2014

Pythagorean Formula in Baseball: Creation of a Better Model Using Backward Elimination Stepwise Regression Analysis

Kellen R. Taylor

PYTHAGOREAN FORMULA IN BASEBALL: CREATION OF A BETTER MODEL USING BACKWARD ELIMINATION STEPWISE REGRESSION ANALYSIS.

Kellen Radale Taylor

Master's Project

Submitted to the School of Human Movement, Sport, and Leisure Studies
Bowling Green State University

In partial fulfillment of the requirements for the degree of

MASTER OF EDUCATION
In
Sport Administration

(December 8, 2014)

Project Advisor

Dr. Amanda Paule-Koba _____

Second Reader

Dr. David Tobar _____

Table of Contents

Abstract

Since the revealing of "Moneyball", baseball organizations have increased its focus on the importance of statistical analysis (Lewis, 2003). This study attempts to create a model using baseball statistics that can predict a team's wins, more accurately than the Pythagorean formula. The Pythagorean formula measures actual or projected runs scored against runs allowed and projects a team's won-loss percentage (James, 1980). While this measure is accurate within reason, it excludes traditional and newer statistical measures from the equation. The author hypothesizes that a better model can be produced from more advanced statistical analysis, since Bill James' developed the formula through experimental observation (James, 1980). This study uses backward elimination regression analysis from batting, pitching, and fielding statistics beginning with the 2005 season through the 2014 season to create a formula. The purpose of this study is to determine whether backward elimination regression will create a model that is more accurate at predicting wins than the Pythagorean formula. An additional forced entry regression analysis finds the amount of variance accounted for by the variables included in the Pythagorean formula. R^2 values from both analyses were compared and the SEE from each equation was compared. The results indicate that a better model was created

W = 28.723 +(.076*runs)+(.148*OPS+)+(.437*saves)-(.065*runs allowed)+(.09*ERA+)+ (1.537*SO/W). *Equation 1*

This model accounted for 92.7% of the variance while the Pythagorean formula variables accounted for 86.8% of the variance. The SEE of each formula resulted in the regression model, SE=2.99, being slightly better than the Pythagorean formula, SE=4.02. This study suggests that the model created through this study is about one game more accurate at predicting wins in a season.

Introduction

In today's baseball world, organizations are using statistics like never before (Lewis, 2003). Entire departments within front offices are dedicated to analyzing statistics to gain a competitive edge over their competition. In a game where intuition has prospered for the majority of its existence, sabermetrics has evolved as a method to discover objective data about baseball (James, 1980). Bill James, the most recognizable name to most people, helped publicize this new method using limited statistics to answer questions about how baseball operates, through his *Baseball Abstracts* (James, 2001). This lead to the method of "Moneyball", the process of finding undervalued baseball players to help build a cost-effective team (Lewis, 2003) and Farrar and Bruggink (2011) discussed how "Moneyball" has yet to diffuse to every team in the Major League Baseball. James discovered that runs scored and runs allowed were the biggest predictors of winning games and created a model using only these statistics called the Pythagorean Formula (James, 1980). Research has been conducted to verify the formula's effectiveness and also add improvements regarding the exponent in the formula (Cha, Glatt, & Sommers, 2006; Miller, 2007). The formula has since been applied to other sports like football (Lewis, 2008), basketball (Ostfield, 2006), soccer (Carlisle, 2008), and hockey (Mason & Foster 2007) and has used the same concept of measuring the relationship of points scored against points allowed.

The Pythagorean formula uses the final outcome to determine won-loss percentage and thus, withholds all of the "how", from the equation. The amount of runs scored is dependent on multiple aspects of a team's offensive performance like the amount of hits a team has, its ability to get on base, drive in runs, steal bases etc. The same is true for the amount of runs allowed or a team's ability to prevent runs from being scored on them (hits allowed, walks allowed, fielding

percentage, etc.). Understanding how to produce wins is instrumental in developing an overall plan for an organization, regardless if that plan comes from the front office or future employees of a team, baseball writers, or diehard fans.

With this void of "how" in mind, this study is expected to add to the baseball statistical literature by having two purposes. The first purpose is to determine if there are any other statistics, in addition to runs scored and runs allowed, that are key predictors in determining wins in a baseball season. The second purpose is to determine if a new model can be built using additional statistics that can better predict wins than the Pythagorean formula and if so, provide the formula and compare them.

Literature Review

This literature review will start by introducing statistics, sabermetrics and Bill James' Pythagorean formula and its contribution to baseball. This will be followed with a description of the alternative method to project wins proposed by this study. Following this, a brief description of the varying types of statistics will be discussed and the review will end with a look at the Pythagorean formula and its recent adjustments from the 1980 version thanks to recent discoveries.

Sabermetrics Births the Pythagorean Formula

Sabermetrics was coined by Bill James in 1980 in honor of the Society for American Baseball Research (SABR) and defined it as "the search for objective knowledge about baseball" (Birnbaum, 2014, pg. 1). Sabermetrics use statistical analysis to answer questions about baseball. How often does a team get on base? How many batters does a team strike out per game? How many runs are yielded because of a team's fielding errors? Sabermetrics answers the type of questions that aren't left up to opinion.

Sabermetrics is relatively new in the mainstream spotlight. This is due to the release of Michael Lewis' book "*Moneyball: The Art of Winning an Unfair Game*" (2003) and its corresponding movie "Moneyball". The book tells the story of how the Oakland Athletics were able to compete with the rest of Major League Baseball using sabermetrics, despite losing notable stars like Jason Giambi, Johnny Damon and Jason Isringhausen and competing against the substantial payrolls of teams like the Yankees at 126 million dollars, with only about 40 million dollars committed to their own roster. The fact is, however, that sabermetrics have been used well before this time.

As baseball developed from its inception, the statistics did as well . The box score, which is used to chart the number of the single count statistics in the game, was created by Henry Chadwick in the mid-19th century (Birnbaum, 2014). The box score is still used today as the trademark way to summarize games. Chadwick developed the box score to closely resemble that of a cricket box score, the game he related baseball to the most (Schiff, 2008). Chadwick's statistics helped front office personal determine what players were effective and which ones weren't. Second, he developed the box score so the fans could better understand the game. Since the fans are the biggest supporters of baseball he wanted the fans to know how to tell which players were the ones that deserved recognition and the ones that didn't (Schiff, 2008, p. 87).

F.C. Lane was a writer for Baseball Magazine from 1912 to 1937 where he developed his thoughts into a book called *Batting: One Thousand Expert Opinions on Every Conceivable Angle of Batting Science: The Secrets of Major League Batting and Useful Hints for Hitters...*, in 1925. This man's book and ideas had great thoughts on the statistics used in that era. He discussed how batting average could be misleading and also suggested that different outcomes of an "at bat" should have different weights on the batting average (Lane, 1925). This discussion eventually led to the creation of the slugging percentage statistic, which is wildly used today (Schwarz, 2004). Ernie Lanigan added the RBI to the box score that Chadwick had developed. In the 1940s Branch Rickey, who boldly took the risk to sign a young 26-year-old Jackie Robinson from the Negro Leagues, hired a statistician named Allan Roth to evaluate the Brooklyn Dodgers' players. Roth was a fan of baseball and obsessed with the statistics surrounding it. Roth had to convince Branch Rickey that he could actually help the team via statistical analysis. He proved this by helping the Dodgers win their first ever World Series in 1955. Almost 50 years before MLB front offices would put an emphasis on OBP (because of "Moneyball"), Roth recommended that

the Dodgers do the same (Soule, 1957). A few decades later Hall of Fame manager Earl Weaver used analytics to critique his platooning system and pitching changes for the Baltimore Orioles. These incidents led a Johns Hopkins, professor, Earnshaw Cook, to write two books about the statistics of baseball and how they are analyzed, *Percentage Baseball* (1966) and *Percentage Baseball and the Computer* (1971). This all lead to one man doing more for the baseball statistics community than anyone else, Bill James.

Bill James attended the university of Kansas and studied three things: economics, literature, and baseball. James' first book, *1977 Baseball Abstract: Featuring 18 Categories of Statistical Information That You Just Can't Find Anywhere Else*, was self-published and largely ignored. From the first book he posed a theory about fielding statistics, more specifically the error. James thought that the error was a lie because it was an arbitrary statistic to whoever was the official scorer of that game. An error is described as a play that should have been made and was not. To this day, an error still is up to the scorer to deduce if a player could have made the play. James devised a different method called the range factor, which measured the amount of successful plays a fielder made in a game (James, 1977). This is one of the first statistics James created to evaluate the game differently, and this statistic is used greatly today to determine a players fielding success.

In James' third book, *Baseball Abstract,* published in 1979 he created the "Runs Created" model. This model was used to predict how many runs a team would score using the team's number of walks, steals, singles, doubles etc. during one season. In 1980 in what was now an annual book, *Baseball Abstract,* he published the Pythagorean formula for predicting wins of a baseball team using a team's runs scored and the number of runs the team allowed (James). The formula was first used to retroactively evaluate a team based on the previous years statistics. For

example, if a team posted a record of 85-77 and their Pythagorean prediction was calculated to be 81-81, then the team was said to exceed expectations for the season. This formula has evolved over time, but major outlets like ESPN use this formula to evaluate a team's season and how offseason additions or subtractions may change a team's outlook for the next year. This gives, reporters, writers, and executives a starting point to discuss the outlook of a team. An additional use for the Pythagorean formula is to insert the amount of predicted runs and runs allowed via a forecasting system like Marcel, Fans, ZiPS, Steamer, Oliver, CAIRO, PECOTA, or CHONE (Fangraphs, 2014). These systems have different algorithms for predicting player and team statistics that can be inserted into the Pythagorean formula. After determining if a better predictive formula can be made through this study, these types of statistics should be used to predict team wins for an upcoming season.

For this reason, it is important to determine if the Pythagorean formula is still the best way to predict expected wins for a baseball team. If a better model were possible it would help everyone that is invested in baseball. The writers and reporters would have a better starting point with their analysis and discussions of MLB teams. The managers will better understand what statistics account for the most variance of wins among the many statistics that we have today. A better model would aid front office personal that are building the team either from scratch through the farm system or looking to add a productive piece by adding a free agent. The playoffs can be lost or won by one game and the importance of a better model is paramount in determining if a team needs more help or better players on their roster. Front office personal will have a better estimate for how their team will perform and how that success or lack there of will affect their ability to make money during the season. Yost and Rainey (2009) discovered fans who can't get invested in the team reported lower levels of conversations about the team, lower

levels of team website visits, and watched fewer games on television. This lack of self-identification leads to reduced ticket sale, and merchandise profit. If a team can figure out if they are going to have a down year, they can prepare their finances and budget accordingly to help with the lack of revenue. Through his research, Surdam (2011), concluded that competitive balance affected individual teams and the league as a whole and affected profits and attendance records. Surdam noticed that in the MLB, a year in which a league has a runaway winner, the league also produced its poorest profits (2011). The Pythagorean Formula needs to be challenged so baseball can continue to develop on the field and off the field and better prepare itself for changes in wins, losses, attendance, television viewership, and profits.

Regression Analysis

Baseball is so overwhelmed with baseball statistics that a regression analysis is needed in order to try to build this better formula for predicting the amount of wins for a team in a season. There are two types of variables within this study, predictor variables and response variables (independent and dependent). The response variable for this study will be the amount of wins in a given season. The predictor variable will be the team statistics accumulated by baseballreference.com for the corresponding year and wins. Regression analysis will identify if there is a functional relationship between the predictor and the response variables and create a model based on these relationships (Draper & Smith, 1998). Building this model will use stepwise regression, which is, a step-by-step approach for either including or eliminating a variable in the model. There are three different approaches to stepwise regression: forward selection, backward elimination, and Efroymson's (a combination between forward, and backward) procedure (Qinggang, Koval, Mills, & Lee, 2008). Forward selection starts with a blank canvas and adds variables to the model until a satisfactory equation is produced. Backward

elimination starts with every stat and eliminates variables based on their F-test value. Draper and Smith (1998) say that backward elimination "is much more economical of computer time and manpower" and "a satisfactory procedure, especially for statisticians who like to see all the variables in the equation once in order 'not to miss anything'" (p. 307). For this reason this study hypothesizes:

H1: Using backward elimination regression, this study will create a model that is more accurate at predicting wins than the Pythagorean formula.

Type of Statistics

There are many statistics that baseball executives, sports writers, and baseball fans can look at when evaluating their team. Two different types of statistics will be used in this study. One statistic is the simple count method, which can be simply calculated by using the sum of individual games statistics at the end of the season. This includes batting, pitching, and fielding statistics (listing of the variables and their abbreviations can be found in the Appendix A). In the batting category, mainstream statistics include the amount of times that a player got up to bat ("plate appearance" or "PA"), the number of hits a batter produced ("hits" or "H"). There is also the kind of hit the batter produced, such as a hit that resulted in reaching all the bases without an error in the field ("homerun" or "HR"), a hit resulting in three bases without an error in the field ("triple" or "3B"), and so on with two bases ("double" or 2B). Runs Batted In (RBI) is also another important offensive statistic, which credits the player at bat for getting a base runner across home plate as a result of their PA. Within each category of statistics the abbreviations for things like hits, walks (BB), and runs are the same, but instead, have a different meaning under a new context. For example, within the pitching category of team statistics, the abbreviation hits,

now means, “hits allowed” by that team, and so on with other statistics like homeruns, walks, and strikeouts (Official Rules).

Other statistics are used to describe an interrelationship between two variables or statistics. Batting average (“BA”) is the relationship between the amount of hits a batter produces and how many times a player gets up to bat (minus walks and hit by pitches). Another example would be On-base percentage (OBP), which describes the relationship between how often a player gets on base (hits + walks + hit by pitch) and how many times a player gets up to bat. These types of statistics help tell a story about how the player or team performs relative to the game or season. Some of these statistics are traditional, ones that have been used for over a century while others are relatively new to the baseball world. Hakes and Sauer recently published *An Economic Evaluation of the Moneyball Hypothesis* (2006) where they tested how much variance OBP and slugging percentage accounted for when using win percentage as the response variable (two interrelationship statistics). They concluded that on-base percentage accounted for 82.5 percent and slugging accounted for 78.7 percent of the variance. These numbers will provide a good comparison to see if on-base and slugging percentage still account for a high amount of variance in the model.

The Pythagorean Formula

In 1980, Bill James developed a method for determining a team’s Won-Lost percentage called the Pythagorean formula. James concluded, since games are won by scoring more runs than the opponent, these two factors are the biggest predictors of won-lost percentage (James, 1980). The Formula is as follows:

$$\text{Won-Lost Percentage} = \frac{(\text{Runs Scored})^{\wedge}2}{(\text{Runs Scored})^{\wedge}2 + (\text{Runs Allowed})^{\wedge}2}$$

Because the Won-Lost Percentage isn't what will be examined in this study, we can simply change this formula to only include the number of games won the formula predicts by multiplying by the number of games played during the season:

Number of Games Won = Won-Lost Percentage * 162 games

Cha, Glatt and Sommers (2006) have published the most recent evidence of Bill James' Pythagorean Formula by testing the baseball seasons from 1950 to 2007. They noted that the formula is still very accurate for predicting wins for a team, but since 2000 the exponent for the formula has changed from 2 to 1.94. With this in mind, this study will alter Bill James' formula to include this new finding.

$$\text{Won-Lost Percentage} = \frac{(\text{Runs Scored})^{\mathbf{1.94}}}{(\text{Runs Scored})^{\mathbf{1.94}} + (\text{Runs Allowed})^{\mathbf{1.94}}} * 162$$

Statistical analysts have not exclusively attempted to create a model that accounts for a larger variance of wins. Instead of investigating this, researchers have assumed that certain statistics are important and have neglected to dig any deeper. It took 23 years, since James wrote his first abstract in 1979 for one team, the Oakland Athletics, to adopt any of Bill James strategies (2002) and another year (2003) before the Boston Red Sox reached out to acquire his services. Baseball has notoriously been slow to develop and to embrace new ideas, and this study will hopefully help speed both processes.

Method

Unlike the formula for the Pythagorean formula this study will use a combination of batting, pitching and fielding statistics to attempt to create a better formula for predicting team wins than the Pythagorean formula.

Sample

The data that will be used are the batting, pitching, and fielding statistics beginning with the 2004 season through the 2014 season. These years were chosen to create a large and current sample size for this study. Some statistics will be excluded because they are constant for each team, such as, games played, games pitched etc. Team wins will be the response variable while the remaining statistics will be used as the predictor variables. The data will correspond with the number of wins the given team had and the statistics that team generated during that season. The identification of a team is irrelevant and will be held out of the process.

Data Collection

Data will be collected from Baseballreference.com from each corresponding season from 2004 through 2014. Team statistics are publicly available for use. The amount of variables used in this study is 67 and consist of batting, pitching, and fielding statistics (all variables can be found in Appendix A). This amount and variety are used to represent the three important phases of baseball contrary to only two variables used in the Pythagorean formula.

Data Analysis

This study will input all data into the Statistical Package for the Social Sciences (SPSS) version 21 and run the analysis through its programming. The study will first attempt to use all of the statistics provided by baseballreference.com. If assumptions are violated then a second test using only variables with a large correlation coefficient (any variable with a Pearson correlation

over +-.5) will be used in the regression analysis (Gujarati & Porter, 2009). Using the constant from the final model of the backward elimination and the unstandardized coefficient B of the included variables, this study will attempt to create a better predictive formula (equation seen below) where Y_i is the dependent variable or wins, b_0 is the constant, and b_1 is the B coefficient representative of the corresponding independent variable that is X_i etc.

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

A second forced entry regression analysis will be used to determine how much variance can be explained using only the variables of the Pythagorean Formula and the results will be discussed. Expected wins will then be calculated using the restructured Pythagorean Formula, amended by Cha, Glatt, and Sommers (2006). Finally a comparison of the two models standard error of the estimate (SEE) will be examined to see if the model built by regression predicts wins more accurately.

Results

After inputting all of the statistics into the regression analysis the model was able to account for 97% of the variance with an F value at 639.68, $p < .001$ and standard error of the estimate at 1.74. The model was reduced from 67 variables to 20 (all variables can be found in Appendix A). The 20 variables were: TRIPLES, BK, HBP, WP, HBP (allowed), SV, IP, BB, TSho, DOUBLES, SH, CG, LOB (offensive), HR, BA, ER (allowed), GF, RpG, PA, and OPS (description of variables can be found with Appendix A). However, a problem occurred regarding the multicollinearity of the variables. Variance Inflation Factors (VIF) were well over the recommended maximum value of 10 for the variables RPERG, PA, DOUBLES, TRIPLES, HR, BB, BA, OPS, LOB, GF, CG, tSHO, and IP (can be seen in Appendix B, Table 1), and therefore a second attempt had to be conducted.

Going back to the correlation results, 11 variables were shown to have a large ($r > +/- .5$) correlation coefficient and all had a significant value ($p < .001$), which can be seen in Appendix B, Table 2. Only these variables were used in the second attempt at the backward elimination regression analysis and this helped the model meet all assumptions. These statistics were: ERA+, SV, WHIP, RALL, OPS+, ERALL, R, RBI, RPG, H9, SOWITHW (all variables are defined within the Appendix A). ERA+ had the largest positive correlation ($r = 0.719$) with saves the second ($r =0.653$) and OPS+ ($r = 0.599$) third. These positive correlations are logical within the baseball context because the better the ERA+ and save value is the better the pitching and the same is true for OPS+ in respect to batting. The largest negative correlations were WHIP ($r = -0.617$), RALL ($r = -0.608$), and H9 ($r = -0.522$). These statistics also make sense from a baseball standpoint; as each of these values increases, the quality of a team's pitching decreases. It is also

interesting to note that of the 11 variables with large effect sizes, four correspond to a team's ability to score runs, and seven correspond with preventing runs.

Next, a backward elimination regression analysis reduced the model from 11 variables to six variables. The six variables included are R, OPSPLUS, SV, RALL, ERAPLUS, and SOWITHW and the model excluded RPERG, H9, ERALL, RBI, and WHIP. The final model (Table 3) shows that these variables (R^2 = .927) account for 92.7% of the variance of wins and has a low standard error of the estimate value (SE = 2.99). By using the constant from the final model of the backward elimination and the unstandardized coefficient B of the included variables this study was able to produce a formula to account for 92.7% of the variance: W = 28.723 +(.076*R)+(.148*OPS+)+(.437*SV)-(.065*RALL)+(.09*ERA+)+ (1.537*SOWITHW). By looking at the ANOVA test in Appendix B (Table 4), we can see that the final model has a powerful F statistic of 688.144 while remaining significant at $p<.001$.

This model was tested to reveal any assumption violations. We can see the tolerance, which measures redundant predictors and VIF levels in Table 4 and neither of them violate their assumptions. That is, the tolerance values were greater than .10 and the VIF values were less than 10. Using a Histogram (Figure 1), the study shows that the results have no outliers and cook's distance shows no values over 1 (Table 5). Next the model was tested for any heteroskedasticity by plotting the residual by the predicted value (Figure 2). From this figure we can see that there is no pattern and the plots are random.

Another regression analysis was used to determine how much variance runs scored and runs allowed accounted for when using wins as the response variable. This analysis resulted in these two variables having an R^2 value of .868, accounting for 86.8% of the variance and has a

standard error of the estimate value at 4.02 (Table 6). This amount of variance is only 6% less than the new model with an additional four variables.

Next the Pythagorean formula was used to predict wins using the statistics accumulated by baseballreference.com. The results showed that the standard error of the estimate was 4.04. Table 7 shows a sample of how the two models compare with the actual wins. By comparing the SEE from each model we can see that the new model, (SEE = 2.99) has a lower value than the Pythagorean formula (SEE = 4.04) and that the new model results in a closer prediction to actual wins by one game.

Discussion

This study was intended to answer the question: are there any other statistics that are key predictors to predicting wins, and provide an alternative formula incorporating these predictors that is better than the Pythagorean formula. The results from this study support the hypothesis that a backward elimination regression analysis can produce a better and more accurate formula to predict wins using additional statistics. When using the regression equation compared to the Pythagorean formula we see that there is a one game decrease in the SEE of the number of games that a team would win in a season. This means that the regression formula is one game better than the Pythagorean formula. The question of the equation's simplicity is easy to answer as well. While the regression equation is more accurate it also uses four more variables than Bill James' Pythagorean formula. Bill James' method compared with this study's results still suggests that the Pythagorean formula is a more simplistic way to predict wins in a season.

From the formula created in this study we see that the positive and negative correlations make sense. As runs scored increases the amount of wins increase. The same is true for OPS+, SV, ERA+, and SOWITHW. The only negative correlation is RALL, which means the more runs a team allows the less wins the team will produce.

The formula produced by the regression analysis seems to tell the same story as previous literature. In the past, research performed by James (1980) and Hakes and Sauer (2006) concluded that runs, runs allowed, slugging, and on-base percentage were the most powerful predictors of wins throughout a season. Within this study, the statistics runs, runs allowed and OPS+, which accounts for on-base percentage and slugging percentage relative to the league average (the higher the points the better the team), are all used in the final formula. This study supports the theory that on-base percentage and slugging percentage are powerful statistics in

determining wins and suggests that using OPS+ is better at evaluating a team than the two former statistics.

This study also suggests the importance of three pitching statistics SV, ERA+, and SOWITHW. These additions to the equation are important when front office personal or analysts of any kind are evaluating a team's pitching staff. By looking at a predicted amount of SV and SOWITHW, using one of the forecasting systems mentioned in the literature review, compared to the rest of the league, the team can begin to improve these statistics through trades, free agency or the amateur or international draft.

Another important note is that direct fielding statistics did not make a large impact in the correlation results or in the equation created by the model. The closest one to be considered, but was ultimately left out because of a medium effect size was rtotPERyr or Total Zone Total Fielding Runs Above Average per 1,200 innings. This means the average runs a team's fielding is worth (values usually range between +/-10, higher than zero being above average).

Conclusion

This study has shown that a better model for predicting wins can be created out of a backward elimination regression analysis than the Pythagorean formula. This study uses six variables to calculate a prediction for wins and accounts for 6% more variance compared to the Pythagorean formula variables. These four additional predictors can have huge future implications for baseball organizations. Since the overall goal of baseball is to score runs on an opponent and to keep the opponent from scoring, one can hardly wonder why runs and runs allowed account for so much of the variance. But now there is additional insight into how to improve a team's win value with the additions of OPS+, ERA+, SOWITHW, and SV. From this study, focus can now be directed to closing tight games and notching a save. With the addition of SOWITHW, front office personnel may now realize it is not the amount of strikeouts a pitcher accumulates, but the amount he has relative to the walks he issues. Focusing on the ratio can give baseball teams a slight edge on the competition.

The one win improvement or 6% increase in variance accounted for can greatly impact baseball organizations. A team that has suffered losing seasons year after year or a team that has been excluded from the playoffs can be influenced by one game. One game can determine if a general manager is selling or buying at the trade deadline, or closing or opening your stadium during the postseason (resulting in playoff income). This is exactly what happened to the 2009 Detroit Tigers when they ended the season one game behind the Minnesota Twins in the American Central Division and again to the Seattle Mariners finishing one game behind the Oakland Athletics in the 2014 Wild Card standings. These type of instances happened eight out of the ten years used in this study (ESPN). Using this formula will help decision makers gain a better understanding of how they compare competitively to the other teams in the league.

Implications

From this research, baseball executives from the top down can begin to restructure their plans to build a better baseball team. By using the results, general managers, scouting directors, player development personnel, hitting coaches, and pitching coaches can be given additional criteria for improving entire teams or players individually that has not been discovered in previous texts. First, OPS+ is a better indicator than on-base percentage and slugging percentage individually. By using this research, statisticians, general managers, and hitting coaches will begin looking at OPS+ as the primary statistic rather than a secondary one. Second, teams need a reliable player that can "close" or save a game. This study has increased the importance of playing an eight inning by having an established closer to hand the ball to in the ninth inning. Lastly is the importance of the SO/W statistic. Pitching coaches and general managers will want to acquire players that have high values of this statistic or start with this statistic when evaluating the baseball team. The higher it is the better.

Limitations

Limitations occurred in the study due to the multicollinearity among the predictor variables. This is probably due to the fact that so many of the interrelationship statistics are derived from the simple count statistics. This lead to being cautious when determining what variables to use in the analysis, which is why only large effect sizes and significant values were chosen to proceed further with the model. Another limitation is the quality of statistics available. Baseballreference.com is a combination of statistics and a lot of the information comes from retrosheet.org. This mixing of statistics could lead to unreliable data within the study.

Future Studies

From this study it has been noted that runs and runs allowed account for a significant portion of the variance of wins. While confirming this theory, studies should dive deeper and find out the powerful predictors that help keep runs off the board. A study should explore using some type of regression analysis using pitching and defensive statistics to determine a model for predicting runs allowed. From this a defensive "Moneyball" may take shape and transform the baseball industry yet again. Another recommended study would use predicted team statistics prior to a season and use them to predict wins using this study's formula and compare them with the correlated season's wins to test if this formula is accurate with predicted results.

Another recommendation for future studies would be to only use simple count statistics in the regression analysis. Additional research can also be done to discover if any other statistics can be added to increase the model's accuracy. Further studies can likewise test the seasons before and after that of this study to test if this formula is still better compared to the Pythagorean formula. This study can also be used to determine if better models can be created in other sports that use the Pythagorean formula. This study also used simple statistics and statistics with interrelationships. Having a different study that created a model for each individual type of statistic may also help identify what the absolute best combination is for estimating wins.

References

Carlisle, J. (2008). Beane brings moneyball approach to MLS. ESPNsoccer. Retrieved from http://soccernet.espn.go.com/columns/story?id=495270&cc=5901

Cha, D. U., Glatt, D. P., & Sommers, P. M. (2006). An empirical test of Bill James's pythagorean formula. *Journal of Recreational Mathematics, 35(2), 117-124.*

Cook, E. (1966). *Percentage baseball* 2d ed: Mass. Inst. Of Technology

Cook, E. (1971). *Percentage baseball and the computer*. Baltimore, MD; :sn.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2d ed.). New York: Wiley.

Farrar, A., & Bruggink, T. H. (2011). A new test of the moneyball hypothesis. *Sport Journal*, 14(1), 1-13.

Gujarati D. N. & Porter D. C. (2009) *Basic econometrics*, 5th edition, Boston: McGraw-Hill Irwin.

Hakes, J. K., & Sauer, R. D. (2006). An economic evaluation of the moneyball hypothesis. *Journal of Economic Perspectives*, 20(3), 173-185.

James, B. (1979). 1979 Baseball abstract. Lawrence, KS: self-published

James, B. (1980). 1980 Baseball abstract. Lawrence, KS: self-published

James, B. (2001). The new Bill James historical baseball abstract. New York, NY: FreePress.

Lane, F. C. (1925). *Batting: One thousand expert opinions on every conceivable angle of batting science: The secrets of major league batting and useful hints for hitters....* Baseball Magazine Company.

Lewis, M. 2003. *Moneyball: The art of winning an unfair game*. New York, NY: Norton

Lewis, M. (2008). *The blind side*. New York, NY: W.W. Norton & Company.

Mason, D. S. & Foster W. M. (2007). Putting moneyball on ice? *International Journal of Sport Finance*, 2, 206-213.

Miller, S. (2007). A derivation of the pythagorean won-loss formula in baseball. *Chance Magazine,* 20(1), 40-48.

MLB Standings. (2014, October 28) Retrieved from http://espn.go.com/mlb/standings/_/date/20141028

Ostfield, A. J. (2006). The moneyball approach: Basketball and the business side of sport. *Human Resource Management*, 45, 36-38.

Official Rules. (2013, December 13). Retrieved November 12, 2014, from http://mlb.mlb.com/mlb/official_info/official_rules/official_rules.jsp

Qinggang, W., Koval, J. J., Mills, C. A., & Lee, K. D. (2008). Determination of the selection statistics and best significance level in backward stepwise logistical regression. *Communication in Statistics: Simulation & Computation,* 37(1), 62-72.

Schiff, A. J. *The father of baseball, A biography of Henry Chadwick*. North Carolina: McFarland & Company, Inc., Publishers.

Schwarz, A. (2004). *The numbers game: Baseball's lifelong fascination with statistics.* New York, NY: Thomas Dunne Books.

Soule, G. B. (1957). How they're using mathematics to win ball games. *Popular Science,* 17(1), 64.

Surdam, D. G. (2011). *Wins, losses, and empty seats: How baseball outlasted the Great Depression*. University of Nebraska Press.

Yost, J. H., & Rainey, D. W. (2014). The consequences of disappointment in team

performance among baseball fans. *Journal of Sport Behavior*, 37(4), 407-425.

Appendix A

* All statistics are reported in the order batting, pitching, and fielding and insertion into SPSS

Batting Statistics

RPG – Runs per game.

PA – Plate Appearences

AB – At bats

R – All Offensive runs scored by the team.

H - Hits

2B - Doubles

3B - Triples

HR – Home runs

RBI – Runs Batted In. The amount of runs a team forces home from an at bat.

SB – Stolen bases

CS – Caught Stealing

BB - Walks

SO – Strike Out

BA – Batting Average

OBP – On-base Percentage

SLG – Slugging Percentage

OPS+ - (On-base percentage (OBP) + Slugging percentage (SLG) / league SLG – 1) *100. OBP + SLG relative to the league average. Values at 100 are said to be average, while values over and under are above average and below average respectively. OPS+ is the adjustment to a team's ballpark.

TB – Total Bases

GDP – Ground into Double Play

HBP – Hit By Pitch

SH – Sacrifice Hits

SF – Sacrifice Flies

IBB – Intentional Walks

LOB – Left On Base

Pitching Statistics

ERA – Earned Run Average

GF – Games finished

CG – Complete Games

TSHO – Team Shutouts

CSHO – Complete Game Shutouts

SV – Saves – given to a pitcher that closes the game under certain conditions. Credit a pitcher with a save when he meets all three of the following conditions:

(1) He is the finishing pitcher in a game won by his club; and

(2) He is not the winning pitcher; and

(3) He qualifies under one of the following conditions:

- (a) He enters the game with a lead of no more than three runs and pitches for at least one inning; or

- (b) He enters the game, regardless of the count, with the potential tying run either on base, or at bat, or on deck (that is, the potential tying run is either already on base or is one of the first two batsmen he faces; or

- (c) He pitches effectively for at least three innings. No more than one save may be credited in each game (Official Rules).

IP – Innings Pitched

HALL – Hits Allowed

RALL – Runs allowed by the team

ERALL – Earned runs allowed by a team. An earned run is one that was not enabled by a fielding error, while RALL accounts for all runs.

HRALL – Hits Allowed

BBALL – Walks Allowed

IBBALL – Intentional Walks Allowed

SOALL – Strike Outs Allowed

HBPALL – Hit By Pitch Allowed

BK - Balks

WP – Wild Pitches

BF – Batters Faced

ERA+ - leagueERA/teamERA *100. ERA+ values at 100 are said to be average, while values over and under are above average and below average respectively. ERA+ is the adjustment to a team's ballpark.

FIP – Fielding Independent Pitching

WHIP- Walks + Hits allowed per inning pitched. The average whip for the 2014 was 1.275

H9 – the amount of hits given up by a team's pitching staff per 9 innings.

HR9 – Homeruns per 9 innings

BB9 – Walks per 9 innings

SO9 – Strike outs per 9 innings

SOWITHW – SO/W – strikeouts/walks – illustrates the amount of strikeouts a teams pitching staff accumulates per walk given.

LOB – Stranded Runners

Fielding Statistics

RA/G – Runs allowed per game

DEFEFF – Defensive Efficiency

CG – Complete Games

INN – Innings Played in the Field

CH – Defensive Chances

PO- Putouts

A - Assists

E - Errors

DP – Double Plays Turned

FLD% - Fielding Percentage

RTOT – Total Zone Total Fielding Runs Above Average

RTOT/YR - Total Zone Total Fielding Runs Above Average per 1200 Innings

RDRS – BIS Defensive Runs Saved Above Average

RDRS/YR - BIS Defensive Runs Saved Above Average per 1200 Inning

Appendix B

Table 1

Coefficients of First Attempt

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
40	(Constant)	-364.742	65.645		-5.556	0	-493.909	-235.575					
	RPERG	45.46	1.732	2.057	26.242	0	42.052	48.869	0.524	0.831	0.229	0.012	80.53
	PA	-0.254	0.012	-2.615	-21.099	0	-0.277	-0.23	0.379	-0.768	-0.184	0.005	201.192
	DOUBLES	-0.074	0.023	-0.189	-3.251	0.001	-0.119	-0.029	0.189	-0.182	-0.028	0.023	44.229
	TRIPLES	-0.135	0.047	-0.108	-2.866	0.004	-0.227	-0.042	-0.041	-0.161	-0.025	0.054	18.436
	HR	-0.184	0.068	-0.557	-2.727	0.007	-0.317	-0.051	0.404	-0.153	-0.024	0.002	546.205
	BB	-0.042	0.019	-0.253	-2.233	0.026	-0.078	-0.005	0.382	-0.126	-0.02	0.006	168.782
	BA	-605.671	237.693	-0.665	-2.548	0.011	-1073.373	-137.969	0.353	-0.143	-0.022	0.001	891.476
	OPS	342.754	124.624	1.203	2.75	0.006	97.535	587.973	0.48	0.155	0.024	0	2507.497
	HBP	-0.044	0.02	-0.054	-2.143	0.033	-0.083	-0.004	0.127	-0.121	-0.019	0.122	8.225
	SH	-0.019	0.007	-0.036	-2.585	0.01	-0.034	-0.005	-0.044	-0.145	-0.023	0.399	2.505
	LOB	0.261	0.012	1.376	21.188	0	0.237	0.285	0.215	0.77	0.185	0.018	55.293
	GF	2.154	0.375	0.574	5.738	0	1.415	2.892	-0.188	0.31	0.05	0.008	131.295
	CG	2.104	0.375	0.56	5.605	0	1.365	2.842	0.203	0.304	0.049	0.008	130.558
	tSho	0.104	0.034	0.037	3.093	0.002	0.038	0.171	0.451	0.173	0.027	0.532	1.881
	SV	0.099	0.022	0.063	4.438	0	0.055	0.143	0.653	0.245	0.039	0.377	2.656
	IP	0.803	0.033	0.99	24.442	0	0.738	0.867	0.476	0.812	0.214	0.047	21.485
	ERALL	-0.017	0.003	-0.121	-5.181	0	-0.023	-0.01	-0.584	-0.283	-0.045	0.141	7.116
	HBPALL	0.018	0.009	0.019	1.905	0.058	-0.001	0.036	-0.08	0.108	0.017	0.757	1.321
	BK	0.071	0.041	0.016	1.755	0.08	-0.009	0.151	-0.21	0.099	0.015	0.877	1.14
	WP	-0.021	0.008	-0.025	-2.601	0.01	-0.037	-0.005	-0.22	-0.146	-0.023	0.834	1.199

a Dependent Variable: W

Table 2

Correlation Coefficients of all 67 Variables

	W		W		W		W		W
W	1								
RPERG	0.524	OPS	0.48	RALL	-0.608	SO9	0.304	RdrsPERyr	0.26
PA	0.379	OPSPLUS	0.599	ERALL	-0.584	SOWITHW	0.505		
AB	0.183	TB	0.43	HRALL	-0.303	LOBALL	-0.376		
R	0.526	GDP	0.045	BBALL	-0.471	PRAPERG	-0.02		
H	0.332	HBP	0.127	IBBALL	-0.252	DefEff	0.001		
DOUBLES	0.189	SH	-0.044	SOALL	0.337	FCG	-0.016		
TRIPLES	-0.041	SF	0.289	HBPALL	-0.08	Inn	-0.002		
HR	0.404	IBB	0.296	BK	-0.21	Ch	0.027		
RBI	0.526	LOB	0.215	WP	-0.22	PO	0.013		
SB	0.094	GF	-0.188	BF	-0.473	A	0.021		
CS	-0.068	CG	0.203	ERAPLUS	0.719	E	-0.015		
BB	0.382	tSho	0.451	FIP	-0.473	DP	-0.311		
SO	-0.139	cSho	0.259	WHIP	-0.617	FldPERC	-0.047		
BA	0.353	SV	0.653	H9	-0.522	Rtot	0.095		
OBP	0.466	IP	0.476	HR9	-0.315	RtotPERyr	0.431		
SLG	0.451	HALL	-0.482	BB9	-0.494	Rdrs	0.165		

Table 3

Model Summary of Final Analysis

Model Summary[g]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.963[a]	.928	.925	3.01269	.928	371.573	11	318	.000	
2	.963[b]	.928	.926	3.00797	.000	.000	1	318	.998	
3	.963[c]	.928	.926	3.00346	.000	.042	1	319	.837	
4	.963[d]	.928	.926	2.99910	.000	.069	1	320	.793	
5	.963[e]	.928	.926	2.99518	.000	.159	1	321	.691	
6	.963[f]	.927	.926	2.99690	.000	1.370	1	322	.243	1.793

a. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, H9, OPSPLUS, RALL, WHIP, ERALL, RPERG, R

b. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, H9, OPSPLUS, RALL, WHIP, ERALL, R

c. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, OPSPLUS, RALL, WHIP, ERALL, R

d. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, OPSPLUS, RALL, WHIP, R

e. Predictors: (Constant), SOWITHW, SV, ERAPLUS, OPSPLUS, RALL, WHIP, R

f. Predictors: (Constant), SOWITHW, SV, ERAPLUS, OPSPLUS, RALL, R

g. Dependent Variable: W

Table 4

Anova of Final Analysis

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37097.683	11	3372.517	371.573	.000[b]
	Residual	2886.268	318	9.076		
	Total	39983.952	329			
2	Regression	37097.683	10	3709.768	410.016	.000[c]
	Residual	2886.268	319	9.048		
	Total	39983.952	329			
3	Regression	37097.302	9	4121.922	456.936	.000[d]
	Residual	2886.649	320	9.021		
	Total	39983.952	329			
4	Regression	37096.677	8	4637.085	515.540	.000[e]
	Residual	2887.275	321	8.995		
	Total	39983.952	329			
5	Regression	37095.250	7	5299.321	590.709	.000[f]
	Residual	2888.702	322	8.971		
	Total	39983.952	329			
6	Regression	37082.959	6	6180.493	688.144	.000[g]
	Residual	2900.993	323	8.981		
	Total	39983.952	329			

a. Dependent Variable: W

b. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, H9, OPSPLUS, RALL, WHIP, ERALL, RPERG, R

c. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, H9, OPSPLUS, RALL, WHIP, ERALL, R

d. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, OPSPLUS, RALL, WHIP, ERALL, R

e. Predictors: (Constant), SOWITHW, RBI, SV, ERAPLUS, OPSPLUS, RALL, WHIP, R

f. Predictors: (Constant), SOWITHW, SV, ERAPLUS, OPSPLUS, RALL, WHIP, R

g. Predictors: (Constant), SOWITHW, SV, ERAPLUS, OPSPLUS, RALL, R

Table 5

Final Model and Variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
6	(Constant)	28.723	8.758		3.28	0.001	11.493	45.952					
	R	0.076	0.006	0.557	13.306	0	0.065	0.087	0.526	0.595	0.199	0.128	7.792
	OPSPLUS	0.148	0.049	0.104	3.04	0.003	0.052	0.243	0.599	0.167	0.046	0.193	5.177
	SV	0.437	0.028	0.279	15.474	0	0.381	0.492	0.653	0.652	0.232	0.691	1.446
	RALL	-0.065	0.006	-0.502	-11.185	0	-0.077	-0.054	-0.608	-0.528	-0.168	0.112	8.957
	ERAPLUS	0.09	0.044	0.08	2.053	0.041	0.004	0.175	0.719	0.113	0.031	0.147	6.822
	SOWITHW	1.537	0.633	0.055	2.427	0.016	0.291	2.784	0.505	0.134	0.036	0.436	2.293

a. Dependent Variable: W

Table 6

Residual Statistics of Final Model

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	52.7154	104.9080	80.9879	10.61669	330
Std. Predicted Value	-2.663	2.253	.000	1.000	330
Standard Error of Predicted Value	.229	.716	.425	.098	330
Adjusted Predicted Value	52.5922	104.9055	80.9866	10.61600	330
Residual	-7.16862	8.72992	.00000	2.96945	330
Std. Residual	-2.392	2.913	.000	.991	330
Stud. Residual	-2.429	2.927	.000	1.002	330
Deleted Residual	-7.38967	8.81398	.00123	3.03444	330
Stud. Deleted Residual	-2.447	2.962	.000	1.005	330
Mahal. Distance	.925	17.777	5.982	3.257	330
Cook's Distance	.000	.033	.003	.005	330
Centered Leverage Value	.003	.054	.018	.010	330

a. Dependent Variable: W

Table 7

Variance Explained Using Pythagorean Formula Variables

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.608[a]	.369	.367	8.76964	.369	191.902	1	328	.000	
2	.932[b]	.868	.867	4.02117	.499	1233.027	1	327	.000	2.037

a. Predictors: (Constant), RALL

b. Predictors: (Constant), RALL, R

c. Dependent Variable: W

Table 8

Comparing SEE of Pythagorean Formula and Regression Formula

	SEE (Standard Error of the Estimate)
Pythagorean Formula	4.03
Regression Formula	2.99

Figure 1

Histogram of Final Model

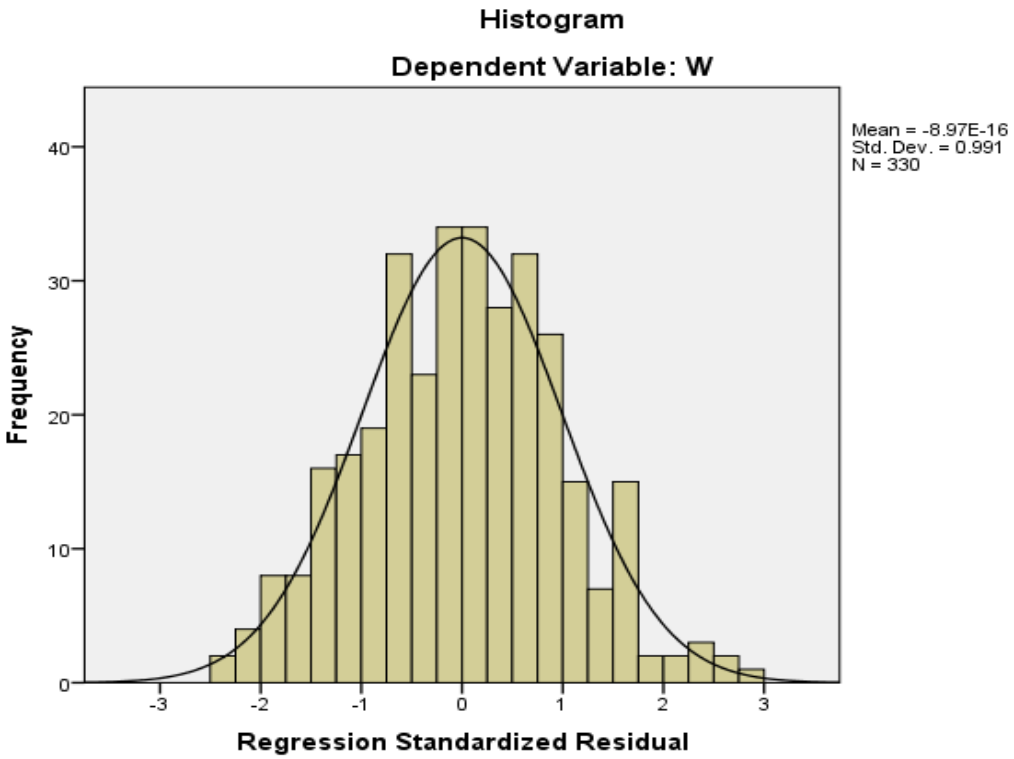

Figure 2

Scatterplot of Final Model