{\text{RS}_2}} \left(\frac{x - \beta}{\alpha_{\text{RS}_2}} \right)^{\gamma - 1} e^{-(\frac{x-\beta}{\alpha_{\text{RS}_2}})^\gamma}. \tag{A.1}$$

From [Mur], and using the fact that the two Weibulls in the linear combination are independent, we obtained the following moments:

$$\text{First Moment} = c_1(\alpha_{\text{RS}_1}\Gamma(1 + \gamma^{-1}) + \beta) + (1 - c_1)(\alpha_{\text{RS}_2}\Gamma(1 + \gamma^{-1}) + \beta)$$

$$\text{Second Moment} = c_1^2 \left[\alpha_{\text{RS}_1}^2 \left(\Gamma(\frac{2}{\gamma} + 1) - (\Gamma(\frac{1}{\gamma} + 1))^2 \right) \right]$$
$$+ (1 - c_1)^2 \left[\alpha_{\text{RS}_2}^2 \left(\Gamma(\frac{2}{\gamma} + 1) - (\Gamma(\frac{1}{\gamma} + 1))^2 \right) \right]$$

$$\text{Third Moment} = (c_1^3 + (1 - c_1)^3) * \frac{g_3 - 3g_1g_2 + 2g_1^3}{(g_2 - g_1^2)^{3/2}}$$

$$\text{Fourth Moment} = (c_1^4 + (1 - c_1)^4) * \frac{g_4 - 4g_1g_3 + 6g_2g_1^2 - 3g_1^4}{(g_2 - g_1^2)^2}$$
$$+ 6 * c_1^2(1 - c_1)^2 * \left[\alpha_{\text{RS}_1}^2\alpha_{\text{RS}_2}^2 \left(g_2 - g_1^2 \right)^2 \right], \tag{A.2}$$

where $g_i = \Gamma(1 + \frac{i}{\gamma})$, and $\Gamma(x)$ is the Gamma function defined by $\Gamma(x) = \int_0^\infty e^{-u}u^{x-1}du$. We next used R to find the observed moments for teams' runs scored from 2007. The code is available in [Luo].

REFERENCES

[DaMil1]   K. Dayaratna and S. J. Miller, *First Order Approximations of the Pythagorean Won-Loss Formula for Predicting MLB Teams Winning Percentages*, By The Numbers – The Newsletter of the SABR Statistical Analysis Committee **22** (2012), no 1, 15–19.

[DaMil2] K. Dayaratna and S. J. Miller, *The Pythagorean Won-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey* (with Kevin Dayaratna), The Hockey Research Journal: A Publication of the Society for International Hockey Research (2012/2013), pages 193–209.

[HJM] C. N. B. Hammond, W. P. Johnson and S. J. Miller, *The James Function*, preprint 2014. http://xxx.tau.ac.il/pdf/1312.7627v2.

[Ja] B. James, *Baseball Abstract 1983*, Ballantine, 238 pages.

[JT] M. A. Jones and L. A. Tappin, *The Pythagorean Theorem of Baseball and Alternative Models*, The UMAP Journal **26** (2005), no. 2, 12 pages.

[Luo] V. Luo, *Relieving and Readjusting Pythagoras*, Senior Thesis (supervised by S. J. Miller), Williams College 2014.
http://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/st/VictorLuo.pdf.

[Mil] S. J. Miller, *A Derivation of the Pythagorean Won-Loss Formula in Baseball*, Chance Magazine (2007), no. 1, 40–48 (an abridged version appeared in The Newsletter of the SABR Statistical Analysis Committee **16** (February 2006), no. 1, 17–22, and an expanded version is available online at http://arxiv.org/pdf/math/0509698.pdf).

[MCGLP] S. J. Miller, T. Corcoran, J. Gossels, V. Luo and J. Porfilio, *Pythagoras at the Bat*, in: Social Networks and the Economics of Sports (edited by Victor Zamaraev), Springer-Verlag, 2014 (to appear).

[Mur] G. Muraleedharan, *Characteristic and Moment Generating Functions of Three Parameter Weibull Distribution – an Independent Approach*, Research Journal of Mathematical and Statistical Sciences **1** (2013), no. 8, 25–27.

E-mail address: victor.d.luo@williams.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267

E-mail address: sjm1@williams.edu, Steven.Miller.MC.96@aya.yale.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267