Spring 2016

Pythagorean Expectation for Baseball

UNLV Economics Department

E. Scott Leavitt[1], Anthony Serrano[2]

Abstract 0

In this paper we look to predict a baseball team's win percentage using readily available statistics. We use Bill James' "Pythagorean expectation" from his Baseball Abstract published in 1982, as modified by Dennis Moy in 2006, to calculate a predicted win percentage. The Pythagorean expectation that a team's win percentage is dependent on the ratio of the runs they have scored in all their games and the total runs scored by both teams in all their games. We found this formula to be very accurate in predicting a team's win percentage, with the mean prediction estimating actual win percentage inside of 0.13%! This predictive

Acknowledgements:
[1] Scott would like to thank his fiancée Colleen Boyle, Dr. Brad Wimmer and the UNLV Economics Club, and the entire UNLV Economics department. He would also like to thank Kevin Costner and Susan Sarandon, but not Tim Robbins, for starring in *Bull Durham.*

[2] Anthony would like to his friends, family, the UNLV Economics department, and utmost his dad for being there for him and supporting everything he does. He loves and thanks you all, and could not have accomplished it without you!

formula can be used by baseball managers to determine how long-term changes in a team’s lineup may affect their long-term success.

Introduction 1.1

The sport of baseball is one of the most heavily analyzed sports in the world. This is in no small part to its discrete nature with each pitch, hit, or error being recorded in numerous statistics. The wealth of data produced by one season of baseball is staggering with each team plays 162 regular season games. These actions are recorded and compiled by an army of statisticians and published for the public to view. Since the advent of computers statistical analysis of baseball has become easier and more popular, with many Major League teams employing full-time statisticians. The 2004 film *Moneyball* is about Sabermetrics, the analysis of baseball-specific statistics.

We will use the Pythagorean expectation to predict a team's win percentage based on team-level offensive and defensive statistics. Our model deconstructs runs scored as the percentage of total at bats a team gets on base in addition to how many bases a team gains for every at bat. It also deconstructs runs allowed by a team as the sum of the amount of runners opposing teams put on base and how many outs the team makes per opportunity to make an out.

Team owners, managers, players, and fans are all concerned with a team's season win percentage as it determines if a team makes the playoffs at the conclusion of the regular season. Separating runs scored into player-level statistics and runs allowed into partial player-level statistics allows us to observe the effects of a change in a team's lineup on predicted win percentage. This is important especially to baseball managers, whose positions require team performance.

We found that using the Pythagorean expectation with our empirical results is accurate at predicting win percentage. On average our model predicted win percentage 0.124% higher than actual data, with a standard deviation of 4.572%. This means that 95% of the time the actual win percentage of a team will be between 9.086% and -8.838% of the prediction.

Our research is expounded upon in the following sections. Included is a glossary for those unfamiliar with baseball statistics, followed by a review of literature relevant to the Pythagorean expectation. We present our predictive model and the data we used in our regression analysis. Relationships between the variables is visualized and reported in table form. We then present our regression and win percentage prediction results in table and narrative form and discuss our findings and their implications.

Glossary 1.2

Batting Average (AVG): The most widely-used statistic for measuring offensive performance. This statistic is calculated by dividing the number of total number of hits by the total amount of at-bats. This statistic can be misleading, as looks at only hits and excludes other data.

On-Base Percentage (OBP): Calculated by dividing the total times a base was reached for any reason by the total number of at bats. This is relevant because a team cannot score if players do not reach a base (a home run counts as a base reached).

Slugging Percentage (SLG): Slugging is calculated by the total number of bases reached divided by the total number of at bats. A player that hits all doubles will have a higher slugging percentage than one who hits all singles.

Walks and Hits per Innings Pitched (WHIP): Calculated by dividing the sum of walks and hits allowed by a pitcher by the number of innings pitched by that player. This metric has allowed sabermetricians to roughly gauge pitching efficiency.

Defensive Efficiency Ratio (DER): A more complicated statistic, this measures the effectiveness of a team's defense, and cannot be applied to any single player. It consists of the total number of hits by opposing teams less home runs divided by the total number of balls hit into play less home runs. Homeruns are excluded because the defense does not have an opportunity to make an out on that play.

Runs scored (RS) Total number of runs scored by a team.

Runs Allowed (RA) Total number of runs allowed by a team.

Literature Review 2

Sabermetrics makes for an interesting research topic. This provides a fair amount of literature available for review, both on sabermetrics in general and the Pythagorean expectation specifically.

The Pythagorean expectation gained its name from its similarity to the famous Pythagorean theorem: $a^2+b^2=c^2$. Bill James also created a "runs created" metric, which attempts to better measure an offensive player's performance. It is not used much despite its accuracy, is difficult to model with traditionally recorded statistics. A less complicated model that predicts runs scored was needed for runs scored, and was provided by Jim Albert and Jay Bennett (Moy, 2006). The breakdown of runs scored as a sum of on base percentage and slugging, or bases per at bat, provides the simplicity needed to collect data from most

databases for analysis. Breaking down runs allowed as a sum of WHIP and defensive efficiency ratio serves as an obtainable defensive model. Many sabermetricians have analyzed the Pythagorean expectation and attempted to make its predictions more accurate. This has been done by changing the exponent used from 2 to roughly 1.8 (Moy, 2006). Though initially created to predict win percentage for baseball teams, it has been modified to predict win percentage for several sports. The exponents are changed and points for and against are substituted for statistics related to scoring and defense in those sports (Gosch, 2015). Rachel Gosch's model involving Weibull probability distributions looks at the effects of designated hitters on winning percentage. She uses a decomposition of the Pythagorean expectation to account for runs scored with and without a designated hitter. This clever use of the formula predicts the effect a designated hitter has on season win percentage. The use of this decomposition can also be used to evaluate the predicted effect of trades or injuries, with the assumption the player will play in each game and perform as they have in the past. It is important to note that the Pythagorean expectation is most accurate at predicting regular season games. This is because teams rely on five starting pitchers during the season, where in the playoffs they depend on only three or four starting pitchers (Winston, 2009).

Model 3

Variables 3.1

The variables used in this experiment, their possible values, and expected signs are listed in Table 1 below. We describe our reasoning behind our expectations of the coefficients' signs and their effects on the dependent variables.

Table 1: Variables

Variables	*Description*	*Values*	*Expected Sign*
Dependent Variables			
rs	Runs Scored	rs ≥ 0	
ra	Runs Allowed	ra ≥ 0	
Independent Variables			
obp	On-Base Percentage	1 ≥ obp ≥ 0	Positive
slg	Slugging Percentage	1 ≥ slg ≥ 0	Positive
whip	Walks Hits Innings Pitched	whip ≥ 0	Positive
der	Defensive Efficiency Ratio	1 ≥ der ≥ 0	Negative

The variable *rs* is dependent on *obp* and *slg*. We expect the coefficient for obp to be positive, as a higher on-base percentage should lead to more runs scored for a team. If a team has a lower on-base percentage they should score fewer runs, as a player must reach base to score. In baseball, a home run counts as reaching a base. Our intuition leads us to believe

the coefficient for *slg* should also be positive. Slugging is a measurement of “power” performance: the more bases per at bat, the higher the slugging percentage. The farther a player advances around the base path, the more likely it is they will score. Both of these being positive means we think that a higher occurrence of gaining at least one base per at bat combined with a higher number of bases gained per at bat will result in more runs scored in total by a team.

The variable *ra* is dependent on the variables *whip* and *der*. We expect *whip* to have a positive coefficient because the more walks and hits per innings pitched, the more likely it is that a run will be allowed to score. Players must make it on base to score, and more players on base should mean more runs allowed. We think *der* should lead to less runs allowed, making the coefficient negative. This is because of the nature of how *der* is calculated. The higher the value for *der*, the more outs per available attempts to get an out. We interpret this model as more base runners means more runs will be allowed. More walks and hits per inning should increase base runners, where more outs per available attempts to get an out should lower the amount of base runners.

Model 3.2

The Pythagorean expectation is that a team’s win percentage for the current season should be equal to the square of the number of runs scored by a team divided by the square of runs scored and the square of runs allowed by a team. Our model, equation 3.1 below, will use an exponent of 1.8 instead of the square of the variables.

$$win\ percentage = \frac{(runs\ scored)^{1.8}}{(runs\ scored)^{1.8} + (runs\ allowed)^{1.8}} \tag{3.1}$$

The reasoning behind this model is fairly intuitive. The percentage of wins should have a direct correlation with the percentage of runs they score in games they have played. The higher the ratio of their run scoring, the more games they should win. For example, if a team scores 30 runs and the total amount of runs scored is 50, that team should win more than one who only scored 10 out of 50 runs. Runs scored and runs allowed are team-level statistics. To allow for better analysis, we break runs scored and runs allowed into as many possible player-level statistics that still describe runs scored and runs allowed. To do this, we break runs scored into the sum of on-base percentage and slugging percentage. These are fairly intuitive pieces of runs scored, as it measures times a player gets on base and how many bases a player hits for. We regress runs scored on on-base percentage and slugging in equation 3.2 below to observe the marginal effect each has on runs scored.

$$rs = \beta_0 + \beta_1 * obp + \beta_2 * slg + \varepsilon \quad 3.2$$

β_1 is the effect a 1 unit increase in on-base percentage has on runs scored. β_2 is the effect a 1 unit increase in slugging percentage has on runs scored. Because both on-base percentage and slugging percentage are less than or equal to 1 by definition, these coefficients will be very large.

We also would like to view the marginal effects of any changes in the defensive statistics. This is done by regressing runs allowed on WHIP and DER, as seen in equation 3.3 below.

$$ra = \beta_0 + \beta_1 * whip + \beta_2 * der + \varepsilon \quad 3.3$$

As with the offensive equation 3.2, β_1 is the effect a 1 unit increase in walks hits over innings pitched has on runs allowed. β_2 is the effect a 1 unit increase in defensive efficiency ratio has on runs allowed. Unlike the offensive equation *whip* is not bound between zero and one, so β_1 should not have a value as large as if it was a percentage or a ratio. β_2, as surmised in the previous section describing the variables, should be negative. This is because as there are more outs per total attempts to gain an out, there should be fewer runs allowed.

Data and Descriptive Statistics 4

Data Sources 4.1

Data for this replication was gathered from MLB.com and consists of team-level hitting, pitching, and fielding data from 2005 through 2015. Historical team win percentage was gathered from Baseball-Reference.com, a leading baseball statistic aggregator. There were 30 teams in each of these 11 seasons, although some teams underwent minor changes that did not affect the continuity of the data. The Tampa Bay Devil Rays shortened their name to the Tampa Bay Rays between the 2007 and 2008 seasons. The Florida Marlins became the Miami Marlins between the 2011 and 2012 seasons.

There was no easy way to export the data from MLB.com or Baseball-Reference.com. We loaded the tables on the website and copied the data into an Excel file. We then compared the data to the original source to ensure integrity and imported it into Stata for use in regression analysis. We used data from 2005 to 2014 in our regression model, excluding 2015 for prediction purposes.

Descriptive Statistics 4.2

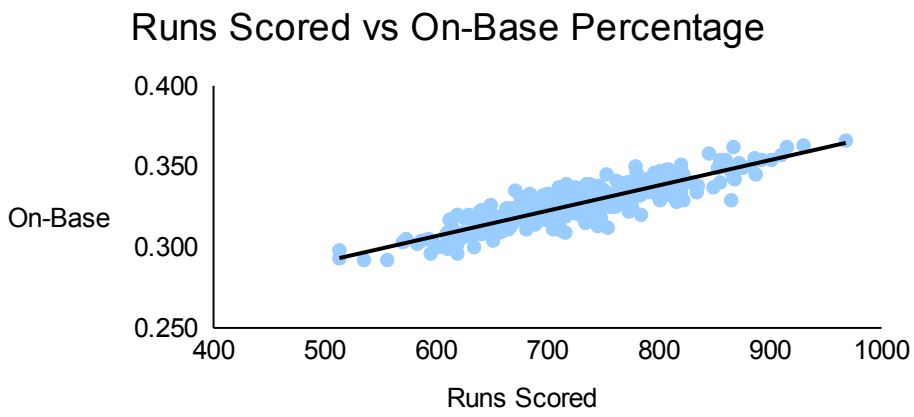
In this section we observe summary statistics of the data used in our regression analysis. We also observe the relationships between the dependent and independent variables in our regression models, both visually and in data tables. Following each graph and table we briefly discuss our findings and their relevance.

Table 2: Summary Statistics

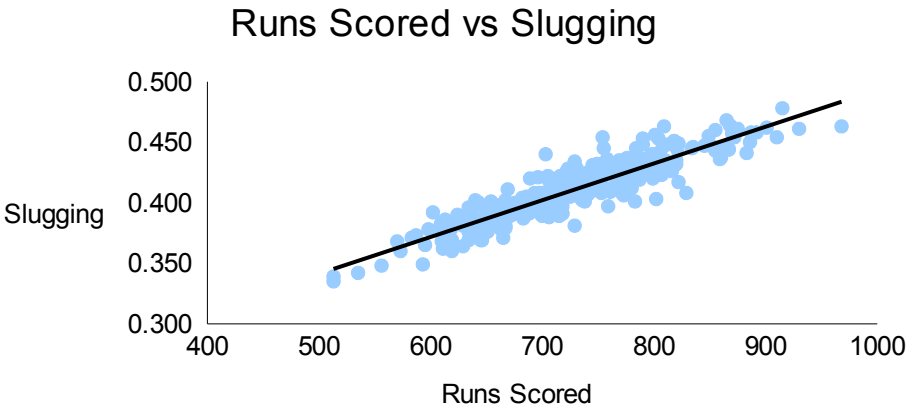
Variable	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
rs	724.660	79.096	513	968
obp	0.326	0.014	0.292	0.366
slg	0.410	0.026	0.335	0.478
ra	724.660	83.807	529	971
der	0.691	0.011	0.655	0.724
whip	1.351	0.090	1.140	1.600

The above Table 2 contains summary statistics for variables in each of our models, the first three contained in the offensive model and the last three in the defensive model. The mean values of our runs scored and runs allowed variables are equal. We would expect this to happen, as a run scored for one team counts as a run allowed for the other. If these were not equal it would signal a data integrity issue. The on-base percentage variable shows very little variation with a standard deviation of 0.014. The mean of *obp* should be interpreted as a runner gets on base once out of approximately three attempts. The mean of *slg* has a slightly different interpretation, which is for each at bat, approximately 0.41 bases are earned. This variable has slightly more variation, but still falls within a fairly tight range. The mean value for der is 0.691. This means that in this sample 69.1% of balls in play with a chance for a defensive out were realized. There is almost no variation in this variable either as the standard deviation is 0.011. This implies defensive consistency between teams across the sample. Finally, the mean value for *whip* is 1.351. This means that on average pitchers allowed over 1.351 base runners every inning. There isn’t much variation across the sample, with the standard deviation at 0.09. This implies that pitchers were fairly consistent between teams across the sample.

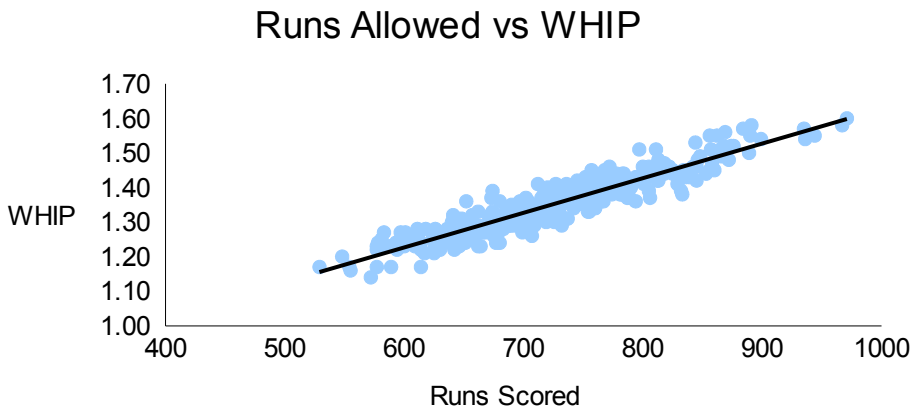
The following graphs provide a visual interpretation of the relationship between dependent and independent variables in both the offensive and defensive models.



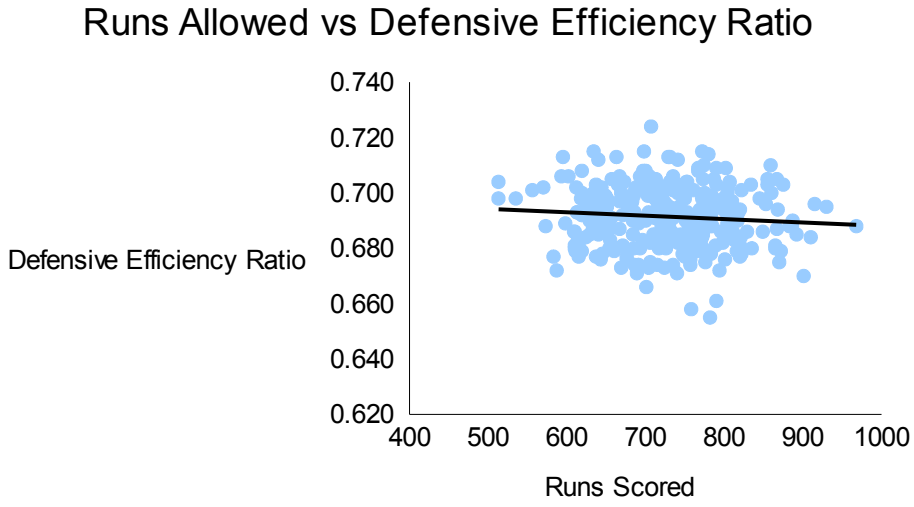
There looks to be a strong positive correlation between the number of runs scored and the percentage with which a team puts a runner on base. This makes intuitive sense as the more base runners a team has, the more likely that team is to score runs. A runner must gain at least one base to have a chance to score.



There is a strong positive correlation between slugging percentage and the number of runs a team scores. This is expected, if one recalls that slugging is total bases divided by total at bats. An increase in slugging percentage is interpreted as more bases gained per at bat, which makes scoring runs more likely.



This graph also shows a strong positive correlation between Walks and Hits per Inning Pitched and runs allowed. We predicted this earlier, reasoning that the more base runners a team allows, the more runs that team will allow. The idea that more base runners scores more runs is present in the offensive model, as well as the defensive model.



This graph shows a very weak negative correlation between *der* and *ra*. We reasoned a negative correlation between *der* and *ra* should occur because a team is recording more outs per attempt, reducing the number of base runners and accelerating the pace of the game simultaneously. There appears to be a very weak correlation between these two variables, leading us to believe that *der* may not explain variation in *ra*.

We constructed the correlation matrices in Tables 3 and 4 below to further study the relationship between the variables in each model.

Table 3: Correlation Matrix, Offensive Model

	rs	*obp*	*slg*
rs	1		
obp	0.888	1	
slg	0.911	0.775	1

The above correlation matrix confirms the observations from the preceding scatterplots. Runs scored are highly correlated with both obp and slg. The two independent variables are also highly correlated, which holds with definitions. Any type of base hit will record as an on-base event.

Table 4: Correlation Matrix, Defensive Model

	ra	*der*	*whip*
ra	1		
der	-0.673	1	
whip	0.933	-0.712	1

This correlation matrix confirms observations about the strong positive correlation between *whip* and *ra*, recalling the earlier reasoning that more base runners results in more runs for that team. A team that allows more base runners will have more runs allowed. The negative correlation between *der* and *ra* was reasoned earlier, however the strength of the correlation was estimated to be weak in the scatterplot. This high negative correlation between these two variables may signal an underlying relationship.

Empirical Results 5

Regression Analysis 5.1

After forming our models, gathering our data, and performing OLS regressions we collected the results in the following Tables 5 and 6. Equation 5.1 reflects the results gathered in Tables 5 and 6 and is used to predict win percentage in Table 7. Descriptions of each table's contents will follow each table.

Table 5: Regression Results, Offensive Model

Variables	Offensive Model
obp	2,570***
	(159.0)
slg	1,679***
	(83.16)
Constant	-801.9***
	(32.27)
Observations	300
R-squared	0.913

Robust standard errors in parentheses*** p<0.01, ** p<0.05, * p<0.1

The offensive regression model yielded highly statistically significant results. The interpretation of these results is for every 0.01 increase in on-base percentage, runs scored should rise by 25.7 over the course of the season. Slugging has a similar interpretation, with every 0.01 increase in slugging percentage yielding 16.8 additional runs. The constant, -801.9, represents the amount of runs predicted when both on-base percentage and slugging percentage are equal to 0. This is nonsensical by itself, but adjusts the marginal effects of both independent variables to fit the runs scored data.

Table 6: Regression Results, Defensive Model

Variables	Defensive Model
whip	858.0***
	(26.90)
der	-139.1
	(200.4)
Constant	-338.6**
	(165.2)
Observations	300
R-squared	0.871

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

The defensive regression model yielded significant results for *whip*, but did not yield individually statistically significant results for *der*. This is most likely due to the same reasons a weak correlation was observed between *der* and *ra* in the scatterplot. An F-test for joint significance found that both *whip* and *der* were highly significant, and thus should both

jointly be included in the model. The constant, as with the offensive model, provides adjustments to provide the best fit for coefficients.

When we used these regression models in our prediction model, we obtained equation 5.1 below.

$$win\ percentage = \frac{(2570 * obp + 1679 * slg - 801.9)^{1.8}}{(2570 * obp + 1679 * slg - 801.9)^{1.8} + (858 * whip - 139 * der - 338.6)^{1.8}}$$

5.1

The values in this formula are based on our prior regression of statistical data from 2005-2014. We used this model with these values to predict each team's win percentage for the 2015 season, shown in Table 7.

Table 7: Predicted Win Percentage

Year	Team	W%	PREDICT	VAR
2015	Arizona Diamondbacks	48.80%	51.25%	2.4500%
2015	Atlanta Braves	41.40%	39.17%	-2.2261%
2015	Baltimore Orioles	50.00%	49.77%	-0.2327%
2015	Boston Red Sox	48.10%	49.80%	1.6992%
2015	Chicago Cubs	59.90%	60.05%	0.1454%
2015	Chicago White Sox	46.90%	44.75%	-2.1525%
2015	Cincinnati Reds	39.50%	45.44%	5.9442%
2015	Cleveland Indians	50.30%	58.43%	8.1251%
2015	Colorado Rockies	42.00%	42.67%	0.6733%
2015	Detroit Tigers	46.00%	50.33%	4.3251%
2015	Houston Astros	53.10%	59.91%	6.8104%
2015	Kansas City Royals	58.60%	53.64%	-4.9615%
2015	Los Angeles Angels	52.50%	50.49%	-2.0137%
2015	Los Angeles Dodgers	56.80%	60.36%	3.5564%
2015	Miami Marlins	43.80%	46.14%	2.3448%
2015	Milwaukee Brewers	42.00%	44.39%	2.3877%
2015	Minnesota Twins	51.20%	46.37%	-4.8268%
2015	New York Mets	55.60%	56.82%	1.2152%
2015	New York Yankees	53.70%	53.42%	-0.2825%
2015	Oakland Athletics	42.00%	48.84%	6.8445%
2015	Philadelphia Phillies	38.90%	37.95%	-0.9527%
2015	Pittsburgh Pirates	60.50%	54.24%	-6.2578%

2015	San Diego Padres	45.70%	44.87%	-0.8277%
2015	San Francisco Giants	46.90%	56.67%	9.7746%
2015	Seattle Mariners	51.80%	49.86%	-1.9365%
2015	St. Louis Cardinals	61.70%	53.17%	-8.5268%
2015	Tampa Bay Rays	49.40%	54.66%	5.2640%
2015	Texas Rangers	54.30%	49.66%	-4.6400%
2015	Toronto Blue Jays	57.40%	64.59%	7.1896%
2015	Washington Nationals	51.20%	56.59%	5.3889%
			Average Error:	1.1434%

Conclusion 6

Our model that was trained on the data from 2005-2014 predicted win percentage an average of 1.1434% higher than actual win percentage, with a standard deviation of 4.6%. We are incredibly pleased with the accuracy of these results and that they mirror those obtained by other studies with different data. We found that we could predict a team's win percentage by using almost exclusively player-level statistics, allowing us to observe the marginal effects of lineup changes with respect to season win percentage. A noted weakness with this analysis is the lack of a player-level data for fielding percentage. It should be noted that fielding, and defense in general, has a smaller effect on win percentage than offense. Another weakness of this study is that qualitative data is difficult to integrate into the model. A bad manager may mismanage situations that should statistically yield a win and induce the team to lose. Extensions of this study may include the exploration and incorporation of managers' performance and its effect on win percentage.

References 7

Gosch, R. O. (2015). Incorporating the Effects of Designated Hitters in the Pythagorean Expectation. *Department of Mathematics, University of Texas at Austin*.

Moy, D. (2006). Regression Planes to Improve the Pythagorean Percentage. *University of California - Berkeley*.

Winston, W. L. (2009). Baseball's Pythagorean Theorem. *Princeton University Press*.