THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

USING CONVENTIONAL AND SABERMETRIC BASEBALL STATISTICS FOR PREDICTING MAJOR LEAGUE BASEBALL WIN PERCENTAGE

VICTORIA DECESARE
SPRING 2016

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Science
with honors in Statistics

Reviewed and approved* by the following:

Andrew Wiesner
Lecturer of Statistics
Thesis Supervisor

David Hunter
Department Head of Statistics
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Major League Baseball is dominated by statistical analysis; one cannot watch a baseball game on the television without hearing and seeing a plethora of statistics such as batting average, runs batted in, earned run average, and the list goes on. In addition to these popular stats that most people are familiar with, there are several, more complex baseball statistics – known as "sabermetric" statistics – that have been developed over the past few decades that seek to evaluate players and the game more scientifically and comprehensively. However, with all of these stats available, it is easy to get caught up in the data and overlook the main goal of MLB teams: to win games. With this in mind, the goal of this research is to explore some of the numerous baseball statistics available, both the traditional and modern ones, and observe which ones are truly the best at predicting wins. Encompassing this, is it better to use the more complex methods in analyzing how teams win, or does it hold true that "less is more"? This research seeks to answer these questions and to provide a unique perspective for fans and managers alike when trying to make use of the ever-growing world of baseball data.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

- To my dad – softball coach for six years but life coach forever. Thank you for being one of the main reasons why I fell in love with the game of baseball (even when you were yelling my nicknames out on the softball field).

- To my mom, for always being my #1 fan. The best compliment I have ever received is that I am just like you. Thank you for all you do for our whole family and me.

- To my sister – occasional teammate, but more importantly my best friend. Thank you for always supporting me.

- To Dr. Wiesner, not only for guiding me throughout the whole thesis process, but also for showing me how statistics can be applied to my love of sports, especially baseball.

- To Dr. Hunter, for helping me to push the boundaries of this thesis until the very end.

- To my friends, to whom I would just have to say the word "Thesis" and they understood.

- To my favorite MLB player Derek Jeter, who said: "There may be people that have more talent than you, but there's no excuse for anyone to work harder than you do."

Chapter 1

Introduction

For more than 150 years, statistics have been used to analyze the game of baseball and provide quantifiable insight into the success and value of individual players and teams. With Henry Chadwick's creation of the box score in the mid-1800s, people were able to see numerical representations of some of the most important defensive and offensive events in a game of baseball, such as runs, outs, strike outs, and times left on bases (Schwarz). Soon enough, other baseball statistics, such as batting average (BA) and Earned Run Average (ERA) were developed based on Chadwick's initial statistics and have remained very popular and widely used through the modern baseball era. However, over the past 40 years, a revolution in baseball statistics has taken place that has greatly altered the way individual players and teams are evaluated. This "sabermetric" revolution, spearheaded by Bill James, has led to the creation of a plethora of additional metrics. New measures such as Weighed On-Base Average (wOBA) and Wins Above Replacement (WAR) offer an even more detailed and complex look into various aspects of the game compared to "conventional" statistics. Presently, there are over 100 statistics used in baseball analysis, and many baseball experts have their reasons why some statistics may be better than others.

Essentially, the ultimate goal for every team in Major League Baseball (MLB) is to win the World Series. The most obvious way to achieve this is by winning the most games possible. Simply put, a team must score more runs and allow fewer runs than their opponent in order to win a game. Ultimately, a team's success at the end of the season is determined by their win-loss

record. This makes understanding tactics which contribute to wins, or at least runs scored or prevented, extremely important.

The goals of this paper are to not only explore which of the many available baseball statistical measures are the best at predicting an MLB team's win-loss percentage, but also discover the simplest yet most effective method or model for predicting winning percentage. In comparison with conventional statistics, are the newer, more complex stats more highly correlated with win percentage? Or, do the conventional stats perform just as well as the sabermetric stats? In addition, while it may be discovered that a model with multiple sabermetric stats is a great predictor of win percentage, can a simpler model (or, simpler stats) perform just as similarly? In order to answer these questions, this paper will explore the background of various baseball statistics and formulas that have been shown to relate greatly to winning games through contributing to scoring runs or preventing runs. The origin of sabermetrics and the baseball analytics revolution will also be investigated more in depth. Season-ending team averages for various statistics were collected dating back to 1998, the year of MLB's expansion to its current thirty teams. Multiple regression models were then developed using select measures and then analyzed in Minitab. Lastly, this paper will conclude with the findings and considerations for future research. Hopefully, this paper can help individuals take on a new and potentially simpler yet more effective approach when it comes to analyzing which factors are truly the most important when quantifying a team's success in the MLB.

Chapter 2

Literature Review

History of Baseball Statistics

Early Baseball Statistics

Baseball and statistics have coexisted for almost as long as the game itself. Baseball statistics has its origin in the box score, the first of which appeared in the *New York Morning News* on October 25, 1845 (Panas). This first box score, representing a game between the New York Ball Club and the Brooklyn Club, consisted solely of hands out (today, just "outs") and runs, reflective of baseball's early association with cricket (Panas). Within the next two decades, box scores began to include additional statistics (detailing various aspects of a player's fielding, batting, and baserunning performance). Soon enough, cumulative statistics for teams' entire seasons became available. This was due largely in part to Henry Chadwick, baseball's first statistician and the "father of baseball", according to numerous baseball historians (Panas). Chadwick designed his own box scores and published these and his other statistical assessments in prominent publications such as the *New York Clipper* and the *Beadle Guide* (Panas). However, one of Chadwick's main goals was to formulate the best set of objective statistical measures that would indicate which players contributed most and least to a team's wins (Schwarz). In 1861, Chadwick presented the total games, outs, runs, home runs, and strike outs for players on five different teams; by 1867, Chadwick had additionally introduced runs per game, outs per game,

hits per game, and total bases per game (Schwarz). The best hitters, according to Chadwick, were those with the greatest runs per game (Schwarz).

Batting and fielding were both equally considered integral parts of the game in the late 1800s. In 1867, Chadwick made the distinction between earned runs and unearned runs in an attempt to evaluate batting and fielding, despite modern-day Earned Run Average (ERA) purely evaluating pitching. However, soon, fielding ability was measured by successful plays divided by total chances, known today as fielding percentage (Fld%) [Schwarz]. In regards to offense, in 1872, Chadwick adopted a new batting statistic from a fan named H.A. Dobson that would become a staple in the game for years to come: batting average (BA). BA, which simply measures hits per times at bat, was deemed by Chadwick as the best indicator of offensive performance (Panas). When the National League formed in 1876, BA and Fld% were the official statistics representing batting and fielding performance respectively (Panas). Pitching, while initially seen as having no impact on a team's ability to win, began to be viewed much more prominently in the 1870s; by 1876, there were eleven official statistics to measure pitching performance including earned runs per game and hits allowed (Panas). Many of these early baseball statistics are still in use to this day, despite the origin of many newer, more complex and comprehensive statistics in the following century.

The Rise of Sabermetrics

During the early twentieth century, many began to question the effectiveness of the early baseball statistics being used to measure batting, fielding, and pitching performance. Ferdinand Cole Lane was one of these individuals who specifically spoke out against BA, discussing the

statistic's disregard for walks and extra base hits, and Fld%. Lane was one of the first individuals to discuss the importance of ballpark factors in addition to weighing the various types of hits (singles, doubles, triples, and home runs) differently based on their respective contribution to scoring runs, forming the basis of the linear weights system that would officially be introduced in 1984 by John Thorn and Pete Palmer in *The Hidden Game of Baseball* (Panas).

Baseball statistical analysis reached another milestone in 1947, when Branch Rickey, the innovative General Manager of the Brooklyn Dodgers at the time, hired Allan Roth as the team's analyst, making Roth MLB's first full-time team statistician (Panas). Both Roth and Rickey were early advocates of what we now call on-base percentage (OBP) and isolated power (ISO) and how these measures were much more effective than BA in evaluating batting ability. In addition, Roth was one of the first individuals to combine multiple player skills into a single measure, although this would not become common practice until many years later.

In 1964, retired engineer Earnshaw Cook published his studies on baseball analysis in his book *Percentage Baseball*. While criticized by baseball fans and statisticians alike for its esoteric language, *Percentage Baseball* was the first book on baseball statistical research to garner national media attention and was highly influential on future sabermetricians. A key component of Cook's studies though was his Scoring Index, a statistic which showed the probability of scoring a run as essentially the product of OBP and ISO (again, concepts which wouldn't become well-known until years later) [Schwarz].

Bill James and the Sabermetric Revolution

Sabermetrics grew in popularity in the 1970s thanks to avid baseball fan and writer Bill James. From 1977 – 1988, James annually published his *Baseball Abstracts* that sought to answer various questions about baseball through statistical analysis (Panas). These *Baseball Abstracts* were hugely successful and attracted a vast audience to the world of advanced baseball statistical analysis (Panas). In 1980, James put a name to this new advanced baseball analysis and coined the term "sabermetrics", derived from SABR, an acronym for the Society for American Baseball Research (Schwarz). He specifically defineD sabermetrics as "the search for objective knowledge about baseball" (Birnbaum). Understandably, Bill James is considered to be the "father of sabermetrics" and is responsible for many of the newer baseball statistics such as the Pythagorean Expectation (PE) and Defensive Efficiency Ratio (DER) [Panas].

Sabermetrics caught the attention of not only journalists, fans, and players, but also managers. Earl Weaver, manager of the Baltimore Orioles from 1968-1982 and 1985-1986, was one of the first managers to rely heavily on statistics for an objective perspective on players and in-game situations such as batter and pitcher matchups (Panas). The Orioles were able to win four pennants during Weaver's tenure, and from this point on, front offices began to greatly value sabermetrics in their decision making. Most prominently, Billy Beane, general manager of the Oakland Athletics from 1998-2015, utilized sabermetrics in his approach to building a successful baseball team on a low budget – a strategy popularized in the 2003 book *Moneyball* (Panas). Beane highly valued on-base percentage (OBP) and slugging percentage (SLG) over traditionally-valued measures such as BA and RBIs, and his approach provided a new perspective on what factors are truly the most important when building a winning team.

Winning Games

The ultimate goal for any Major League Baseball team is to win the World Series; to potentially achieve this, first it is necessary for a team to win enough games to make the postseason. Therefore, it is understandable why a team would want to maximize their winning potential. Teams win games by outscoring their opponents; in baseball specifically, a team must score more runs compared to the runs they allow to win a game, and then do this over the course of an entire season to win the most games possible. According to Lee Panas, a team's run differential – that is, runs scored minus runs allowed - is very closely related to a team's wins, and teams with greater run differentials tend to have more wins. In support of the importance of the relationship between run differential and wins, Pete Palmer went even further and estimated that a run differential of ten (scoring ten more runs than allowed) is equivalent to one win, and vice versa (Panas).

Pythagorean Expectation

One of the most popular models demonstrating the relationship between runs scored, runs allowed, and winning is Bill James' Pythagorean Expectation. Introduced in his 1980 *Baseball Abstract*, the Pythagorean Expectation predicts a team's winning percentage based solely on runs scored and runs allowed. The formula is as follows:

$$\text{Pythagorean Expectation (PE): Winning Percentage} = RS^x / (RS^x + RA^x)$$

The exponent x can vary; however, when Bill James first introduced the formula, he used $x = 2$ (the equation then had three squares, reminding James of trigonometry's Pythagorean Theorem, hence the name of his formula) [Costa]. While relatively simple, James' formula

correlates fairly accurately with winning percentage, although an error of approximately three games off has been observed. The Pythagorean Expectation does have its flaws, however. Using this formula, it would be mathematically impossible for a team that has a positive run differential to have a winning percentage below 0.500, and a team that has a negative run differential to have a winning percentage above 0.500, both of which could happen in real life (Costa).

Offensive Measures

As discussed previously, a win is the result of more scored runs over allowed runs in a game. Scoring runs is the result of offensive (batting) events, while allowing runs is the result of defensive (pitching and fielding) events. Therefore, it is necessary to evaluate individual offensive statistics at the team level to gain a better understanding of how runs are scored.

The following offensive statistics have been shown to be related to scoring runs and thus to winning games:

Conventional Statistics:

- Batting Average (BA)
- Runs Batted in (RBIs)
- On-Base Percentage (OBP)
- Slugging Percentage (SLG)
- On-Base Plus Slugging (OPS)

Sabermetric Statistics:

- Isolated Power (ISO)
- Weighted On-Base Average (wOBA)
- Weighted Runs Above Average (wRAA)
- Weighted Runs Created (wRC)
- Weighted Runs Created Plus (wRC+)

Batting Average

BA is the most widely used statistic to measure a batter's performance, and the one with which individuals are the most familiar (Gerard). It is calculated as:

$$BA = H / AB$$

According to Panas, one way that BA is useful is because it measures a batter's ability to make contact with the ball and hit it out of a fielder's reach; in addition, teams with higher BAs tend to score a lot of runs. However, BA fails to take non-events, such as walks or being hit by a pitch, into account. These events have the same result as a single but BA does not recognize this. Furthermore, BA counts all hits equally, but in terms of scoring more runs, extra-base hits tend to be more valuable than singles. Many in the baseball and statistical community, such as Billy Beane and Bill James, have publicly deemed BA as an ineffective statistic, despite its ongoing use in the present baseball era.

Runs Batted In

In addition to BA, RBI is one of the most famous and commonly referenced statistics in baseball. RBI measures the number of runners who score due to a hit, walk, sacrifice, or fielder's choice (Weinberg). Although the MLB has recognized RBI as an official statistic since 1920, many in the baseball community heavily criticize RBI due to its context-dependent nature. First, a player's RBI is heavily influenced by his order in the batting lineup; for instance, a leadoff hitter has significantly fewer RBI opportunities compared to a hitter deeper in the lineup. Second, a player can have a high RBI total solely due to favorable circumstances, such as a lot of runners being on base when they were batting. While the ability to hit with runners on base is

without a doubt a valuable skill for a batter to have, many agree that RBI fails to accurately measure that skill due to the unequal opportunities batters have to gather RBI (Weinberg). Despite this, RBI is still used today to offensively evaluate players; in fact, BA, RBI, and HR are the three statistics that make up MLB's "Triple Crown" in batting (Keri).

On-Base Percentage

OBP measures the frequency of a batter safely reaching base. OBP, like SLG (discussed below), attempts to account for some of the drawbacks of BA. It is calculated as:

$$OBP = (H + BB + HBP) / (AB + BB + HBP + SF)$$

Perhaps the most well-known proponent of OBP was Billy Beane, whose statistically-focused management style of the Oakland A's was popularized through *Moneyball*. Beane believed that OBP and SLG represented the two most important parts of offensive performance: getting on base and hitting for power (Keri). Beane used OBP, undervalued at the time, to build his team, and the A's were very successful during his time as General Manager. Jonah Keri et. al. in *Baseball Between the Numbers* summarizes the effectiveness of the OBP statistic:

> "Among traditional offensive statistics, it's the most important; the higher a player's OBP, the less often he's cost his team an out at the plate and the more he's prolonged innings and created more runs, which leads to more wins."

Slugging Percentage

Quite simply, SLG measures a batter's ability to hit for power, and is calculated as:

$$SLG = (1B + 2 \times 2B + 3 \times 3B + 4 \times HR) / AB$$

As mentioned previously, SLG is the second component of what many consider to be ideal offensive performance. It also tries to account for BA's inability to distinguish between different types of hits by giving different weights to singles, doubles, triples, and home runs. However, there has been much debate over the choice of weights for the various hits; for instance, is a homerun exactly four times more valuable than a single (Gerard)? Like BA, SLG also does not include walks in its calculation.

Isolated Power

ISO measures a hitter's raw power and his ability to hit for extra bases. It is calculated as:

$$ISO = (2B + 2 \times 3B + 3 \times HR)/AB \quad \text{or} \quad ISO = SLG - BA$$

Developed by Branch Rickey and Allan Roth in the 1950s and reintroduced by Bill James in the 1970s, ISO, like SLG, stresses the importance of *extra* base hits over singles when it comes to scoring runs (Panas). However, since SLG is in ISO's formula, ISO faces the same issue of the weight choices for the different types of hits.

On-Base Plus Slugging

Like its naming suggests, OPS is simply the sum of OBP and SLG. Introduced by Pete Palmer in 1984, OPS combines the two key elements of offense – getting on base and power hitting – into one statistic that can more effectively describe a team's ability to score runs. OPS has many devout advocates, including Peter Gammons, a famous MLB analyst and baseball writer, and other sports journalists; in addition, many MLB stadiums now display OPS on their scoreboards (Schwarz). However, one of the downsides of OPS is the equal weighting of OBP

and SLG in the formula; in relation to scoring runs, OBP is undervalued compared to SLG (Panas). Paul DePodesta, analytic assistant to Billy Beane during the *Moneyball*-highlighted era, even went as far to say that OBP was worth three times as much as SLG when predicting run production (Lewis). Others have analyzed OBP to be worth approximately no more than 1.5 times SLG, but regardless, it is general expert consensus that OBP should be weighted more than SLG in the OPS formula (Lewis).

Weighted On-Base Average

Like OPS, wOBA measures total offensive contribution per player. However, unlike OPS, wOBA assigns more appropriate linear weights to the various offensive events based on each event's potential run value. wOBA, wRAA, and wRC were all introduced by Tom Tango, a sabermetric analyst, in his 2006 book *The Book: Playing the Percentages in Baseball* (Panas). wOBA is calculated as[1]:

$$\text{wOBA} = [0.71 \times (\text{BB} - \text{IBB}) + 0.74 \times \text{HBP} + 0.89 \times 1\text{B} + 1.26 \times 2\text{B} + 1.58 \times 3\text{B} + 2.02 \times \text{HR} + 0.24 \times \text{SB} - 0.51 \times \text{CS}] / (\text{PA} - \text{IBB})$$

The individuals at FanGraphs.com, a baseball statistics and analysis site, are huge proponents of the wOBA statistic. Neil Weinberg, site educator of FanGraphs and maintainer of the FanGraphs statistical library, noted that "players should get credit for the degree to which their actions lead to run scoring, and wOBA offers a much more complete accounting of that than something like RBI, BA, or OPS" (Weinberg).

[1] Weights change every year based on the run environment

Weighted Runs Above Average

wRAA is a metric that measures the number of runs a batter contributes to his team beyond what an average player would have contributed in his place (Panas). Essentially, wRAA is wOBA but converted into runs, which are direct components of wins. wRAA is calculated as[2]:

$$wRAA = (wOBA - MLB\ wOBA) / (1.21 \times PA)$$

wRAA also adjusts for league and ballpark differences.

Weighted Runs Created

In 1982, Bill James created the Runs Created (RC) statistic, which "estimates a player's offensive contribution in terms of total runs" ("Weighted Runs Created Plus"). wRC measures the same thing as RC but is based on linear weights and thus is a more accurate measure. It is calculated as[3]:

$$wRC = wRAA + (MLB\ runs / PA)$$

[2] The "1.21" value changes every year but is usually approximately 1.2
[3] wRC is equivalent to wRAA with MLB run average scaled to zero

Weighted Runs Created Plus

wRC+ is the same measure as wRC, except it compares wRC with the league average and adjusts for league and ballpark differences. It is calculated as:

$$wRC+ = 100 \times (wRC/ MLB\ wRC)$$

wRC+ is largely considered to be the best and most comprehensive offensive metric; according to MLB.com, "wRC+ quantifies the most important part of a batter's job – creating runs – and normalizes it… wRC+ is about as good as it gets when it comes to assessing hitters in a vacuum…" ("Weighted Runs Created Plus").

Defensive Measures

The use of metrics to analyze a team defensively has been much more recent and complex compared to offensive analysis; whereas an offensive event can be attributed to a single batter, there are many factors (and players) which collectively contribute to how runs are prevented. Still, there is a plethora of measures, both old and new alike, that seek to describe some aspect of defensive performance, and many of those are detailed below.

While many individuals separate pitching statistics from fielding statistics, for the purposes of this research, both are collectively classified as "defensive statistics" since both are integral for decreasing runs allowed. We will focus on the following defensive statistics:

Conventional Statistics:

Earned Run Average (ERA)
Strikeouts per Nine Innings (K/9)
Walks per Nine Innings (BB/9)
Fielding Percentage (Fld%)

Sabermetric Statistics:

Defensive Efficiency Ratio (DER)
Batting Average on Balls in Play (BABIP)
Fielding Independent Pitching (FIP)
Left-on-base Percentage (LOB%)
Walks plus Hits per Inning Pitched (WHIP)

Earned Run Average

According to David Gerard in *Baseball GPA: A New Statistical Approach to Performance and Strategy*, ERA is the most widely used statistic to measure pitcher performance. It is defined as the average number of runs charged to a pitcher per nine innings pitched and is calculated as[4]:

$$ERA = (Earned\ Runs \times 9) / IP$$

One of the reasons why ERA was created was to isolate a pitcher's ability to prevent runs from his teammates; however, according to Keri, "tracking ERA lessens the problem of teammate reliance but does not eliminate it." Despite some of the backlash against traditional ERA, many still cite the statistic as a helpful tool for evaluating pitchers. David Gerard describes why ERA has been such a popular statistic in the baseball community:

[4] Earned runs are runs resulting from pitching and not due to fielding errors

> "ERA meets many of the criteria we would expect in a useful baseball statistic. It has the advantage of being measured over many innings and reflects a pitcher's ability to prevent hitters from being productive, runners from advancing, and is easy for the casual fan to understand."

Since ERA's creation, there have been many newer statistics, such as Defense-Independent ERA and ERA+, that are improvements upon traditional ERA. However, the traditional ERA's strength is in its simplicity and its ability to emphasize the main goal of the defense: to allow the least amount of runs possible.

Strikeouts per Nine Innings

Strikeouts are very important for a pitcher and his defense; however, it is important to analyze strikeouts in terms of strikeout *rate* because a pitcher who pitches more innings is likely to have more strikeouts than a pitcher who pitches fewer innings solely because he has more opportunities to do so (Albert). K/9 is standardized, scaling a pitcher's strikeout rate to nine innings, and is calculated as:

$$K/9 = (K \times 9) / IP$$

K/9 is a helpful indicator of runs prevented; if a batter strikes out, he does not get on base and current baserunners do not advance as well, therefore not allowing runs (Weinberg).

Walks per Nine Innings

Like K/9, BB/9 is a standardized statistic and is calculated as:

$$BB/9 = (BB \times 9) / IP$$

BB/9 is essentially the opposite of K/9; when a batter gets a walk, he gets on base and increases his team's chances of scoring (Weinberg). Ideally, there should be a wide spread between a pitcher's K/9 and BB/9 rates (Albert).

Fielding Percentage

As discussed previously, Fld%, created in 1876, is the oldest officially recognized fielding statistic (Schwarz). Essentially, a team's Fld% is the percentage of plays successfully converted without error and is calculated as:

$$Fld\% = (PO + A) / (PO + A + E)$$

While Fld% is still referenced in modern analysis, it is widespread consensus throughout the baseball community that Fld% is not as useful as was once thought (Schwarz). One of the fundamental problems with Fld% is that it is rooted in the idea that all fielders have the same opportunities to make plays, which is not true. Second, it does not account for a fielder's range; for instance, a shortstop has a much greater range compared to a first baseman, and therefore is more likely to have more difficult fielding opportunities (Schwarz). Bill James proposed a new metric, called Range Factor (RF), that accounted for this shortcoming and measured instead plays made per game (Schwarz). However, since Fld% is a readily available statistic online and since it has been so prevalently used over the past 140 years, it is applicable in this study.

Defensive Efficiency Ratio

Created in 1978 by Bill James, DER is the percentage of batted balls in play (not including HR) which are converted to outs by the fielders. Essentially, DER is a more efficient

version of Fld% since it is quantifiably based in the idea of range instead of errors (Panas). It is calculated as:

$$DER = (BFP – H – K – BB – HBP – 0.6 \times E) / (BFP – HR – K – BB – HBP)$$

Like Fld%, a strong DER means that a team is effective at converting balls to outs, which helps prevent runs. DER does have its flaws – for instance, DER has no bearing on if bases are loaded and a batter gets a home run – but it is undoubtedly a better measure of defensive prowess compared to Fld% (Panas).

Batting Average on Balls in Play

BABIP measures how often batted balls fall for hits (excluding home runs) – essentially the opposite of DER. BABIP can be observed from both the pitching/defense and batting perspectives (in this study, it will be analyzed from the defensive perspective); from a pitcher's perspective, it is a "pitcher's BA allowed" (Cockcroft). In both cases it is calculated the same[5]:

$$BABIP = (H – HR) / (AB – K – HR + SF)$$

BABIP has become a very popular statistic in the sabermetric community during the past decade, but it is still largely misinterpreted (Keri). In the early 2000s, baseball researcher Voros McCracken was the first to suggest that pitchers actually had very little influence on their BABIP – that is, pitchers have minimal control over whether batted balls become hits or outs (McCracken's research soon became formally known as the Defense Independent Pitching Statistics, or DIPS, theory) [Keri]. It is therefore important to understand that a pitcher's BABIP is greatly affected by his defense, pure luck, and changes in talent level (Weinberg). However, pitching BABIP still provides insight into hit frequency, a vital component of allowing runs.

[5] BABIP can be approximated by 1 - DER

Fielding Independent Pitching

Inspired by McCracken's research and the DIPS theory, the FIP statistic "estimates a pitcher's run prevention independent of the performance of their defense" (Weinberg). Developed by Tom Tango, it is designed to look like ERA and is calculated as[6]:

$$FIP = [HR \times 13 + (BB + HBP - IBB) \times 3 - K \times 2] / IP + FIP\ constant$$

Based on BABIP's shortcomings, FIP only includes those items which a pitcher directly controls (K, BB, HR, and HBP) and does not include events potentially influenced by the defense (i.e. hits) [Panas]. According to *The Hardball Times*, "FIP helps you understand how well a pitcher pitched, regardless of how well his fielders fielded" (Judge). Essentially, FIP is another statistic that provides greater insight into the events that contribute to runs allowed.

Left-on-base Percentage

LOB% is similar to the Strand Rate statistic, but while strand rate is the percentage of baserunners who fail to score, LOB% is the number of baserunners who did not score divided by the total number of baserunners (excluding those who scored home runs) [Panas]. LOB% is calculated as:

$$LOB\% = [H + BB + HBP - RS] / [H + BB + HBP - (1.4 \times HR)]$$

While extremely high or low LOB% values may be the result of a matter of luck, according to Dave Studeman, owner of baseball statistical analysis website *The Hardball Times*, LOB% is

[6] The FIP constant varies each year

still a powerful statistic in measuring the degree to which pitchers can keep baserunners from scoring (Studeman).

Walks plus Hits per Inning Pitched

WHIP is a measurement of the total number of baserunners a pitcher has allowed per inning pitched (conveniently, its calculation is also its name). Invented in 1979 by Daniel Okrent, writer and inventor of Fantasy Baseball, WHIP is one of the sabermetric measures that has made its way into the mainstream, due mainly in part to its ease in calculation and understanding (Di Fino). Compared to other pitching metrics, such as a pitcher's wins, WHIP is not context-specific and effectively shows a pitcher's ability to keep runners off the bases. If runners are kept off bases, then runs cannot be scored, so WHIP is helpful to understand when trying to decrease runs allowed.

Total Contribution

When baseball statistical analysis first took form, it was common practice to analyze players and events in isolation. The sabermetric revolution helped bring about statistics which describe multiple facets of players and the game itself, since there are multiple factors which contribute to how teams win games. WAR, discussed below, is one of the most heralded Total Contribution statistics today.

Wins Above Replacement

WAR is an all-inclusive statistic that attempts to summarize a player's total contribution to his team (Weinberg). More specifically, it is the number of additional wins a player would contribute to his team if he was substituted by a replacement-level player (for instance, a freely available minor league player) [Weinberg]. WAR incorporates both offensive and defensive elements for position players, and so it was not discussed in previous sections.

WAR data can be found on both FanGraphs.com (fWAR) and Baseball-Reference.com (bWAR or rWAR), but both sources calculate WAR slightly differently (we will use the fWAR definition throughout this paper, although just indicated as "WAR"). WAR involves many elements from other metrics in its calculation, but its simplest calculation is as follows[7]:

$$\text{fWAR} = \text{wRAA} + \text{UZR} + \text{Position} + 20/600 \times \text{PA}$$

Although not recognized as an official statistic by the MLB, WAR is one of the few but most popular statistics in the sabermetric community today that quantifies a player's skill in terms of wins and not just runs. In regards to WAR's ability to predict wins, Dave Cameron, managing editor and senior writer of FanGraphs, has stated that "WAR does an impressive job of projecting wins and losses" (Weinberg). One of the additional benefits of WAR is that it is park-

[7a] The UZR calculation is extremely thorough and therefore will not be included in this thesis. What is important to know is that UZR (Ultimate Zone Rating) is the defensive component of WAR and seeks to put run value to defensive events.

[7b] The "Position" component assigns different values for each player based on his position:

Catcher: +12.5 per 162 games
1B: -12.5
2B: +2.5
3B: +2.5
SS: +7.5
LF: -7.5
CF: +2.5
RF: -7.5
DH: -17.5

[7c] The addition of the "20/600" value represents that a typical replacement player is considered to be 20 runs below average over a full season, assuming 600 plate appearances

adjusted, league-adjusted, and context neutral, making it efficient for comparison purposes. Despite its inherent complexity, WAR is definitely one of the most important statistics to understand for sabermetric analysis.

Chapter 3

Methodology

Statistical Methods

The main purpose of this thesis is to analyze the relationship between a variety of baseball statistics and winning percentage; in other words, we want to predict winning percentage from these metrics and see which metrics actually have the greatest impact on winning percentage, since winning games will help a team get to the playoffs and potentially the World Series. Because of this purpose, regression analysis is the most applicable statistical method for this study (discussed more in depth below).

One of the underlying goals of regression analysis is to fit a parsimonious model that explains variation in the response with a small set of predictors (Kutner). Parsimony indicates that the fitted model is not only powerful, but also simple (i.e. having low complexity). This is extremely important for our study because of how these results can be applied in the real world; for instance, while a complex regression model might explain a lot about winning percentage, a general manager of an MLB team would much more prefer to utilize a model that is simpler (and easier to understand) but just as effective when deciding how to build his team and incorporating run scoring and run preventing strategies.

With our overall goal in mind, throughout our data analysis we will:

1. Observe which statistics are best correlated with winning percentage

2. Discover which statistics are best correlated with scoring runs, the offensive component of winning games
3. Determine if the sabermetric statistics are completely dominant over the conventional statistics in terms of association with winning percentage and scoring runs
4. Develop parsimonious models for winning percentage and runs scored

In order to accomplish these tasks, we will explore the statistical concepts and methods behind correlation, the multiple linear regression model itself, model building, model selection, and the issue of multicollinearity.

Correlation

Correlation measures the strength of the relationship between two variables. Specifically, the Pearson correlation coefficient, denoted as R, is an indicator of the strength and direction of the linear relationship between two continuous variables (Kutner). The formula for R is as follows:

$$R = [\sum^{n}_{i=1}(x_i - \bar{x})(y_i - \bar{y})] / [\sqrt{\sum^{n}_{i=1}(x_i - \bar{x})^2 \sum^{n}_{i=1}(y_i - \bar{y})^2}]$$

where:

x_i = value of the predictor variable in the *i*th case
$\bar{x}$ = mean value of the predictor variable
y_i = value of the response variable in the *i*th case
$\bar{y}$ = mean value of the response variable

In terms of strength, the coefficient can range from -1 to 1; the larger the absolute value of the coefficient, the stronger the relationship is between the two variables (so, a correlation coefficient of -1 or 1 denotes a perfect linear relationship and a coefficient of 0 indicates that no

linear relationship exists) [Kutner]. In terms of direction, a positive coefficient represents a positive relationship between the variables, and vice versa. Scatterplots are useful tools for visualizing this strength and direction. Lastly, for simple linear regression, the Pearson correlation coefficient is the square root of the coefficient of determination, or R^2 (to be discussed later in this section):

$$R = \pm \sqrt{R^2}$$

In relation to our data set, the Pearson correlation coefficients will be used to measure the relationship between the each of the various baseball statistics and W-L% and RS respectively. For the purposes of our data set, the following ranges will be used as cutoffs for the degrees of correlation (Hopkins):

Table 1: Correlation Coefficient Ranges

Correlation Coefficient Range (in terms of absolute value)	**Degree of Correlation**
0.9 – 1.0	Nearly perfectly correlated
0.7 – 0.9	Very highly correlated
0.5 – 0.7	Highly correlated
0.3 – 0.5	Moderately correlated

One important caveat of correlation is that correlation does not imply causation; in other words, just because two variables are highly associated with each other does not mean that one directly causes the other. While causation *may* exist, correlation coefficients themselves do not provide any information about potential causation.

The Multiple Linear Regression Model

Because we are using multiple baseball metrics to predict win percentage (and runs scored), in this study multiple regression models will be used. Multiple linear regression models the relationship between multiple independent variables (explanatory variables, or predictors) and a dependent variable (the response). Simple linear regression, on the other hand, uses only one independent variable to explain the variation in the response.

The general multiple linear regression model is as follows:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i \qquad i = 1,2\ldots n$$

where:

y_i = the response variable
$\beta_0, \beta_1 \ldots \beta_{p-1}$ = the unknown parameters
$x_{i,1}, x_{i,2} \ldots x_{i,p-1}$ = the predictor variables
ϵ_i = the random error assumed to have a normal distribution, zero mean, and constant variance

The beta (β) values are estimated by the model and represent how strongly their corresponding predictor variables predict the response; the higher the β value, the greater that variable's impact on the response, and vice versa (Kutner). For our study, the β values will be used to analyze the impact of each predictor on W-L% and RS respectively.

In this study, two distinct regression models will be developed. The first model will use win percentage (W-L%) as the response. Win percentage is used as the response instead of just simply wins because win percentages continuously fall between 0.2 and 0.8, making linear regression applicable (Long). The second model is a sub-model of the win percentage model. Based on the offensive component of the PE, an offensive model was developed using Runs Scored (RS) as the respective response. The predictors for the offensive model were the statistics (both conventional and sabermetric measures) that relate to batting.

Table 2: List of Response and Predictors for Distinct Regression Models

	First Model	**Second Model**
Response	W-L%	RS
Predictors	X_1: PE X_2: RS X_3: RA X_4: BA X_5: RBI X_6: OBP X_7: SLG X_8: ISO X_9: OPS X_{10}: wOBA X_{11}: wRAA X_{12}: wRC X_{13}: wRC+ X_{14}: ERA X_{15}: K/9 X_{16}: BB/9 X_{17}: BABIP X_{18}: LOB% X_{19}: WHIP X_{20}: FIP X_{21}: Fld% X_{22}: DER X_{23}: WAR	X_1: BA X_2: OBP X_3: SLG X_4: ISO X_5: OPS X_6: wOBA X_7: wRAA X_8: wRC X_9: wRC+

Model Building

The number of possible models that can be developed from *p* number of variables = 2^{p-1} (Kutner); so, for the overall W-L% model, there are 2^{22} = 4,194,304 possible models and for the RS model, there are 2^8 = 256 possible models. From all these possible models, it is necessary to identify the combination of independent variables that is the best at explaining the variation in

the dependent variable, therefore leading to the best possible model. There are two statistical methods that can help us do this: Stepwise Regression and Best Subsets Regression procedures.

In stepwise regression, predictors are added to or removed from a model in a "stepwise" manner until no more predictors can be justifiably added or removed. More specifically, predictors are added to the model when their p-values are less than or equal to the specified alpha-to-enter significance value, and predictors are deleted from the model when their p-values are greater than the specified alpha-to-remove significance value. In Minitab, the default alpha-to-enter and alpha-to-remove values for stepwise regression are both 0.15 (we will use these default alpha values in our study). Once no more variables can be added or removed, a final model is produced (although this final model is not guaranteed to be the absolute best model for predicting the response).

Conversely, the Best Subsets procedure assesses all possible models from the candidate predictors and provides the best models. It is up to the individual to judge the best model from the models given; generally, the approach is to choose the model with the smallest subset of predictors that also fulfills certain statistical criteria. These statistical criteria include R^2, adjusted R^2, Mallow's C_p, and S.

Both Stepwise regression and Best Subsets procedures will be used in our study to build regression models for W-L% and RS. However, there are a few key differences between the two procedures that are important to note. First, in the stepwise procedure, Minitab actually produces a regression equation along with the coefficients, standard errors of the coefficients, p-values, and VIFs of each term. The Best Subsets procedure does not provide any regression equation nor any information about the specific coefficients in the model. On the other hand, Best Subsets

provides a variety of models to choose from and makes it possible to compare potential models, whereas the stepwise procedure only produces one final model.

Model Selection Criteria

There are various statistics that can be used to compare models and to select the model that fits the data the best. Some of these selection criteria that will be used throughout our data analysis are: R^2, adjusted R^2, MSE, AIC, BIC, and Mallow's C_p.

R^2, also known as the coefficient of determination, is defined as: "the percentage of the response variable variation that is explained by a linear model" (Kutner). It always ranges in value from 0% - 100%; generally, the higher the value, the better the model fits the data. While R^2 is extremely helpful for assessing a model's goodness of fit, it is not a perfect statistic and should be utilized in conjunction with other evaluative measures. R^2 is calculated as:

$$R^2 = SSR / SST = \sum^{n}_{i=1} (\hat{y}_i - \bar{y})^2 / \sum^{n}_{i=1} (y_i - \bar{y})^2$$

where:

SSR = Regression Sum of Squares

$\hat{y}_i$ = predicted value of the response variable in the *i*th case
$\bar{y}$ = mean value of the response variable

SST = Total Sum of Squares

y_i = value of the response variable in the *i*th case
$\bar{y}$ = mean value of the response variable

Quite simply, SSR measures the amount of variance in the modeled values, while SST is the sum of the squared differences of each observation from the overall mean. Ideally, both SSR and SST

quantities should be small, indicating that the model fits tightly to the data and thus maximizing R^2.

One of the relevant issues with R^2 is that every time a predictor is added to the model, the R^2 increases; therefore, if solely observing the R^2 value, a model with more predictors may seem to fit the data better than a model with fewer predictors. This is problematic in regard to this data set because our goal is to find not only a good-fitting model, but also a simple model for predicting win percentage. To account for this, adjusted R^2 can be used. The purpose of adjusted R^2 is to compare regression models that contain different amounts of predictors. Like R^2, we seek to find the model with the highest adjusted R^2 value. Adjusted R^2 is calculated as follows:

$$R^2\,(\text{adj}) = 1 - [(1 - R^2)\,(n - 1) / (n - p - 1)]$$

where:

n = sample size
p = number of explanatory variables in the model (not including the constant)

While it is always lower than R^2, adjusted R^2 increases when adding a predictor improves the model not solely by chance (Kutner). In this way, a model with fewer predictors can be observed as actually being more useful and practical than a comparable model with more predictors.

The Mean Squared Error, or MSE, can also be used to assess goodness of fit. As a simple definition, the MSE is an estimator of the variance of the errors. In comparing linear regression models, the smaller the MSE of a model, the better the model explains the data. The MSE is calculated as:

$$\text{MSE} = \text{SSE} / (n - p)$$

where:

SSE = Sum of Squared Errors
n = sample size
p = number of β coefficients in the model (including the constant)

While the MSE is directly available in the Stepwise regression output in Minitab, one can also look at the standard deviation of errors, denoted a "S", in Minitab's Best Subsets and Stepwise regression outputs. S is simply the square root of the MSE:

$$S = \sqrt{MSE}$$

Based on this calculation, S has the same regression interpretation as the MSE – a smaller S value indicates a better-fitting model (Kutner).

Another way of comparing the goodness of fit of various regression models is by using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). While each is calculated slightly differently, both information criteria use the SSE, the number of parameters in the model, and the sample size to provide information about the relative quality of each model (Kutner). AIC and BIC are respectively calculated as follows:

$$AIC = n \ln(SSE/n) + 2p$$

$$BIC = n \ln(SSE/n) + p \ln(n)$$

where:

n = sample size
SSE = Sum of Squared Errors
p = number of β coefficients (including the constant)

The model with the lowest AIC or BIC is the preferred model. Similar to adjusted R^2, the AIC and BIC methods penalize models for simply having more parameters (if those parameters don't provide that much more significance to the model). One of the main differences between AIC and BIC is that AIC tends to overfit the data while BIC tends to underfit the data (Chakraborty).

Lastly, unique to the Best Subsets output is the Mallow's C_p statistic. In choosing between models using Best Subsets, a smaller Mallow's C_p value is better. The formula for the

Mallow's C_p statistic is:

$$\text{Mallow's } C_p = (SSE / \sigma^2) - (n - 2p)$$

where:

SSE = Sum of Squared Errors
σ^2 = MSE for the full model
n = sample size
p = number of β coefficients (including the constant)

Multicollinearity

One of the issues relevant to multiple linear regression that we want to be aware of throughout our data analysis is the issue of multicollinearity. In multiple linear regression, multicollinearity occurs when two or more predictor variables are highly correlated with each other, which makes analyzing a final regression model quite difficult. One of the ways to detect multicollinearity is by looking at the Variance Inflation Factors (VIFs) of each variable in the model. According to Michael Kutner, "VIFs measure how much the variance of the estimated regression coefficients is inflated as compared to when the predictor variables are not linearly related" (Kutner). The VIF for the *k*th predictor is calculated as:

$$VIF_k = 1 / (1 - R^2_k)$$

where:

R^2_k = the R^2 value obtained by regressing the *k*th predictor on the remaining predictors

As a general rule of thumb, any VIF greater than 10 is a sign of serious multicollinearity (Kutner). Some of the problematic impacts of multicollinearity are increasing the standard error of the coefficients, causing coefficients to switch signs in a model, and overall making the model difficult to interpret.

Data Collection

Data were collected from all thirty MLB teams from the years 1998-2015. This time frame was chosen because 1998 was the year the MLB expanded to its current thirty teams (Zimniuch). Each team's W-L%, total runs scored, and total runs allowed were collected for each of the eighteen seasons. Because the overarching goal of this study is to analyze wins, which are the result of an entire team's effort, team averages were taken for each of the individual offensive and defensive metrics across all thirty teams for the eighteen seasons, resulting in 540 data points (throughout the data analysis, it is important to keep in mind that all data are team-based).

Data were obtained from FanGraphs.com and Baseball-Reference.com and imported as CSV files. While a majority of the team averages for the various statistics were available on these sites, some statistics had to be calculated manually once in Excel (for instance, OPS was calculated manually by summing OBP and SLG).

Minitab 17 was the statistical software used for all statistical procedures. Because Minitab was not able to directly produce the AIC and BIC values, these had to be manually calculated from data available in the regression output (for these calculations, see Appendix B).

Chapter 4

Data Analysis and Results

Correlations between All Measures and W-L%

First, we wanted to observe how each metric correlated individually with W-L%. Through Minitab, we were able to find the correlation coefficient (R) values for each metric with W-L%. Table 3 contains all R values, ordered from highest to lowest, of the various statistics with W-L%:

Table 3: Correlation Coefficients with W-L%

Name of Metric	R
PE	0.940
WAR	0.744
wRC+	0.673
RA	-0.611
ERA	-0.609
WHIP	-0.602
wRAA	0.596
LOB%	0.578
RBI	0.526
RS	0.522
wRC	0.509
wOBA	0.502
FIP	0.501
OPS	0.498
OBP	0.492
SLG	0.465
BB/9	-0.449
ISO	0.419
BABIP	-0.416
DER	0.414
BA	0.375
Fld%	0.362
K/9	0.326

As seen in Table 3, the PE formula was the most highly correlated with W-L%. With an R value of 0.940, the PE formula had an extremely strong positive correlation with W-L%; understandably, the R value is so high because the PE formula is specifically designed as a W-L% predictor. WAR had the next highest R value of 0.744. Interestingly, RA had a higher R value than RS. RA had an R value of -0.611, demonstrating a fairly strong negative correlation between RA and W-L%. In addition to RA, ERA, WHIP, BB/9, and BABIP all had negative correlations as well. The negative correlations for these metrics make sense because if run prevention metrics have high values, this means more runs are allowed which is negatively related to winning. RS had an R value of 0.522; although not as high as RA's R value, 0.522 still

represents a fairly strong correlation between RS and W-L%, and the positive value represents that run scoring is positively related to winning. The three lowest correlations were from BA, Fld%, and K/9, all conventional statistics. Overall, most of these statistics had high correlations ($R > |0.5|$) with W-L%, and all of the statistics were at least moderately correlated ($R > |0.3|$) with W-L%. This coincides with our literature review, since all of these statistics have been discussed to somehow be related to winning.

Again, it's important to note that the higher correlation coefficient values do not mean that these metrics cause W-L% more than the metrics with the lower correlation coefficient values; in fact, *none* of these values imply anything about causing W-L%. However, these values can help us understand how these metrics might perform in a predictive model with W-L%, at which point we can identify potential causation between W-L% and these metrics.

Regression Analysis: W-L% as the Response

First, we performed the Best Subsets procedure using W-L% as the response and all 23 metrics as the predictors. The model that we chose had the highest adjusted R^2, lowest Mallow's C_p value, and one of the lowest S values. These values with the chosen predictors are detailed in Table 4:

Table 4: Model Selection Criteria for W-L% Model, Best Subsets Regression

Number of predictors	**Predictors**	**R^2**	**R^2 (adj)**	**Mallow's C_p**	**S**
15	PE, RS, RA, SLG, OPS, wOBA, wRC, wRC+, wRAA, K/9, LOB%, ERA, WHIP, FIP, and Fld%	89.8%	89.5%	14.5	0.023173

We then ran a stepwise regression procedure on all 23 predictors and obtained the following output:

Table 5: Variables in Final W-L% Model, Stepwise Regression

Predictor	Coefficient	Standard Error (SE) of Coefficient	T-value	P-value	VIF
(Constant)	-1.391	0.473	-2.94	0.003	
PE	0.281	0.130	2.16	0.031	84.58
RS	0.000324	0.000099	3.29	0.001	73.36
OPS	0.425	0.193	2.20	0.028	59.67
wOBA	1.311	0.565	2.32	0.021	79.81
wRC	-0.000261	0.000104	-2.50	0.013	91.55
wRC+	0.000853	0.000281	3.04	0.002	5.82
wRAA	-0.000174	0.000044	-3.94	< 0.001	9.21
BB/9	-0.01077	0.00338	-3.19	0.002	2.17
LOB%	0.333	0.153	2.18	0.030	12.52
ERA	-0.0389	0.0125	-3.12	0.002	46.30
FIP	-0.02513	0.00810	-3.10	0.002	12.34
Fld%	0.964	0.484	1.99	0.047	1.80

Table 6: Model Selection Criteria for W-L% Model, Stepwise Regression

R^2	R^2 (adj)	MSE	S
89.61%	89.37%	0.000542	0.0232785

Final Stepwise W-L% Regression Model:

W-L% = -1.391 + 0.281 PE + 0.000324 RS + 0.425 OPS + 1.311 wOBA - 0.000261 wRC + 0.000853 wRC+ - 0.000174 wRAA - 0.01077 BB/9 + 0.333 LOB% - 0.0389 ERA - 0.02513 FIP + 0.964 Fld%

While the model we chose from the Best Subsets procedure had 15 predictors, this Stepwise procedure produced a model with 12 predictors. In addition, the difference in the adjusted R^2 values is very slight: 89.5% for the Best Subsets 15-variable model and 89.37% for the Stepwise 12-variable model.

Using the stepwise model, 89.61% of the variation in W-L% is explained by this model. In addition, the variables all had p-values < 0.05, meaning that all of the variables together were significant predictors of W-L% and therefore should be kept in the model. However, eight of these twelve variables had VIFs above 10, representing a sign of serious multicollinearity; in fact, wRC had the highest VIF value of 91.55. While it is not surprising that many of these variables were correlated with each other (such as wRC and wRC+, for instance), it is important to take caution when interpreting this model.

As mentioned previously, the variables with the largest coefficients (in terms of absolute value) should explain the greatest variation in the response. According to the model, wOBA, with $\beta = 1.311$, had the largest coefficient; this can be interpreted as if all other variables were held constant, a 1.311-unit increase in wOBA would lead to a 1-unit increase in W-L%. wOBA is followed respectively by Fld% ($\beta = 0.964$) and OPS ($\beta = 0.425$). This means that these three measures, two of which are conventional statistics, explained the most amount of variability in W-L% compared to the other measures. Surprisingly, measures such as wRC ($\beta = -0.000261$) and wRAA ($\beta = -0.000174$), both sabermetric statistics, explained the least amount of variability in W-L% compared to the other measures. However, our interpretations of the β coefficients are limited because of the varying nature of the units (i.e. W-L as a percentage and wRC as runs). Nevertheless, each variable does have some impact on W-L% and so it's still important to analyze these variables in the model.

In the final model, most of the signs of the coefficients make sense; for example, RS and OPS had positive signs (meaning that *ceteris paribus,* an increase in these values would lead to an increase in W-L%), and BB/9 and ERA had negative signs (meaning that *ceteris paribus,* a decrease in these values would lead to an increase in W-L%). However, the signs of the

coefficients of wRC and wRAA were negative in the model, even though we expected them to be positive based on their R values and based on our intuition that these measures contribute to scoring runs. The existence of multicollinearity could have contributed to the switching of these signs, as it does not make sense that a decrease in either wRC or wRAA would lead to an increase in W-L%.

Interestingly, even though WAR was shown to be very highly correlated with W-L% individually (R= 0.744), WAR was not included as a predictor in either model, even though it is heralded as one of the most effective sabermetric statistics. Furthermore, both Best Subsets and Stepwise models contained a mixture of conventional and sabermetric measures. In the Best Subsets model, eight variables (PE, wOBA, wRC, wRC+, wRAA, LOB%, WHIP, and FIP) of the total fifteen are considered sabermetric measures, while in the Stepwise model, seven variables (the same eight variables in the Best Subsets model, minus WHIP) of the total twelve are considered sabermetric measures.

Overall, this 12-variable Stepwise model is rather difficult to interpret and quite complex, and so we wanted to see if there was simpler model for predicting W-L% (recalling that this was one of our main goals). Previously, we discovered that the PE variable was the most highly correlated – and the only metric that can be classified as being "nearly perfectly correlated" (R = 0.94) – with W-L%. With this in mind, we ran a simple linear regression using the Stepwise procedure with PE as the single predictor and obtained the following statistics:

Table 7: Model Selection Criteria for Simple W-L% Model

R^2	R^2 (adj)	MSE	S
88.31%	88.29%	0.00060	0.0244360

Looking at the R^2 value in Table 7, the single-variable PE model explained 88.31% of the variation in W-L% - only 1.3% less than the variation explained by the complex Stepwise model. Finally, Table 8 below is a comparison of all three W-L% models that we have developed. Not included in Table 8 are R^2 (since each model has a vastly different amount of predictors), Mallow's C_p (only applicable to Best Subsets), and S (which is simply the square root of the MSE):

Table 8: Comparison of Simple and Complex W-L% Models

Regression procedure used	**Response**	**Predictors Used**	**Number of Predictors in Best/Final Model**	**R^2 (adj)**	**MSE**	**AIC**	**BIC**
Best Subsets	W-L%	All metrics (23)	15	89.5%	0.000537	-4050.190695	-3981.528803
Stepwise	W-L%	All metrics (23)	12	89.37%	0.000541	-4048.212107	-3992.421708
Stepwise	W-L%	PE	1	88.29%	0.00060	-4006.720636	-3998.137497

Comparing the three models, we can see that the single-variable PE model had very similar adjusted R^2, MSE, AIC, and BIC values with the more complex models. While the Best Subsets model had the highest adjusted R^2, lowest MSE, and lowest AIC value, the single-variable PE model actually had the lowest BIC value. Despite the single-variable PE model having the lowest adjusted R^2 value, it is only slightly lower than the more complex models, which is important to note because this model has significantly less predictors than either complex model. This single-variable model is also much easier to interpret compared to the Stepwise model since there is no issue of multicollinearity. Therefore, it is best to use the single-variable model, with PE as the explanatory variable, to predict W-L% since it is the most parsimonious of the three models, and therefore more helpful for teams and managers to utilize.

While general managers want to increase their team's W-L% (or simply win more games), they also want to know how they can incorporate this goal into an applicable game strategy. If we look back at the PE formula [$PE = RS^2 / (RS^2 + RA^2)$], we see that its two components are RS and RA. If a team scores more runs than they allow in a game, then they win that game. Therefore, a general manager would be more focused on how their team can score runs and prevent runs on a daily basis, instead of just aiming to increase a winning percentage number. With this in mind, it's necessary to understand which metrics are the most predictive of RS and RA respectively – that is, which metrics are the most correlated to scoring more runs and allowing fewer runs, thus increasing the Pythagorean Expectation. In this study, we will just focus on analyzing all the offensive metrics that relate to scoring runs.

Correlations between Offensive Measures and RS

Similar to W-L%, we wanted to observe how each offensive metric correlated individually with RS. While we know that all of these offensive metrics will have high positive associations with RS (since the nature of all these statistics are related to offensive events and run scoring), our purpose here is to observe which of these conventional and sabermetric statistics are *more* associated with RS compared to others, since all these statistics vary in what they measure. So, with Minitab we found the R values for these ten metrics, ordered from greatest to least R values in Table 9:

Table 9: Correlation Coefficients with RS

Name of Metric	R
RBI	0.997
wRC	0.965
wOBA	0.957
OPS	0.956
SLG	0.914
OBP	0.905
BA	0.822
wRAA	0.791
ISO	0.772
wRC+	0.698

With the R values ranging from 0.698 to 0.997, all offensive metrics are strongly correlated (if not nearly perfectly correlated) with RS, which is not surprising at all. The RBI metric has a near perfect correlation (R = 0.997) with RS, which is not surprising because RBI is itself a variation of the RS metric. wRC, wOBA, and OPS have the next highest correlations respectively, all having extremely strong positive correlations with RS. This is not very surprising either because these metrics have become heavily favorited and supported by many baseball analysts and sabermetric websites like FanGraphs.

There are, however, some interesting and surprising results from this RS correlation data. First, wRAA was less correlated with RS than wOBA, even though wRAA incorporates wOBA into its formula. Similarly, wRC+ is much less correlated with RS compared to wRC, which is especially surprising because wRC+ is simply an adjusted (and supposedly improved) form of wRC. Also, while SLG and OBP are equal components in the OPS statistic, many have argued that OBP should be valued more highly than SLG in terms of scoring runs, However, these data show that individually, SLG is slightly more correlated with scoring runs than OBP, with R values of 0.914 and 0.905 respectively.

Based on these correlations, the newer, sabermetric statistics aren't necessarily more associated with scoring runs compared to the conventional statistics. In fact, conventional statistics such as SLG, OBP, and BA were more associated with scoring runs compared to wRAA and wRC+, both of which have quite complex calculations. This is very important to note because with all of the statistics available, in looking at how runs are scored, the conventional statistics should not be disregarded.

Again, from this specific correlation information, it's relevant to note that these metrics don't necessarily *cause* runs scored; rather, they are just simply highly associated with scoring runs, and to varying degrees. However, next we will use these highly correlated statistics to observe which ones cause the most impact on RS in a predictive model.

Regression Analysis: RS as the Response

We used the Best Subsets procedure to observe which subset of offensive predictors were the best in a model. We entered nine of the offensive metrics as predictors: BA, OBP, SLG, ISO, OPS, wOBA, wRAA, wRC, and wRC+. We did not select RBI as a predictor in this model due to its repetitiveness and almost perfect correlation with RS, the response. In observing the adjusted R^2, Mallow's C_p, and S values from the Best Subsets output, we chose a model with the following criteria as the best model:

Table 10: Model Selection Criteria for RS Model, Best Subsets Regression

Number of Predictors	**Predictors**	**R^2**	**R^2 (adj)**	**Mallow's C_p**	**S**
4	BA, OBP, wRC, and wRAA	93.5%	93.5%	1.6	22.239

Next, we performed Stepwise regression, again using the same nine offensive metrics used in Best Subsets for our predictors here. The results are detailed below:

Table 11: Variables in Final RS Model, Stepwise Regression

Predictor	Coefficient	SE of Coefficient	T-value	P-value	VIF
(Constant)	78.7	43.9	1.79	0.074	
BA	353	154	2.30	0.022	3.80
OBP	-696	215	-3.24	0.001	11.16
wRC	1.0809	0.0434	24.93	< 0.001	17.30
wRAA	-0.1564	0.0286	-5.48	< 0.001	4.20

Table 12: Model Selection Criteria for RS Model, Stepwise Regression

R^2	R^2 (adj)	MSE	S
93.53%	93.48%	495	22.2388

Final Stepwise RS Regression Model:

RS = 78.7 + 353 BA - 696 OBP + 1.0809 wRC - 0.1564 wRAA

The Best Subsets and Stepwise procedures produced models with the same predictors and almost identical adjusted R^2, MSE, and S values, so we will base the rest of our RS analysis on the Stepwise output considering Stepwise regression also produced information on the predictor variables themselves.

Four of the nine originally chosen offensive metrics are included in the final RS regression model: two of which (BA and OBP) are conventional measures and two of which (wRC and wRAA) are sabermetric measures. With p-values < 0.05, all four measures together are statistically significant predictors of RS in this model, so they should be kept in the model. In addition, this four-variable model explains 93.53% of the variation in RS, which is very strong.

Based on the magnitude of the coefficients, the variable that explains the greatest variation in RS is OBP (β = -696), followed by BA, wRC, and wRAA respectively. However,

like the W-L% models, it is difficult to interpret the magnitude of these coefficients (especially with OBP being extremely large compared to the other variables) due to scale and the varying nature of the units. In addition, the signs of the OBP and wRAA coefficients in the final model are negative, which does not make sense at all; for instance, with all other variables held constant, decreasing team OBP by 696 units should not lead to one more run scored. Again, this sign switching can be due to the issue of multicollinearity; as seen in Table 11, the OBP and wRC variables have respective VIF values of 11.16 and 17.30. In theory, because only offensive metrics were included in this model and based on the positive R values from Table 9, all of the coefficients should be positive, as all of these metrics are positively associated with scoring runs. Overall, this final regression model is too difficult to interpret accurately (even with only four variables). Intuitively, this model does not make much since either because in whatever context, more runners on base (OBP) should not negatively impact runs scored (from an offensive perspective). Therefore, this model is not very valuable for a general manager to use in order to increase his team's run scoring.

Due to this RS model being not only difficult to interpret but also rather complex, we wanted to see if there was a simpler and more effective way of predicting RS (just like what we did with W-L%). Ignoring RBI, the offensive metrics that were nearly perfectly correlated ($0.9 < R < 1.0$) with RS from Table 9 were wRC ($R = 0.965$), wOBA ($R = 0.957$), OPS ($R = 0.956$), SLG ($R = 0.914$), and OBP ($R = 0.905$). Using all five of these measures, we performed Stepwise regression on each of these variables to see how each of them would perform in a simple linear regression model with RS. Some of the summary statistics for the five produced models are found in Table 13:

Table 13: Model Selection Criteria for Simple RS Models

Predictor	R^2	R^2 (adj)	MSE	S
OBP	81.87%	81.84%	1378	37.1198
SLG	83.48%	83.45%	1255	35.4301
OPS	91.47%	91.46%	648	25.4559
wOBA	91.59%	91.57%	639	25.2872
wRC	93.11%	93.09%	524	22.8865

According to the data in Table 13, the best simple linear regression model uses wRC as the predictor, as it has the highest R^2 and adjusted R^2 values and lowest S and MSE values. In addition, this wRC model explained 93.11% of the variation in RS, which is only 0.42% less than the variation explained by the four-variable RS model above – an extremely miniscule difference. In terms of explanatory power, the wRC model is followed respectively by the wOBA and OPS simple models, each explaining more than 90% of the variability in RS.

However, even though wRC and wOBA are better single predictors of RS than OPS, it is important to understand what these measures actually mean. wOBA and wRC are both statistics that use linear weights in their calculation, which change each year based on the run environment of each season (Panas). In addition, OPS is definitely the most straightforward in terms of understanding and calculation compared to wOBA and wRC; it is simply the summation of OBP and SLG, whose respective formulas are relatively intuitive as well. Even with a slightly smaller R^2 value compared to wRC and wOBA, OPS accounts for this with its straightforwardness. Furthermore, with OPS (a conventional statistic) being around for a longer time than wOBA or wRC (both sabermetric statistics), general managers are more likely to understand (or just simply recognize) the OPS statistic compared to wOBA and especially wRC. Therefore, in terms of explanatory power *and* real-world practicality, OPS is the most effective predictor among the five chosen simple RS regression models.

Lastly, we wanted to compare the simple RS model (with OPS as the single predictor) with the multivariable RS model found previously:

Table 14: Comparison of Simple and Complex RS Models

Regression procedure used	**Response**	**Predictors Used**	**Number of Predictors in Final Model**	**R^2 (adj)**	**MSE**	**AIC**	**BIC**
Stepwise	RS	All offensive metrics excluding RBI (8)	4	93.48%	495	3354.960544	3376.418389
Stepwise	RS	OPS	1	91.46%	648	3497.900361	3506.483499

Based on this Minitab output (and without understanding the nature of our data set), overall the four-variable model fits the data better than the single-variable model, as it has a higher adjusted R^2 value and lower MSE, AIC, and BIC values. However, through understanding the baseball context of the data, it is still preferred to use the parsimonious model – i.e. the model that explains the data well but has the least amount of predictors – if at all possible. Despite the multivariable model explaining approximately 2% more of the variation in RS, this difference is very slight given the three-variable difference between the models. The simple OPS model still performs very well, accounting for more than 90% of the variation in RS and lacking any multicollinearity issues (since it only has one predictor).

Given all of this information, we prefer to use the basic OPS model to predict RS because it is much simpler and easier to interpret compared to the multivariable model, making it much more effective for statistical purposes and real-life usage. For instance, from a general manager's perspective, it is much easier to incorporate OPS into a "run scoring" strategy compared to the information conveyed in the multivariable RS model – it would be extremely difficult for a

general manager to derive any practical information from the combination of BA, OBP, wRAA, and wRC. More broadly, this RS information can help general managers select players to build a team focused on increasing the amount of runs they score – a very crucial aspect to winning games and thus winning percentage.

Chapter 5

Discussion and Conclusion

Summary of Results

One of the primary goals of this research was to analyze whether the newer sabermetric measures were both statistically and practically more effective than the simpler, traditionally recognized measures in terms of explaining a team's ability to increase their winning percentage. Individually, the sabermetric statistics were generally more highly correlated with W-L% compared to the conventional statistics. Our findings show that Bill James's Pythagorean Expectation, one of the key formulas developed during the sabermetric revolution, is the single best predictor of W-L% compared to other metrics. This is not surprising because the formula was specifically designed to predict a team's winning percentage, whereas the other metrics in this study describe some aspect of a player's or team's offensive or defensive (or both, if using WAR) performance. However, the complex multivariable regression model predicting W-L% contained a mixed number of sabermetric and conventional statistics, so it cannot be confirmed that the sabermetric statistics are completely dominant over the conventional statistics in terms of predicting W-L%.

Based on the PE formula's composition and coinciding with mainstream belief, runs are what matters most when it comes to winning games. The goals of the offense and defense align with this as well; the offense's objective is to score runs while the defense's objective is to prevent runs. More runs scored than allowed in a given game result in a win for that team; the

more wins a team has in a season, the better their winning percentage. However, it appears that decreasing runs allowed might be correlated more to increasing W-L% than increasing runs scored, as shown by the respective correlation coefficients of RA and RS with W-L%. With this in mind, it may be more important for general managers to focus on run prevention strategies compared to run scoring strategies, if preventing runs is more highly associated with winning than scoring runs. This is quite contrary to our original hypotheses that were based on the offensively-focused mindsets of previous general managers like Billy Beane.

With this in mind, it is still much easier to quantify and analyze a team's offensive performance compared to their defensive performance as a whole. So, similarly to W-L%, it was important to understand the measures that were most related to scoring runs. In observing the correlation between the offensive measures and RS, the ranking of highest to lowest correlations was fairly mixed between the sabermetric and conventional statistics. In addition, the RS model contained an equal mixture of sabermetric and traditional measures. Therefore, as with W-L%, the sabermetric offensive measures are not necessarily better than the traditional offensive statistics for predicting RS, contrary to what many sabermetrically-minded individuals would like to think.

The other goal of this research was to find the best parsimonious model in terms of predicting our main response of W-L% and our sub-response of RS. While Minitab produced strong multivariable models to explain W-L% and RS, these models were extremely complex. In addition, these models were difficult to interpret due to the high existence of multicollinearity among our predictors and the varying nature of the statistics themselves. Taking a step back, we looked at simpler regression models and found that the simple models didn't vary much from the more complex models; specifically, the simple models of PE explaining W-L% and OPS

explaining RS were each comparable with their multivariable counterparts. Consequently, it is much more valuable to use these simpler models when evaluating the factors that contribute to winning and scoring runs.

Real World Applications

The overall purpose of this statistical study was to observe the factors that relate to winning percentage through regression analysis. Practically, teams and managers want to understand such models in order to see which statistics may be more important than others in terms of their impact on winning percentage. Because each statistic describes a unique aspect of offensive or defensive performance, this may help teams understand which areas of the game should be focused on more. For instance, the simple RS model with OPS as the predictor indicates that understanding the OPS statistic should be a priority for the offense and a team's offensive strategy. The OPS statistic is equally divided between simply getting on base (OBP) and hitting for power (SLG), so teams should focus on both of these components in order to potentially increase runs scored.

As mentioned throughout our data analysis, this information is also extremely useful for general managers. If a general manager understands which statistics are most associated with winning and scoring runs, then he will try to build a team that would have the highest averages for these statistics. Of course, this would involve selecting players that rank highly in those statistics themselves. In addition, these data can also help managers understand that they don't necessarily need to use the newest or most complex statistics to formulate successful winning or

run scoring strategies; oftentimes, the simpler statistics and simpler models can provide the exact information they need.

Limitations and Future Considerations

Throughout data collection and analysis, we encountered some limitations that hopefully can be addressed in future research.

Defensive Statistics

Throughout our research and analysis, we discovered how difficult it was to quantitatively analyze a team's defense in order to observe the measures which contribute to decreasing runs allowed. The tracking of defensive metrics has always lagged behind the tracking of offensive metrics for numerous reasons. In *Baseball Between the Numbers*, Keri discussed one of the underlying problems of defensive metrics:

> "Commonly used defensive statistics – the building blocks of any metric or measure of performance – are terribly and perhaps irrevocably flawed. The mainstream measure of defensive performance – the error – is a judgment call by the game's official scorer, who is guided only by his gut and a vague reference in the rules to 'ordinary effort'…The official statistics are not an objective reflection of what's happening on the field. These discrepancies are the primary reason that standard defensive statistics cannot be used to determine the quality of defensive performance – the same play on the same field can be seen differently by different scorers depending on the interpretation of the word 'ordinary.'"

In addition, while offensive performance is more player-based (i.e. one player bats per play), defensive performance (i.e. what contributes to preventing runs) can really only be

analyzed on a team-oriented basis – before and potentially once a ball is put into play, the pitcher and eight other fielders are responsible for what happens to that ball and that runner.

Due to the ambiguity and potential misinterpretation of the defensive data, we decided not to develop an RA model in this study. In addition, we initially believed that scoring more runs would be more important than allowing fewer runs – Billy Beane even made it clear how highly he valued offensive performance, and how defensive performance was not nearly as important to him when it came to winning (Lewis). Based on this, initially it was more important for us to analyze those metrics which contribute to scoring more runs.

Albeit a limitation in this study, the defensive measures initially described – in addition to some of the newer defensive measures being developed each day – should definitely be analyzed more closely in future research in order to observe which defensive metrics most effectively explain run prevention and thus help decrease RA. In addition, hopefully a future model can be developed that provides insight into whether pitching or fielding is more important in predicting RA.

Multicollinearity

As discussed throughout our data analysis and results, one of the significant issues that we came across in our regression analysis was the issue of multicollinearity. Given the nature of our data, it is clear why multicollinearity would be evident; many of these baseball metrics are extensions or variations of other metrics, such as wRC and wRC+, or OPS and OBP/SLG. However, Minitab does not understand the nature of our data and can still put highly correlated

predictors in a model together, which greatly impacted the coefficients in the final models and thus affected our interpretation.

Two ways to solve the issue of multicollinearity are ridge regression and principal components regression (Kutner). Ridge regression penalizes the size of the regression coefficients and is used to "shrink" the estimated coefficients towards zero. This serves to reduce the variance of the estimates (in multicollinearity, variance is inflated among the estimates), and so multicollinearity is reduced as a result (Kutner). On the other hand, principal components regression makes use of principal components analysis (PCA), which attempts to find linear combinations of the predictors that are uncorrelated with each other but still explain as much of the variance in the response as possible; these "new" predictors can then be entered into a regression model through the principal components regression procedure (Kutner). These are simplified explanations, but hopefully in future research these regression techniques can be used in order to create more accurate estimates of the regression coefficients and establish more viable regression models as a result.

Data Time Frame Selection and Time Related Data

Because we chose to collect data from 1998-2015, we were limited to analyze statistics that were only available on FanGraphs.com and Baseball-Reference.com back to 1998. Thus, we weren't able to include many of the newer pitching and fielding statistics that were developed in the early 2000s because we'd be missing years of data for those metrics. For instance, FanGraphs includes player and team data for xFIP, an adjusted version of FIP, and DRS (Defensive Runs Saved) and UZR (Ultimate Zone Rating), both of which are considered "better representations of

defensive value than something like fielding percentage" (Weinberg). However, these statistics are tracked back only to 2002. In the future, in order to analyze these more recent statistics accurately, one could analyze data beginning the year they were developed, but with this comes the disadvantage that fewer years of data are included in the analysis.

In addition, throughout our analysis, we did not give consideration to the time-related nature of the data set. In other words, if a team's management does not change across a time period, then the data are likely to be extremely correlated between these years. From another perspective, team management can change drastically over an eighteen-year time period, and so a team's stats may drastically change across these years. A change in management can greatly impact a team's success in a given year, but our data do not account for this. Future research can attempt to account for the fact that data for each team may not be completely independent from year to year.

Use of Alternative Predictors and Responses

While this study only included various conventional and sabermetric baseball statistics as predictors (all of which were continuous), it would be interesting to include other predictors to see 1) how much (if at all) they correlated which W-L% and 2) how predictive they would be in a W-L% model.

One interesting predictor could be league – National League (NL) or American League (AL). League, a binary predictor, could be included in the model to see if any variation in winning percentage can be partially explained by what league a team is in. If league proved to be a significant predictor, then it'd also be interesting to develop two W-L% models to see which

factors were more relevant in predicting winning percentage based on each league. Hypothesizing, these two models would differ in which metrics were significant due to the different league structures (generally speaking, the AL tends to be more offensively-focused while the NL is more pitching-focused).

While winning and thus a strong winning percentage is of paramount importance for teams, winning a lot of games does not necessarily guarantee a spot in the playoffs (while winning the World Series is the *true* ultimate goal for MLB teams and the front office, securing a spot in the playoffs is the only way to possibly achieve this goal). For a team to make the playoffs (excluding the two wild card teams), a team simply has to have won the most games in their division by season's end. So, if a team with an ending W-L% of 0.650 is in a strong division but in third place in that division, and a team with an ending W-L% of .500 is in a weaker division but in first place, then the 0.500 W-L% team will still make the playoffs while the team with a 0.650 W-L% (and thus more wins) may not make the playoffs (depends on wild card opportunities). With this in mind, a potential alternative response for this study could be if a team made the playoffs or not. A "playoff" binary variable could be introduced in the dataset that would be coded "0" if the team in that year did not make the playoffs or coded "1" if the team in that year did make the playoffs. Because the response would be binary in this case, linear regression would not be applicable and logistic regression is suggested. Still, such analysis would be interesting to observe, especially if this method's results differed significantly from the W-L% results found in this study. If so, teams might have to adjust what metrics they focus on depending on whether they would rather have a high winning percentage overall or would rather just strive to be the best in their division in order to make the playoffs.

Appendix A

Glossary of Baseball Statistics

1B: Single
2B: Double
3B: Triple
A: Assists
AB: At-Bats
BA: Batting Average
BABIP: Batting Average on Balls in Play
BB: Walks
BB/9: Walks per 9 innings
BFP: Batters Facing Pitcher
CS: Caught Stealing
DER: Defensive Efficiency Ratio
E: Errors
ERA: Earned Run Average
FIP: Fielding Independent Pitching
Fld%: Fielding Percentage
H: Hits
HBP: Hit by Pitch
HR: Home Run
IBB: Intentional Walks
IP: Innings Pitched
ISO: Isolated Power
K: Strikes
K/9: Strikeouts per 9 innings
LOB%: Left On-Base Percentage
OBP: On-Base Percentage
OPS: On-Base Plus Slugging Percentage
PA: Plate Appearances
PE: Pythagorean Expectation
PO: Putouts
RA: Runs Allowed
RBI: Runs Batted In

RF: Range Factor
RS: Runs Scored
SB: Stolen Bases
SF: Sacrifice Flies
SLG: Slugging Percentage
UZR: Ultimate Zone Rating
WAR: Wins Above Replacement
WHIP: Walks plus Hits per Inning Pitched
W-L%: Win-Loss Percentage
wOBA: Weighted On-Base Average
wRAA: Weighted Runs Above Average
wRC: Weighted Runs Created
wRC+: Weighted Runs Created Plus
xFIP: Expected Fielding Independent Pitching

("Baseball Abbreviations")

Appendix B

AIC and BIC Calculations

$$AIC = n \ln(SSE/n) + 2p$$

$$BIC = n \ln(SSE/n) + p \ln(n)$$

Model 1: W-L% regressed on all metrics

Stepwise Regression:

AIC = 540 x ln (0.28557/540) + 2(13) = -4048.212107
BIC = 540 x ln (0.28557/540) + 13ln(540) = -3992.421708

Best Subsets Regression:

In the Best Subsets output, the only way we are able to derive the SSE value is from the S value:

$S = \sqrt{MSE}$
$S^2 = MSE$
$MSE = SSE / (n-p)$
$SSE = MSE(n-p)$
$SSE = S^2(n-p)$
$SSE = (0.023173)^2 (540-16)$
$SSE = 0.28138$

AIC = 540 x ln(0.28138/540) + 2(16) = -4050.190695
BIC = 540 x ln(0.28138/540) + 16ln(540) = -3981.528803

Model 2: W-L% regressed on PE

AIC = 540 x ln(0.3212/540) + 2(2) = -4006.720636
BIC = 540 x ln(0.3212/540) + 2ln(540) = -3998.137497

Model 3: RS regressed on offensive metrics

AIC = 540 x ln(264591/540) + 2(5) = 3354.960544
BIC = 540 x ln(264591/540) + 5ln(540) = 3376.418389

Model 4: RS regressed on OPS

AIC = 540 x ln(348626/540) + 2(2) = 3497.900361
BIC = 540 x ln(348626/540) + 2ln(540) = 3506.483499

BIBLIOGRAPHY

Albert, Jim and Jay Bennett. *Curve Ball.* New York, NY: Springer-Verlag, 2001. Print.

"Baseball Abbreviations." *Baseball Almanac.* Baseball Almanac, Inc. N.d. Web. 30 January 2016.

Birnbaum, Phil. "A Guide to Sabermetric Research." *SABR.org.* SABR. N.d. Web. 10 Feb 2016.

Chakraborty, Ranajit, C. R. Rao, and Pranab K. Sen. *Handbook of Statistics: Bioinformatics in Human Health and Heredity*. Amsterdam, Netherlands: North Holland Publishing, 2012. Web.

Cockcroft, Tristan H. "A primer on BABIP." *ESPN Fantasy Baseball Draft Kit*. ESPN. 15 February 2012. Web. 10 February 2016.

Costa, Gabriel B., Michael R. Huber, and John T. Saccoman. *Understanding Sabermetrics: An Introduction to the Science of Baseball Statistics.* Jefferson, NC: McFarland & Company Publishers, 2008. Print.

Di Fino, Nando. "The Secret History of WHIP." *The Wall Street Journal.* Dow Jones & Company, Inc. 3 August 2009. Web. 10 February 2016.

Gerard, David P. *Baseball GPA: A New Statistical Approach to Performance and Strategy.* Jefferson, NC: McFarland & Company Publishers, 2013. Print.

Hopkins, Will G. "A New View of Statistics: A Scale of Magnitudes for Effect Statistics." *Sportscience*. Will Hopkins. 2002. Web. 27 February 2016.

Judge, Jonathan. "FIP, In Context." *The Hardball Times*. The Hardball Times. 11 March 2015. Web. 10 February 2016.

Keri, Jonah, James Click, James Davenport, Neil Demause, Steven Goldman, Dayn Perry, Nate Silver, and Keith Woolner. *Baseball Between the Numbers*. New York, NY: Basic Books, 2006. Print.

Kutner, Michael H., Chris Nachtsheim, and John Neter. *Applied Linear Regression Models*. 4th ed. Boston: McGraw-Hill/Irwin, 2004. Print.

Lewis, Michael. *Moneyballl: The Art of Winning an Unfair Game.* New York, NY: W.W. Norton, 2003. Print.

Long, J. Scott. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, California: SAGE Publications, 1997. Print.

Panas, Lee. *Beyond Batting Average.* Lulu.com, 2008. Print.

Schwarz, Alan. *The Numbers Game.* New York, NY: St. Martin's Press, 2004. Print.

Studeman, Dave. "Left on Base." *The Hardball Times.* The Hardball Times. 22 December 2005. Web. 10 February 2016.

"Weighted Runs Created Plus." *MLB.com Glossary – Advanced Stats*. MLB.com. N.d. Web. 10 February 2016.

Weinberg, Neil. "BABIP." *FanGraphs Sabermetrics Library.* FanGraphs. N.d. Web. 10 February 2016.

Weinberg, Neil. "FIP." *FanGraphs Sabermetrics Library.* FanGraphs. N.d. Web. 10 February 2016.

Weinberg, Neil. "Stats to Avoid: Runs Batted In (RBI)." *FanGraphs*. FanGraphs. 14 October 2014. Web. 10 February 2016.

Weinberg, Neil. "Strikeout and Walk Rates." *FanGraphs Sabermetrics Library*. FanGraphs. N.d. Web. 10 February 2016.

Weinberg, Neil. "UZR." *FanGraphs Sabermetrics Library.* FanGraphs. N.d. Web. 10 February 2016.

Weinberg, Neil. "WAR." *FanGraphs Sabermetrics Library.* FanGraphs. N.d. Web. 10 February 2016.

Weinberg, Neil. "wOBA." *FanGraphs Sabermetrics Library.* FanGraphs. N.d. Web. 10 February 2016.

Zimniuch, Frank. *Baseball's New Frontier: A History of Expansion, 1961-1998*. Lincoln, NE: University of Nebraska Press, 2013. Print.

ACADEMIC VITA

Academic Vita of Victoria DeCesare
vdecesare428@gmail.com

Education:

The Pennsylvania State University – University Park, PA May 2016
Bachelor of Science in Science, General Option
Minors: Business in the Liberal Arts, Statistics, and Spanish; Smeal College Business Fundamentals Certificate

Relevant Experience:

AstraZeneca – Wilmington, DE June 2015 – August 2015
Managed Markets – Capability Building and Customer Insights Intern

- Assisted in creating web-based analytical tools and dashboards that provided visualizations of data from national and regional health care payers and providers for key insights into AstraZeneca's market access
- Completed a long-term assessment of health care quality measures and the impact of health care quality on the pharmaceutical industry and AstraZeneca's customers
- Generated reports of market access percentage for AstraZeneca's entire drug portfolio and competitor data

Major League Baseball Advanced Media (MLB.com) – New York, NY June 2014 – August 2014
Social Media Analytics Intern

- Analyzed statistical social data from six MLB clubs using various social data analytic and visualization tools
- Evaluated social media trends and created marketing strategies for how the MLB clubs and sponsors can most effectively utilize social media to connect more with fans and drive ticket sales
- Planned social media promotions for teams and events; individually implemented MLB's Random Acts of FanFest All Star Weekend Twitter promotion, which had approximately 280 eligible participants

Honors:

Phi Kappa Phi – National Honor Society
Mu Sigma Rho –National Statistics Honorary Society

Skills:

- Statistical programming languages and packages: SAS, R, and Minitab
- Spanish language skills: Written (Business level) and Spoken (Conversational – Business Level) fluency

Additional Leadership Experience:

Apollo – Penn State Dance Marathon (THON) organization August 2012 – May 2016
Alternative Fundraising Co-chairperson, Subgroup leader, THON 2016 dancer

- Lead a subgroup of 50 – 60 members within Apollo, a 250-member organization that financially and emotionally supports children with cancer through THON and the Four Diamonds Fund

Penn State Athletic Department January 2015 – May 2015
Customer Relations Representative

Relay for Life of Penn State October 2014 – April 2015
Mission and Awareness Captain