Title
Understanding the Market Value and Utility of High-Variance Starting Pitchers

Author
Lepore, James Francis

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Understanding the Market Value and Utility of High-Variance Starting Pitchers

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

James Francis Lepore

2018

ABSTRACT OF THE THESIS

Understanding the Market Value
and Utility of High-Variance
Starting Pitchers

by

James Francis Lepore

Master of Applied Statistics
University of California, Los Angeles, 2018
Professor Frederic R. Paik Schoenberg, Chair

While Major League Baseball has long been at the forefront of sports analytics, the most commonly used metric to evaluate the quality of a pitcher's performance is called the "Earned Run Average" (ERA) which is a global figure with no consideration for the distribution of values. This paper investigates the relationship between observed win percentages among pitchers that exhibit high start-to-start variation in the quality of their starts (consistent pitchers) relative to pitchers that exhibit low start-to-start variation in the quality of their starts (inconsistent pitchers). Logistic regression is first leveraged to transform the traditional components of ERA, runs allowed and innings pitched, into a proxy for expected win probability. Then, through the use of bootstrap sampling, a simulation of hypothetical pitchers is created to compare the difference in actual and expected win totals among pitchers that exhibit different distributional characteristics. Finally, the "Probability of Superiority" between opposing pitchers in individual games is calculated in an attempt to identify optimal matchups. In summary, a statistically significant pattern of outperforming expected win totals is found among pitchers with higher start-to-start variation. However, over the time period analyzed, pitchers at the major league level are much more similar than they are different. As a result, the magnitude of the difference in actual and expected win totals

between consistent and inconsistent pitchers — while statistically significant — is arguably not large enough to be considered practically significant.

The thesis of James Francis Lepore is approved.

Yingnian Wu

Juana Sanchez

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2018

To the never-ending support of my family,

for always giving me the courage to pursue my dreams...

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to thank every professor that I had during my time at the University of California, Los Angeles, over my fours years of undergraduate studies, and my two years of graduate studies. Over eight years ago I first chose to attend UCLA because I wanted to be challenged and pushed to be the very best that I could possibly be. Without the constant guidance and support of my professors, I would not be where I am today academically or professionally.

I would also like to thank the volunteer team of baseball enthusiasts at Retrosheet for taking the time to compile every play in Major League Baseball since 1930, and for being generous enough to make all of their hard work publicly available. Without them my thesis, and much of the research that has driven innovation in baseball over the last decade, would not have been possible. The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at www.retrosheet.org.

CHAPTER 1

Introduction

Even before Michael Lewis put the frugal, yet perennially overachieving, Oakland Athletics and their analytically inclined general manager Billy Beane on the map in 2003 with his book *Moneyball: The Art of Winning an Unfair Game*, statistics have long been at the center of America's pastime. The box score, a collection of statistics that memorialized the performances of pitchers and hitters for specific games, was developed by Henry Chadwick in 1858 [Puerzer, 2002]. In the early 1970s, a second baseman, and future manager, of the Baltimore Orioles Davey Johnson wrote a baseball simulation in FORTRAN to try to convince his manager at the time, Earl Weaver, to bat him second in the order [Porter, 1984]. Around the same time, in 1971, the Society for American Baseball Research (SABR) was founded, and the term "Sabermetrics" (statistics for baseball) was coined by Bill James when he began writing his annual *Baseball Abstracts* in 1977 [Puerzer, 2002].

While statistics are used in all sports to try to gain a competitive advantage, the appeal and relatively rapid adoption in Major League Baseball is likely the product of two major factors. The first factor is that, perhaps more than any other major sport, baseball comes down to a series of individual matchups between a hitter and a pitcher. While winning and losing is most definitely a team effort, to have a game decided by such controlled, individualized events makes it ripe for analytics. The second factor is that baseball is one of the few major American sports with no salary cap. The idea behind *Moneyball* was that small market teams like the Oakland Athletics, who in 2002 had a $44 million payroll, needed to exploit market inefficiencies in order to compete with large market teams like the New York Yankees, who in 2002 had a payroll over $125 million [Lewis, 2003].

While countless new statistics have become commonplace in a sabermetrician's lexicon

over the last few decades, the sabermetrics community still favors summary statistics over distributional statistics. The goal of this paper is to examine whether there is any additional utility, both in terms of comparative market value and in-game strategy, in understanding the difference between starting pitchers who are consistent (a low game-to-game variation in quality of performance), and pitchers who are inconsistent (a high game-to-game variation in quality of performance). To this end, logistic regression is employed to develop a measure of quality for each start, and analyze the start-to-start variation in starting pitchers from 1998 until 2017. Then, through the use of simulation techniques and bootstrap random sampling, the difference in actual and expected win percentages among starting pitchers with different distributional characteristics is explored. Finally, whether the use of distributional statistics, such as the probability of superiority, can be leveraged for in-game strategy is examined.

1.1 Data from Retrosheet

Creating distributional statistics requires access to the individual plays that represent every matchup between a hitter and a pitcher. Retrosheet is a website run by a team of volunteers that has compiled essentially every play in Major League Baseball since 1930. There are currently 30 teams in Major League Baseball, and each team plays 162 games a season. A single season can have almost 200,000 plays across all of the games. For the following analyses, research is limited to the 1998 season through the 2017 season. In 1998, Major League Baseball underwent an expansion to the current 30 team format with the introduction of the Arizona Diamondbacks and the Tampa Bay Devil Rays (later renamed as just the Rays), so this year serves as a good starting point for analysis. In all, there were 3,855,054 individual plays across 48,588 games in this time frame that were analyzed.

1.2 Background and Related Work

Traditionally, the primary measure used to evaluate a pitcher's quality is called their "Earned Run Average" (ERA). An ERA is essentially the number of earned runs a pitcher allows

normalized by 9 innings (the length of a game). For example, if a pitcher gives up 3 earned runs in 6 innings, that equates to a half of a run per inning, which would be 4.5 earned runs allowed per 9 innings (a 4.50 ERA). One of the primary concerns with this statistic is that there are multiple ways to achieve the same ERA. For example, a 4.50 ERA can also be the result of pitching 8 innings and giving up 4 earned runs. It is unclear if starts with these equivalent ERAs actually have the same effect on helping their team win. Another potential issue is that ERA is not robust to outliers so, when averaged out over multiple starts, ERA gives unequal weight to different starts because runs and innings pitched are simply considered on the aggregate. This means that one bad start can impact an ERA in a way that is not consistent with a pitcher's overall value to their team. While many of the ERA's sabermetric cousins, such as "Fielding Independent Pitching" (FIP) and "Skill-Interactive ERA" (SIERA), provide more predictive measures by taking out some of the randomness in ERA, they still suffer from these same basic concerns.

In 1985, a sportswriter for the *Philadelphia Inquirer* named John Lowe made the first attempt at looking at starts on a more individual basis with a metric that he called a "Quality Start" (QS) [Neyer, 2006]. A QS is a counting statistic that is awarded to a pitcher for every start he pitches at least 6 innings while giving up 3 or fewer earned runs. As mentioned, a start with 6 innings and 3 earned runs allowed yields an ERA of 4.50. However, a start with 9 innings and 4 earned runs allowed yields an ERA of 4.00, but does not qualify as a "Quality Start." From 1998-2017, the historical win percentage for starts of exactly 6 innings and 3 runs allowed was under 48%, while the historical win percentage for starts of exactly 9 innings and 4 runs allowed was about 66%. While the QS is certainly a step in the right direction in terms of accounting for start-to-start performance, it is clear that it still reflects an element of arbitrariness.

With so much publicly accessible data available, baseball is fortunate to have a deep community of sabermetrically inclined fans who publish opinions and research on numerous well-respected websites. One of those websites is called *Baseball Prospectus*, which was founded in 1996, and publishes daily articles online, as well as an annual book. In 2004, Michael Wolverton posted an article on *Baseball Prospectus* that focused on the relation-

ship between a starting pitcher's win/loss record and ERA. Much like ERA, conventional pitcher win/loss records have long been considered to be subject to too much random noise and factors that are out of the pitcher's control (like the defense behind him, or the runs his team scores, or even the ballpark he plays in) to be considered useful in player evaluations. In Wolverton's article, he proposes a set of alternative statistics he calls the "Support Neutral Win/Loss" (SNWL) and the "Support Neutral Value Added" (SNVA). Wolverton accounts for some of the problems associated with the conventional statistics by first adjusting for park and league scoring levels and considering all runs (earned or unearned) as equal, but the real "benefit of the Support-Neutral numbers is that they look at each start's contribution to winning individually rather than a season's run total cumulatively, so a single disastrous outing can't have the disproportionate impact that it can have on a starter's ERA" [Wolverton, 2004]. To illustrate this, Wolverton posits an example, reproduced here in Table 1.1.

Table 1.1: *Wolverton Example of Two Pitchers with Identical ERAs*

	Start 1	**Start 2**
Pitcher A	0 Innings Pitched, 10 Runs	8 Innings Pitched, 0 Runs
Pitcher B	4 Innings Pitched, 5 Runs	4 Innings Pitched, 5 Runs

The point of Wolverton's example is that innings pitched and runs allowed can be distributed across starts in many different ways. The simple fact that two pitchers have pitched the same number of innings and given up the same number of runs should not necessarily indicate that they both provided equal value to their respective teams. Wolverton explains:

> Their ERAs are equal, but Pitcher A's starts are likely to lead to more wins than Pitcher B's. An average team has a good chance of going 0-2 behind Pitcher B's two starts, but that same team is almost guaranteed to win Pitcher A's second game. The Support-Neutral stats account for the fact that the 10 runs

> concentrated in Pitcher A's one start don't do the same amount of damage as the ten runs spread among Pitcher B's two starts [Wolverton, 2004].

Wolverton, unfortunately, stops short of fully exploring this relationship between wins and start-to-start variation, instead focusing on a few specific examples for the given season he was analyzing. Another article, published in 2013 on *FanGraphs* by Matt Hunter, also explores differences from expected win percentages by using historical averages to do case studies of specific pitchers widely considered by baseball fans to be either consistent or inconsistent [Hunter, 2013]. Both authors reached the similar theoretical conclusion that it is better to be inconsistent than consistent. This paper further investigates this theory on a more generalized level.

Finally, on *FanGraphs* in 2015 (citing the Hunter article from 2013), Henry Druschel attempted to answer the question of whether or not consistency/inconsistency is even a trait that is predictable from year to year [Druschel, 2013]. That is to say, even if Wolverton and Hunter are correct in their hypotheses that there is more value added by inconsistent pitchers, the results would only be useful if consistency, or the lack thereof, is a non-random trait that certain pitchers exhibit more than others. Druschel used a metric developed by Bill James called the "Game Score" that essentially sums points assigned to certain aspects of a starter's performance (such as positive points for innings pitched and strikeouts, and negative points for runs allowed, walks, and hits) into one aggregate score using weights that tend to put things on a scale from 0 to 100 with a 50 being average. Looking at the standard deviation in average "Game Scores" from 2013 to 2014, Druschel claimed that the resulting correlation "is a pile of random points" with an "R^2 value that is basically 0," ultimately coming to the conclusion that, "while inconsistency is a hidden way for a pitcher's results to be better than they look, it doesn't appear to be a skill."

While this is a useful result, there are a couple factors that Druschel fails to consider when reaching his conclusion that consistency is not a predictable trait. The first is with regard to his use of "Game Score" as his measure of quality. "Game Score" intentionally includes inputs into its calculation that are fielding and park independent that a pitcher

largely controls himself, like walks and strikeouts, which would inherently reduce start-to-start variation. Druschel's analysis also only looks at the correlation between two seasons. In any given season, from 1998-2017, pitchers only start an average of about 16 games a season, and *at most* make 36 starts. It would not be surprising to observe large variances both within and between seasons given the small sample sizes that Druschel is dealing with.

1.3 Problem Statement

Specifically, this paper addresses the question of whether traditional statistics, like ERA, tend to undervalue high-variance pitchers relative to low-variance pitchers. This question can best be illustrated on a two-dimensional plane. On the horizontal (x-axis), there is some measure of a starting pitcher's average quality across all of their starts. On the vertical (y-axis), there is the start-to-start variation associated with that average quality. Assuming that quality and variance are monotonically increasing, at the origin there are starting pitchers that are consistently bad; their average quality is bad, and they have no start-to-start variation. On the other end of the x-axis, there are pitchers that are consistently good. The phrases "inconsistently bad" and "inconsistently good" are almost oxymorons. There is only one way for a pitcher to give his team, on average, a 0% or 100% chance to win, and that is by performing at that level in every one of their starts. On the other hand, pitchers who give their teams, on average, a 50% chance of winning can achieve this average in many different ways. In the middle, but still laying on the x-axis, are those starting pitchers that can be considered "consistently average" by giving their team exactly a 50% chance of winning every game. Alternatively, a starting pitcher could give their team a 0% chance to win in half of their starts, and a 100% chance to win in the other half of their starts, and still arrive at the same average 50% chance of winning. This last case represents the maximum variance possible, which will be discussed in more detail in Section 2.1. The result is a triangle that represents the entire spectrum of possible combinations of quality and consistency that starting pitchers can achieve, which is illustrated notionally in Figure 1.1.

Figure 1.1: *Theoretical Spectrum of Starting Pitcher Performance*

Consider a metric of quality that measures expected win percentage using traditional summary statistics that do not account for variation. Then, if one were to draw a straight line up from any point on the x-axis until it hit the side of the triangle, every pitcher that falls on that line would have the same expected win percentage. The question becomes, do any of the pitchers consistently outperform their expected win percentage in reality? It is obviously most preferable to be consistently good, and least preferable to be consistently bad, so this paper focuses on those pitchers that tend towards the average (represented by the yellow and purple sub-sections in Figure 1.1). Is it better to be a pitcher who is consistently average, or to be a pitcher who is sometimes above average and sometimes below?

CHAPTER 2

Methods

2.1 Deriving a Metric for Quality Using Logistic Regression

The first step in attempting to answer the question of whether consistency matters, is defining what is meant by "quality." Traditionally, Major League Baseball has used ERA, or more generally "Runs Allowed per 9 Innings" (RA/9) which includes both earned and unearned runs. Wolverton proposed a "Support Neutral" alternative, Hunter used historical win percentages, and Drushel borrowed Bill James's "Games Score." While all of these methods have their merits, in many ways they go too far in their adjustments to exclude factors that are beyond a pitcher's control. The components of RA/9, namely runs allowed and innings pitched, are factors that are well understood to directly relate to wins and losses. The problem is simply the way RA/9 aggregates and normalizes over 9 innings confounding the relative value of similar RA/9 over different numbers of outs recorded. Additionally, it is possible for a single game RA/9 to be infinite if the pitcher gave up runs and recorded no outs, so calculating variances would technically be impossible. Ideally, we could convert the components of RA/9 into a metric that gives each start an equal weight, and at the same time gives some relative value to starts such that two starts with different inputs, but similar impacts on wining, are close in scale. To derive this metric, this paper employs the use of logistic regression [Faraway, 2016].

Logistic regression is used when the dependent variable is dichotomous (i.e. wins versus losses). In this case, the relationship between the number of runs given up and the number of outs recorded by a starting pitcher, and whether or not the team ultimately ended up winning the game, is of interest. In order to handle the categorical dependent variable, it is

transformed using a "logit" link function which is the natural log of the odds that the team ultimately ended up winning the game.

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{2.1}$$

Equation 2.1 shows the form of the logistic regression with the "logit" link function. As mentioned, P in this case is defined as the probability that the team ultimately ends up winning the game, X_1 is the number of outs a pitcher records, X_2 is the number of runs that the pitcher allows, and ϵ is the error associated with the standard logistic distribution. The regression boils down to finding the β parameters that best fit the distribution of wins. Ultimately, the equation gives back coefficients in terms of log odds. Equation 2.2 illustrates how the log odds can then be combined and re-transformed to back out predicted probabilities.

$$P = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}} \tag{2.2}$$

One of the benefits of this approach is that the sum of all of the predicted probabilities will be exactly equal to the number of wins in the data. This approach can be thought of as dividing the wins across each start proportional to the relationships between the inputs and the actual results of the game. Traditionally, logistic regression is used for predictive analytics. In this case, the emphasis is on creating a metric that gives some idea of the relational quality of the starts that were used as inputs in the model. To achieve this, logistic regression is leveraged as a sort of transformation technique to convert outs recorded and innings pitched into a predicted winning percentage. That said, it could be dangerous to extend these coefficients into making predictions for future starts because a key assumption of logistic regression was violated. In this data, there are matched pairs of starters by each game. What this means is that the results are not independent of each other within each game because if one starter wins, it means that the other must lose. Table 2.1 shows an example of the output of the logistic regression (Expected Win Probability) next to historical win probabilities for starts made from 1998-2017.

Table 2.1: *Starts with an Expected Win Percentage Greater than 75%*

Outs Recorded	Runs Allowed	Expected Win Probability	Historical Win Probability
30	0	0.911	0.800
27	0	0.890	0.988
26	0	0.882	0.960
25	0	0.874	0.876
24	0	0.865	0.899
30	1	0.864	0.667
23	0	0.855	0.918
22	0	0.845	0.812
28	1	0.844	0.000
21	0	0.835	0.858
27	1	0.834	0.928
20	0	0.824	0.852
26	1	0.822	0.964
19	0	0.812	0.789
25	1	0.811	0.838
18	0	0.800	0.831
24	1	0.798	0.800
17	0	0.787	0.779
23	1	0.785	0.823
16	0	0.773	0.732
22	1	0.772	0.760
15	0	0.759	0.764
21	1	0.757	0.751
27	2	0.756	0.830

Looking at Table 2.1, the expected win probabilities and historical win probabilities are fairly consistent. However, there are a few examples that make it clear why it can be dangerous to simply use historical win probabilities. For example, there is very little reason why recording 25 outs and giving up no runs should result in a lower win probability (0.876) than recording 24 outs and giving up no runs (0.899) other than the fact that a starter would only pitch into the 9th and get pulled after recording one out if the game is close. Likewise, recording 28 outs and giving up one run is clearly not a game that the starter should have a 0% chance of winning. Using historical averages can leave you susceptible to small sample sizes and random noise. A typical game only has 27 outs, so it is not surprising that there is only one outing in the entire 20-year sample of data where a pitcher recorded exactly 28 outs and gave up one run. This small sample size issue is likewise the case with all observed starts with more than 27 outs recorded. Another observation that can be observed from the table is that giving up exactly one fewer run is roughly considered equivalent in terms of expected win probability to getting 6 more outs (2 more innings). As an example of this pattern, the last two records in the table show that the expected win probability of recording 21 outs and allowing one run is 0.757, while the expected win probability of recording 27 outs and allowing two runs is right behind at 0.756.

Now that a metric has been derived that gives equal weight to each start, and provides some level of meaningful comparison between starts with different characteristics, where pitchers fall on the theoretical spectrum of starting pitcher performance can be observed. The other benefit from using a metric that is on the scale of 0 to 1, aside from interpretability, is that the maximum variance can be easily computed. If X is always between 0 and 1, then it must be true that $X^2 \leq X$, and therefore $E[X^2] \leq E[X]$. The maximum variance possible given X then becomes:

$$Max\ Var[X] = E[X] - E[X]^2 = E[X] * (1 - E[X]) \tag{2.3}$$

Equation 2.3 then implies that the tip of the triangle will be at $0.5 * 0.5 = 0.25$. Technically speaking, the variances that will be calculated in this paper are sample variances and so 0.25

is not actually the maximum because we are dividing by $(N-1)$ rather than just N, but for illustration purposes it is a close enough approximation. Figure 2.1 shows how starting pitchers that made at least 36 starts (essentially one full season's worth of starts) from 1998-2017 fall on the spectrum. What becomes immediately clear is that, while the triangle represents the full range of theoretical outcomes, in reality starting pitchers at the major league level, both "good" and "bad," are a lot more similar than they are different.

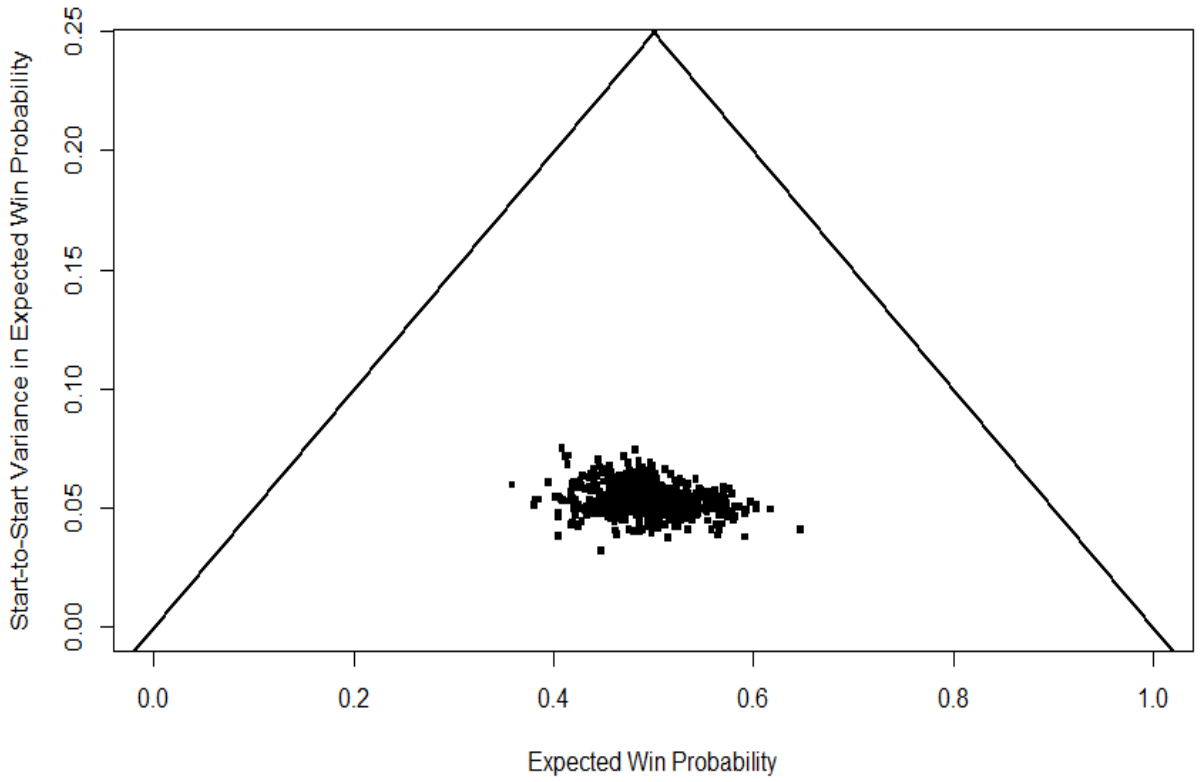

Figure 2.1: *Theoretical Spectrum of Starting Pitcher Performance Among Pitchers with At Least 36 Starts (1998-2017)*

It is now important to show that the derived metric for expected win probability is in fact correlated with actual win percentage. By construction, this should be the case, but that does not necessarily imply that the correlation is strong, or even meaningful. Figure 2.2 shows the proportion of actual wins versus expected win probability, as well as the line of

best fit between the two, for all 627 starters who made at least 36 starts from 1998-2017. The coefficient for the expected win probability is 1.02, which means that there is roughly a one-to-one relationship between actual win proportion and expected win probability. This value is statistically significant at the $\alpha = 0.05$ level of significance, providing evidence to reject the null hypothesis that there is not a linear relationship between a starter's expected win probability and their actual win percentage. The R^2 value of 0.372 can be interpreted to mean that 37.2% of the variation in actual win percentage can be explained just by the variation in the expected win probability.

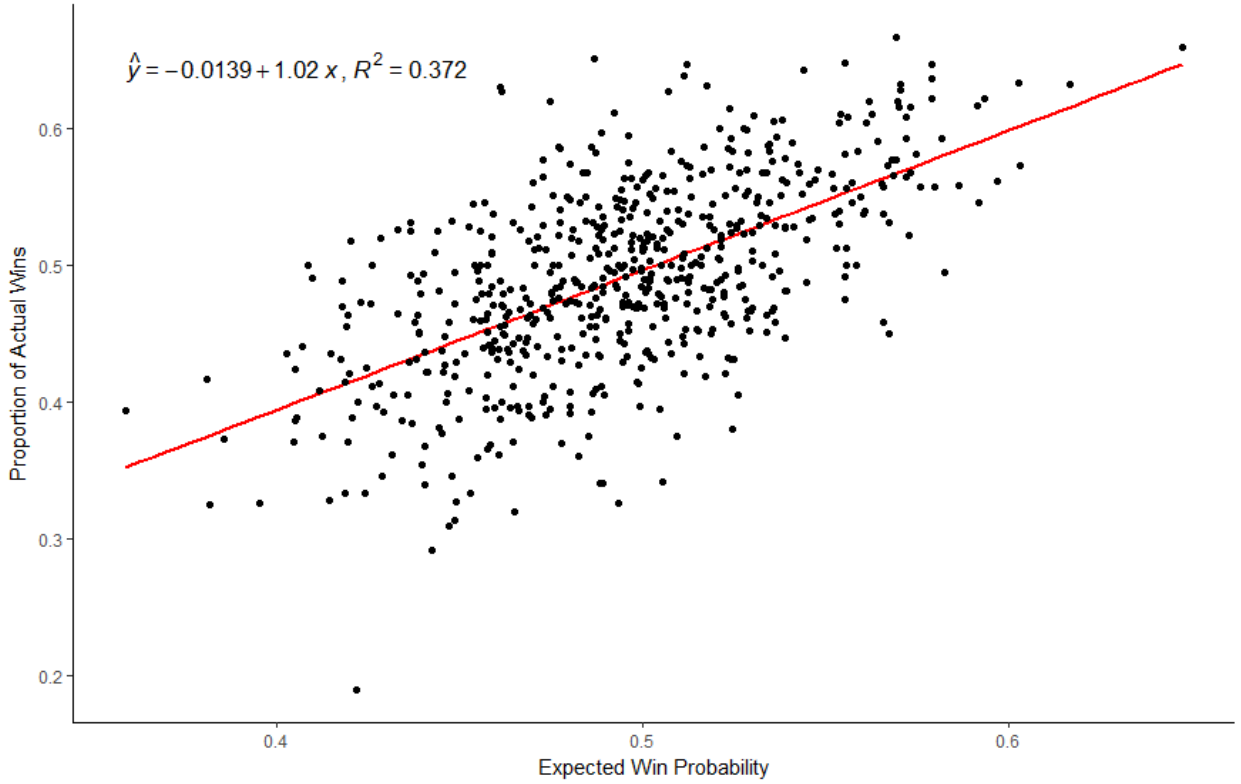

Figure 2.2: *Actual Versus Expected Win Probability Among Pitchers with At Least 36 Starts (1998-2017)*

The absolute correlation between RA/9 and expected win probability is 0.99, which further confirms the point that the logistic regression approach is simply transforming the components of RA/9 into a metric with more desirable properties. RA/9 has an R^2 value of 0.358 when regressing it on actual win percentage, meaning that just by transforming the

components of RA/9 using logistic regression explains an additional 1.4% of the variation in actual win percentage. The question is now boiled down to, can any of the additional variation observed in Figure 2.2 be explained by taking into account start-to-start variation? Put another way, does the inclusion of start-to-start variation significantly improve our ability to predict a pitcher's actual win percentage?

2.2 Simulating Theoretical Distributions with Bootstrap Random Sampling

In order to fully test the theory that inconsistent starting pitchers are more likely to outperform their expected win totals, bootstrap random sampling to simulate four hypothetical starting pitchers is employed [Efron and Tibshirani, 1994]. The advantage of this approach is two-fold. While innings and runs allowed are likely independent of team and ballpark effects over the long run, a starter's actual win percentage is not. Some starting pitchers may pitch their entire career for a perennial championship contender, while others may never come close to experiencing the postseason. Simply looking at the magnitude and significance of the regression coefficient for the start-to-start variation could obscure some of this dependency. If instead the entire pool of starts is considered and then randomly assigned to different pitchers, this dependency can be detached from individual pitchers. Essentially, a sort of controlled experiment can be artificially created by randomizing out the things that are not of interest, while fixing the things that are. In this case, the start-to-start variation is the parameter of interest, so if all starts from 1998-2017 are classified as either being "good," "bad," or "average" based on their expected win probability, what pools to sample from to get desired expected win probabilities and start-to-start variation can be controlled. The second advantage of this approach is that, given how close the observed starting pitchers are on the spectrum, the hypothesis that high-variance pitchers are more likely to outperform their expected win percentages can be tested on a more global and theoretical level by simulating pitchers with a more diverse set of characteristics.

For each of the simulated pitchers, 100 starts will be sampled, with replacement, from

specified combinations of different pools of starts (good, bad, or average) 100,000 times. For each of the 100 start samples, the difference between the pitcher's expected win totals and actual win totals will be calculated. After all 100,000 samples are collected, the resulting distributions can be visualized, and 95% confidence intervals can be approximated by looking at the 2,500th and 97,500th largest values. Figure 2.3 shows the average distributional characteristics of each of the four simulated pitchers across all of the bootstrap random samples. The red star in the bottom left-hand corner symbolizes a hypothetical pitcher that is consistently bad, the green star in the bottom right-hand corner symbolizes a hypothetical pitcher that is consistently good, the purple star in the middle at the bottom symbolizes a hypothetical pitcher that is consistently average, and the yellow star in the middle close to the top symbolizes a hypothetical pitcher that is inconsistently average. Through simulation, the observed games are stretched into distributions that inch closer to the theoretical edges of the triangle such that it will be easier to observe the relationship between start-to-start variance and the difference between actual and expected win percentages.

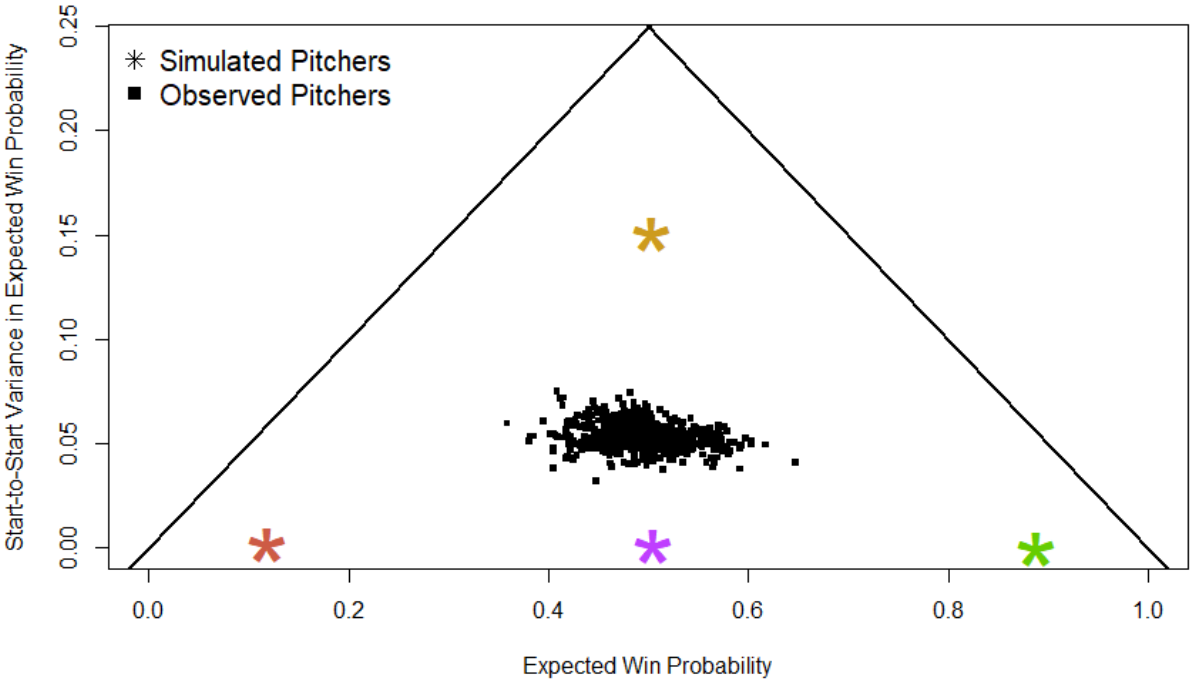

Figure 2.3: *Theoretical Spectrum of Starting Pitcher Performance with Simulated Pitchers*

2.3 Playing Matchups with the Probability of Superiority

While understanding the value of pitcher volatility on an aggregate level is the main focus of this paper, there is also the possibility that pitcher consistency has individual game utility. Consider, for example, the Major League Baseball (MLB) playoffs. There are six divisions in the MLB across two leagues. The winner in each division makes the playoffs, and then the two next best teams in each league earn the right to play each other in a one game wild card. It may be a simple enough solution for each team to just send out their best starter in terms of ERA, but is it possible that in certain situations understanding the starter's distribution of starts relative to their opponent's distribution of starts can provide an edge?

Consider you are the manager of a team in this one game wild card matchup facing a consistently good pitcher who always gives his team a 75% chance of winning. Your team's two best options both have an average win percentage of 50%, but one is perfectly consistent (all starts at 50%), while the other is perfectly inconsistent (half of his starts he gives his team a 0% chance of winning, and the other half he gives his team a 100% chance of winning). In this extreme scenario, the inconsistent pitcher actually has a 50/50 shot at posting a better start by expected win probability than the consistently good pitcher, as compared to the 0% chance that the consistently average pitcher has.

In 1992, a paper by Kenneth McGraw and S.P. Wong gave a name to this type of statistic called the "Common Language Effect Size," or the "Probability of Superiority." The probability of superiority is quite simply defined as the probability that a randomly selected observation from one distribution will be greater than a randomly selected observation from another distribution [McGraw and Wong, 1992]. For every matchup in the data, one can determine a pitcher's probability of superiority for that game based on each of the starters' distribution of the quality of their starts prior to that game. Whether these past distributions are predictive for the current start when considered together can then be examined.

For this paper, probabilities of superiority were calculated for every game in the data where both starters recorded at least 36 starts since 1998 prior to that game. For example, Figure 2.4 depicts the distributions for Mike Mussina's and Gustavo Chacin's 36 starts prior

to their matchup on April, 30, 2004. In the 36 starts prior to this game, Gustavo Chacin, of the Toronto Blue Jays, averaged a 0.521 expected win probability, compared to the virtually identical 0.520 expected win probability of Mike Mussina, of the New York Yankees. Despite having a slight edge in average expected win probability, Gustavo Chacin's probability of superiority was only 0.480. The overlayed histograms in Figure 2.4 show that Gustavo Chacin had a disproportionate number of starts that yielded an expected win percentage between 50% and 60% compared to Mike Mussina (9 starts to 3 starts). Meanwhile, Mike Mussina was much more variable, with 18 starts above a 60% win expectancy (compared to 13 starts for Chacin), and 15 starts below a 50% win expectancy (compared to 14 starts for Chacin). In the matchup between the two on April 30, 2004, Gustavo Chacin pitched six innings while allowing two runs (corresponding to a respectable expected win percentage of 60.3%), while Mike Mussina outperformed him with six innings pitched while allowing only one run (a 71.1% expected win percentage). The Yankees ended up winning that game by a score of 4-1. While this is only one example, and there are countless other factors at play, it illustrates the potential of considering distributional statistics for individual game matchups.

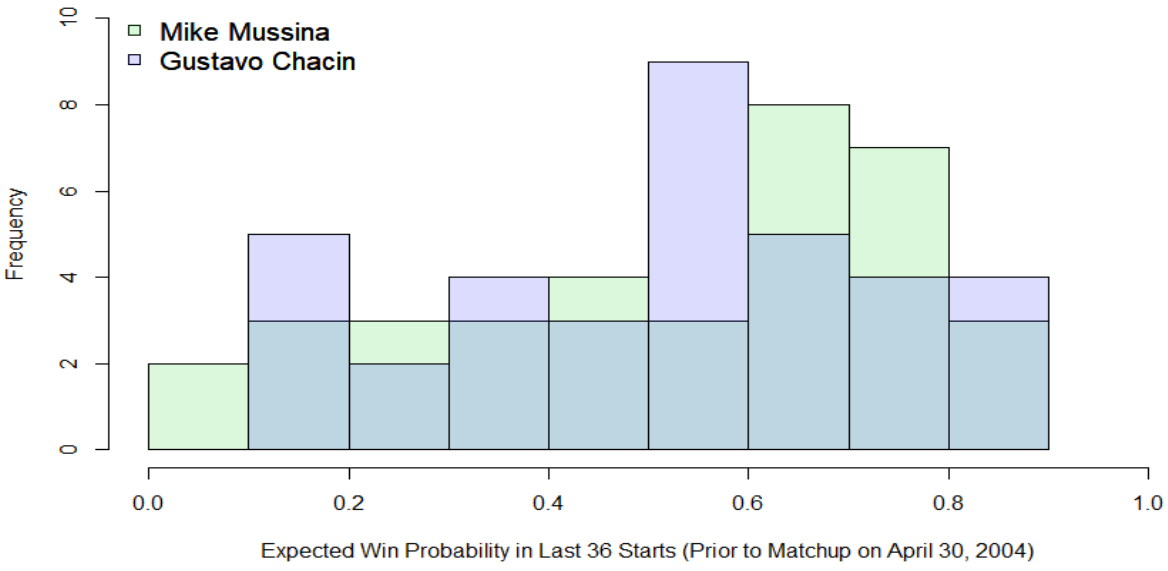

Figure 2.4: *Distributions of Expected Win Probabilities for Example of Matchup Where One Starter Has a Higher Average Expected Win Probability, but a Lower Probability of Superiority*

CHAPTER 3

Results

3.1 Evaluating the Relationship Between Pitcher Volatility and Actual Win Percentage

To investigate the hypothesis that high-variance pitchers systematically outperform similarly situated low-variance pitchers, recall the scatterplot in Figure 2.2. The difference between the points and the line of best fit are called the residuals (actual - predicted). These residuals represent the amount that the regression over-, or under-, predicted each starter's actual win proportion given their average expected win probability. If the hypothesis is true, high-variance pitchers would be consistently underpredicted (positive residuals) relative to low-variance pitchers, who would be overpredicted (negative residuals). Table 3.1 shows this relationship between the residuals shown in Figure 2.2 and pitcher variance for the 627 pitchers that made at least 36 starts from 1998-2017.

Table 3.1: *Regression of Residuals on the Standardized Percentage of Maximum Possible Variance Among Pitchers with At Least 36 Starts (1998-2017)*

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-0.0000	0.0022	-0.00	1.0000
Standardized % of Max Possible Variance	0.0048	0.0022	2.17	0.0303*

Recall that the spectrum of theoretical combinations of quality and consistency for pitchers (Figure 1.1) formed a triangle. This implies that the domain of variances is not constant across all average expected win probabilities. Equation 2.3 showed the formula for deriving

maximum variances given values on a scale from 0 to 1. Using this formula, each pitcher's observed variance can then be divided by their maximum possible variance, given their average expected win probability, resulting in the percentage of maximum possible variance for each pitcher. These values are then further standardized by subtracting them from the overall mean and dividing by the overall standard deviation for ease of interpretation.

Table 3.1 indicates that for a one unit increase in the standard deviation of maximum possible variance, the residual is underpredicted by an average of 0.0048 (or $\sim$0.5%). This increase, while modest, is significant at the $\alpha = 0.05$ level of significance. This means that the difference between actual win proportions and expected average win probabilities is in fact more positive (underpredicted) as start-to-start variance, relative to the maximum, increases. Put another way, pitchers with a higher start-to-start variance, relative to the maximum possible variance, tend to outperform their expected average win probabilities.

Having such a small relative effect is likely due to the fact that there simply is not a lot of variation between pitchers at the major league level after at least a season's worth of starts (Figure 2.1). So, while this approach generalizes the relationship between start-to-start variance and actual win percentage, the theory can be tested further using the bootstrapping technique described in Section 2.2 to simulate pitchers with much more distinguishable characteristics. Threshold values were chosen such that a "good" start was the largest subset of starts that averaged at least an 88% win expectancy, a "bad" start was the largest subset of starts that averaged at most a 12% win expectancy, and an "average" start averaged around a 50% win expectancy.

Table 3.2 shows these distributional summary statistics for the four simulated pitchers outlined in Section 2.2. For each of those four simulated pitchers, the average number of actual wins, the average number of expected wins, and the average start-to-start variation across the 100,000 samples was calculated. Of interest, is how the difference between the expected number of wins and the actual number of wins varies across the random samples. If the hypothesis that high-variance pitchers outperform similarly situated low-variance pitchers is true, the expected difference between actual and expected wins across the samples would be larger for the "Inconsistently Average" pitcher. Each sample consisted of 100 ran-

domly sampled starts for each of the simulated pitchers, and so for ease of interpretation the actual and expected counts in Table 3.2 can be considered as percentages.

Table 3.2: *Distributional Summary Statistics for Simulated Pitchers Across 100,000 Samples*

Simulated Starter Type	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Consistently Bad Actual Wins	2.00	11.00	13.00	13.54	16.00	30.00
Consistently Bad Expected Wins	9.98	11.65	11.97	11.97	12.29	14.11
Consistently Bad Variation	0.00	0.00	0.00	0.00	0.00	0.00
Consistently Good Actual Wins	89.00	97.00	98.00	97.83	99.00	100.00
Consistently Good Expected Wins	88.67	88.86	88.89	88.89	88.92	89.07
Consistently Good Variation	0.00	0.00	0.00	0.00	0.00	0.00
Consistently Avg. Actual Wins	25.00	45.00	49.00	48.80	52.00	70.00
Consistently Avg. Expected Wins	49.21	50.38	50.60	50.60	50.83	51.98
Consistently Avg. Variation	0.00	0.00	0.00	0.00	0.00	0.00
Inconsistently Avg. Actual Wins	45.00	54.00	56.00	55.69	57.00	69.00
Inconsistently Avg. Expected Wins	48.88	50.20	50.43	50.43	50.66	51.86
Inconsistently Avg. Variation	0.14	0.15	0.15	0.15	0.15	0.16

Table 3.2 confirms that the simulated distributions converged around the anticipated values for win expectancy and variation (essentially 0 for all but the "Inconsistently Average" pitcher). It can also be observed that for all but the "Consistently Average" pitcher, the actual win percentage converged to a higher number than the expected win percentage. This lends credence to the theory that being consistently average is the worst of the four major categories on the spectrum in terms of outperforming expected wins. Using the results of the simulation, the difference between actual and expected wins can be tested by calculating the difference in these averages within each of the 100,000 samples and taking the 2,500th and 97,500th largest values to get an approximation of the 95% confidence interval. If the confidence interval contains 0, the null hypothesis that the particular simulated distribution consistently results in more wins than expected would be rejected. Figure 3.1 shows the box-

plot of these differences for both the simulated "Inconsistently Average" and "Consistently Average" pitchers.

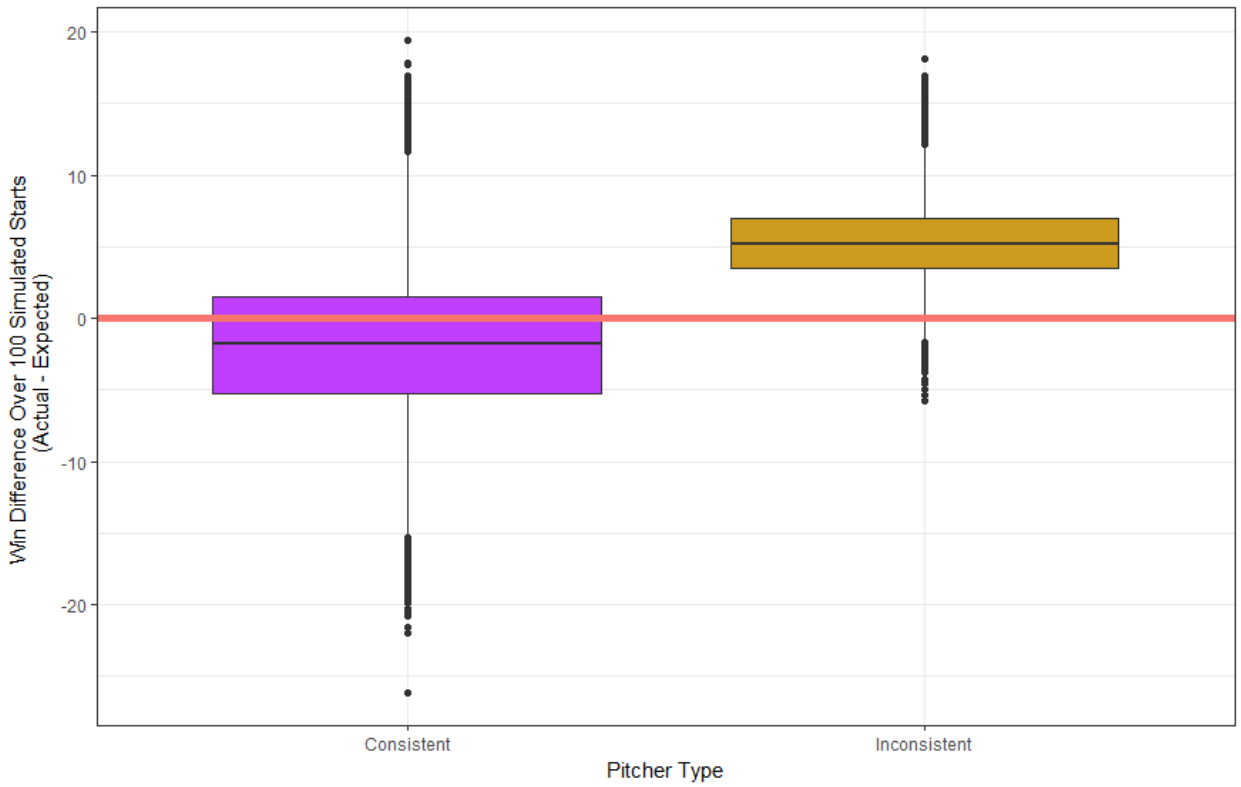

Figure 3.1: *Simulated Difference Between Actual and Expected Win Totals by Pitcher Type*

The results are seemingly paradoxical at first glance. The inconsistent pitcher appears to actually be more consistent in terms of his difference from expected, while the consistent pitcher has a much wider range of outcomes. In fact, the interquartile range (the spaces that are colored in Figure 3.1) for the consistent pitcher is almost double the size of the inconsistent pitcher (6.75 to 3.46, respectively). This phenomenon can be understood by thinking back to the extreme example of a pitcher with a 50% expected win probability with half of his starts yielding a 100% chance of winning, while the other half of his starts yield a 0% chance of winning. In this case, the difference between his actual and expected wins must be exactly 0 since he wins all of his good starts, but he has no chance to win any of his bad starts. On the other hand, the pitcher that gives his team exactly a 50% chance to win every game he pitches in could *conceivably* win or lose all of his starts. Basically,

by pitching at the extremes, inconsistent pitchers eliminate a lot of the luck associated with winning and losing. What this means is that, while providing more consistent results, an inconsistent pitcher's upside may actually be more limited compared to a similarly situated consistent pitcher.

While certainly an interesting result, the primary motivation behind this exercise was to see if the inconsistent pitcher significantly outperformed his expected number of wins. To this end, the 95% confidence interval for the difference in actual and expected wins for the inconsistent pitcher ranged from 0.38 to 10.59. Meanwhile, the 95% confidence interval for the difference in actual and expected wins for the consistent pitcher ranged from -11.54 to 8.00. This means that there is reason to believe that, at the $\alpha = 0.05$ level of significance, inconsistent pitchers systematically outperform their expected win totals while consistent pitchers do not.

While it is not surprising that the consistent pitcher had a much wider confidence interval, the fact that the intervals are not centered around the same point is important. This fact implies that there may actually be an advantage to pitching at the extremes, as was hypothesized. Figure 3.2 shows a 5,000 point moving average of actual and expected win percentages that can be used to investigate the hypothesis that there are certain levels of expected win percentage that consistently over-, or under-, estimate actual win percentage. To create this, all starts were first sorted from lowest to highest expected win percentage, and then the average expected win percentage for the first 5,000 observations was calculated. Then, moving up one observation at a time, the average expected win percentage was recalculated using the last 5,000 observations from that point. Similarly, the 5,000 point moving average of actual win percentage was calculated, and then paired to their associated average expected win percentage over the same set of observations. This smooths out the averages to help identify expected win percentages that are consistently above or below the actual win percentages observed.

Figure 3.2 shows that groups of starts that averaged a 25.4% win expectancy or lower had consistently higher actual win percentages. Likewise, groups of starts that averaged a win expectancy of 77.5% or higher had consistently higher actual win percentages. Mean-

while, groups of starts that had an average win expectancy between 44.5% and 66.4% had consistently lower actual win percentages.

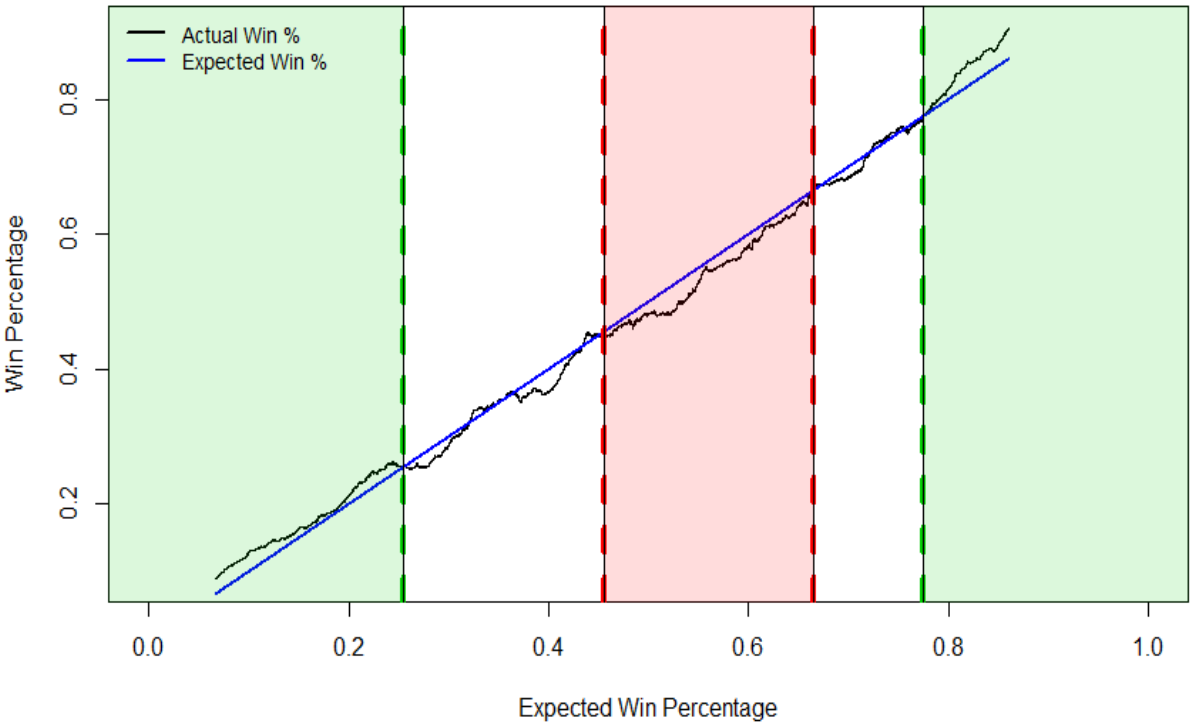

Figure 3.2: *5,000 Point Moving Average of Actual and Expected Win Percentages*

The three colored sections of Figure 3.2 illustrate exactly why the simulated consistently average pitcher was the only one of the group that did not outperform their expected wins. The thresholds for the "average" starts pool in the simulation were win expectancies between 44% and 56%, a range that is almost entirely within the red region of Figure 3.2. The "good" and "bad" pools, on the other hand, were entirely in the green regions. This means that there are thresholds at both the lower and upper tails where the return, in terms of actual wins, is greater than the cost of runs allowed and innings pitched. In other words, there is not the same "return on investment" across all average expected win probabilities. This pattern leads to the result that pitchers who are consistently average underperform relative to pitchers that pitch at the extremes.

3.2 Evaluating Whether Pitcher Volatility is a Predictable Trait

We are now armed with the knowledge that start-to-start variation explains a statistically significant, albeit relatively small, amount of the variation left over from the difference between actual and expected win percentages. It is next prudent to investigate whether pitcher volatility is a trait pitchers exhibit that can be relied upon as a factor when evaluating them. To this end, we can now look at whether a pitcher's start-to-start variance in their last 36 starts is predictive of start-to-start variance in their next 36 starts. Essentially, does a full season's worth of starts give any indication about what can be expected of them in terms of volatility in their next full season's worth of starts?

The predictability of start-to-start variation can be evaluated using linear regression. Framed in this way, a pitcher's last 36 starts will be the predictor, and the pitcher's next 36 starts will be the dependent variable. To achieve this, all of a given pitcher's starts from 1998 to 2017 will be split into a 36-point moving variance. A 36-point moving variance means that the start-to-start variance in a pitcher's first 36 starts will be the first predictor variable, and their variance in their next 36 starts will be the first dependent variable. Then, moving one game at a time, these variances will be recalculated and treated as new observations for inputs into the regression. If N_i is the number of starts that pitcher i makes from 1998-2017, then each pitcher with at least 72 starts (36 to predict for, and 36 to predict with) will have $(N_i - 71)$ different observations in this moving variance regression. Table 3.3 shows the results of this regression after adjusting for the maximum possible variance given the pitcher's average expected win probability in the respective 36-start grouping.

Table 3.3: *Regression of the Percentage of Maximum Possible Variance in Last 36 Starts on Percentage of Maximum Possible Variance in Next 36 Starts (1998-2017)*

		Regular		**Cluster-Robust**		
	Estimate	SE	t-value	SE	t-value	Pr(>|t|)
(Intercept)	19.1587	0.0965	198.60	0.4795	39.95	0.0000***
% of Max Possible Variance	0.0924	0.0045	20.56	0.0227	4.08	0.0000***

In linear regression, one of the primary assumptions is that each observation is independent of the next. Here, this assumption is violated as multiple observations per pitcher that are, by definition of a moving variance, correlated with each other have been introduced. This violation must be accounted for or else the confidence interval associated with the predictor will be too narrow, and thus the hypothesis test associated will lead to a conclusion of statistical significance too frequently.

To handle this violation, "Cluster-Robust Standard Errors" are used which allow correlation within a "cluster" of observations, which in this case are represented by each unique starting pitcher [Millo, 2017]. This means that within a pitcher the moving variances are allowed to correlate, while moving variances between pitchers are still considered independent. Table 3.3 shows the standard error (SE) and the t-value for both the regular regression and the "Cluster-Robust" regression. The t-value after adjusting the standard error drops from 20.56 to 4.08. Even after this adjustment, the corresponding p-value still implies that, at the $\alpha = 0.05$ level of significance, there is enough evidence to conclude that the percentage of the maximum possible variance in a pitcher's last 36 starts is a statistically significant predictor of the percentage of the maximum possible variance in the pitcher's next 36 starts.

Once again, while statistically significant, the magnitude of this impact is relatively small at 0.0924. This means that for a 1% increase in the percentage of maximum possible variance in a pitcher's last 36 starts, the percentage of maximum possible variance in the pitcher's next 36 starts would be expected to increase, on average, by roughly 0.09%. While nowhere near a one-to-one relationship, there does appear to be some positive linear relationship with start-to-start variation from one season's worth of starts to the next.

Due to the fact that the start-to-start variances themselves are small in magnitude, and the fact that pitchers in reality tend to be fairly clumped together on the spectrum, one other approach is to do an analysis of variance (ANOVA) [Dobson and Barnett, 2008]. An ANOVA tests for the difference between different group means by analyzing the within group variance and the between group variance relative to each other. If the variance between groups is small relative to the within group variance, then there is not enough evidence to suggest that the group means themselves differ from each other. On the other hand, if the between

group variance is much larger than the within group variance, it is reasonable to conclude that the group means are in fact different. In other words, groups are considered different if the observed variance between group means is much larger than the variance exhibited between observations within groups.

Put into the context of this paper, if consistency is indeed a predictable trait, the year-to-year variance in start-to-start variance within pitchers would be small relative to the variance between the mean year-to-year variance in start-to-start variance between pitchers. Table 3.4 shows the results of the ANOVA after dividing each pitcher's starts into groups of 36 among pitchers who made at least 72 starts from 1998 to 2017. In this case, starts are simply ordered by date and split into groups of 36, so there is no overlap of starts like there is with the moving variance approach.

Table 3.4: *Analysis of Variance for the Mean Percentage of Maximum Variance Possible by Every 36 Starts Among Pitchers Who Made at Least 72 Starts (1998-2017)*

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Between Pitcher	412	1.02	0.00	1.32	0.0001***
Within Pitcher	1904	3.57	0.00		

Table 3.4 indicates that the variance between pitchers was significantly larger than the variance within pitchers. Broken down, the "Mean Sq" column is the "Sum Sq" column divided by the "Df" (degrees of freedom) column, and the "F-value" test statistic is simply the discussed ratio of the mean square of the "Between Pitcher" and the "Within Pitcher" mean square. The exact conclusion that can be drawn from an ANOVA is that at least one pitcher had a sample mean percentage of maximum variance possible that was statistically different than the others, at the $\alpha = 0.05$ level of significance. While the ANOVA does not allow for a general conclusion, it does at least indicate that start-to-start variation is not an entirely random process, and that for at least for some pitchers there is a consistency of variance relative to the others.

3.3 Investigating the Market Value of Pitchers by Start-to-Start Variation

Now that we have explored the effects of start-to-start variation on actual win percentages, and determined that there is at least some degree of predictability of start-to-start variance within pitchers, we can turn our attention to how variation effects a pitcher's monetary value. To analyze this, we can focus on total earnings over the 20 years covered by the data from 1998-2017. Using this as the dependent variable, a model can be constructed to make an inference on the effect start-to-start variation had on total earnings over this period.

The salary that a player earns, especially over the course of multiple seasons, can be highly dependent on unique circumstances. Quality of performance, injury history, the year a player hits free agency, or whether a player values length over average annual salary in a contract can all dramatically affect how much a player makes over a given time frame. As such, building a model to predict career earnings would be an onerous task. In inferential statistics, we only wish to determine the relationship between a predictor variable and the dependent variable. To make this inference, the model only needs to account for factors that are correlated with both the predictor of interest and the dependent variable. In this case, how start-to-start variation impacts a pitcher's earnings is of interest. Therefore, everything that has an effect on both start-to-start variation and total earnings needs to be considered. Additional factors may increase the overall predictive power of the model, but if the factors do not correlate to start-to-start variation then that relationship and its corresponding inference will remain unchanged. For this exercise, the analysis is limited to the 469 pitchers with salary data who debuted after the start of 1998 and made at least 36 starts.

Once again using the standardized percentage of maximum possible variance as the predictor of interest, variables with which it might be correlated were identified. To start, it is still likely that there is some relationship between average expected win probability and start-to-start variation even though it has already been converted to a percentage of the maximum. Indeed, the correlation between the two variables is roughly -0.263, meaning that as average expected win probability increases, a pitcher is more likely to be consistent

even relative to the maximum. Also, while the sample is limited to only pitchers who made at least 36 starts, it is still expected that pitchers closer to that minimum may exhibit more of the extreme variances observed in the data before it has fully stabilized. The correlation between these two variables is roughly 0.116 and does in fact display the pattern of dispersed variances at low numbers of starts before becoming more dense at higher numbers of starts.

Particularly with salary data, it is also important to check for any necessary transformations before modeling. Often, salaries are right-skewed (most people earn relatively little, while a few people make a lot). Figure 3.3(a) shows that the earnings in this data are indeed right-skewed, and Figure 3.3(b) shows these earnings with a logarithmic transformation applied to help normalize the distribution.

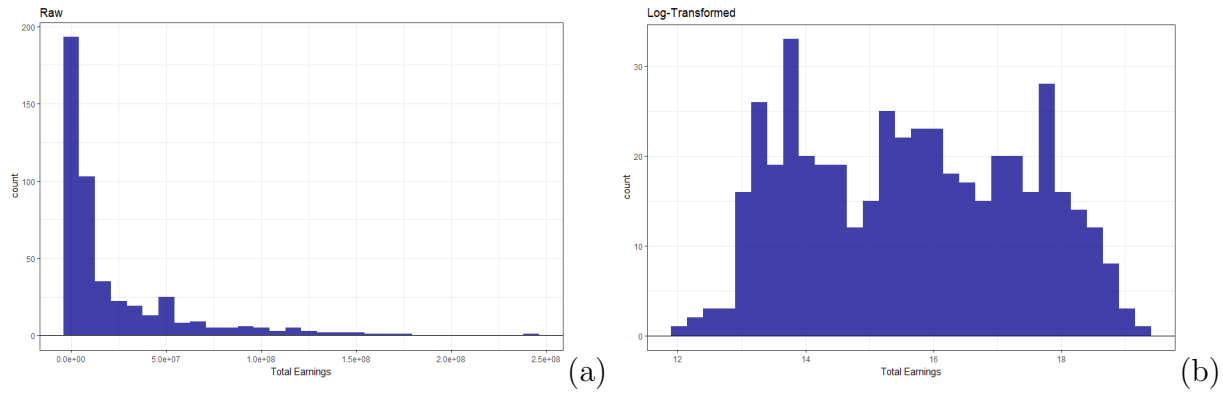

Figure 3.3: *Distribution of Total Earnings (1998-2017)*

Figure 3.4 illustrates the relationships between the log-transformed total earnings and the predictor variables identified. As shown in Figure 3.4(a), the relationship between log total earnings and average expected win probability is positive and linear as anticipated. Figure 3.4(b) shows that the number of starts that a pitcher makes appears to have a clear quadratic relationship to the logarithm of total earnings, meaning that it increases fast at low quantities of starts before tapering off as the pitcher makes more starts. When the dependent variable is log-transformed, the log-linear coefficients can be approximately interpreted as percent changes when the coefficient is small. Therefore, the resulting quadratic relationship between number of starts and total earnings makes intuitive sense as early on a player should

have a larger percent increase in total earnings before the additional yearly salary associated with more starts only becomes a small fraction of the total earnings. Finally, Figure 3.4(c) shows a parabolic relationship between log total earnings and the standardized percentage of maximum possible variance. For average percentages of maximum possible variance (those around 0), there appears to be no discernable pattern with regard to log total earnings, but for standardized percentages of maximum possible variance that are large in magnitude, there appears to be a negative correlation with log total earnings.

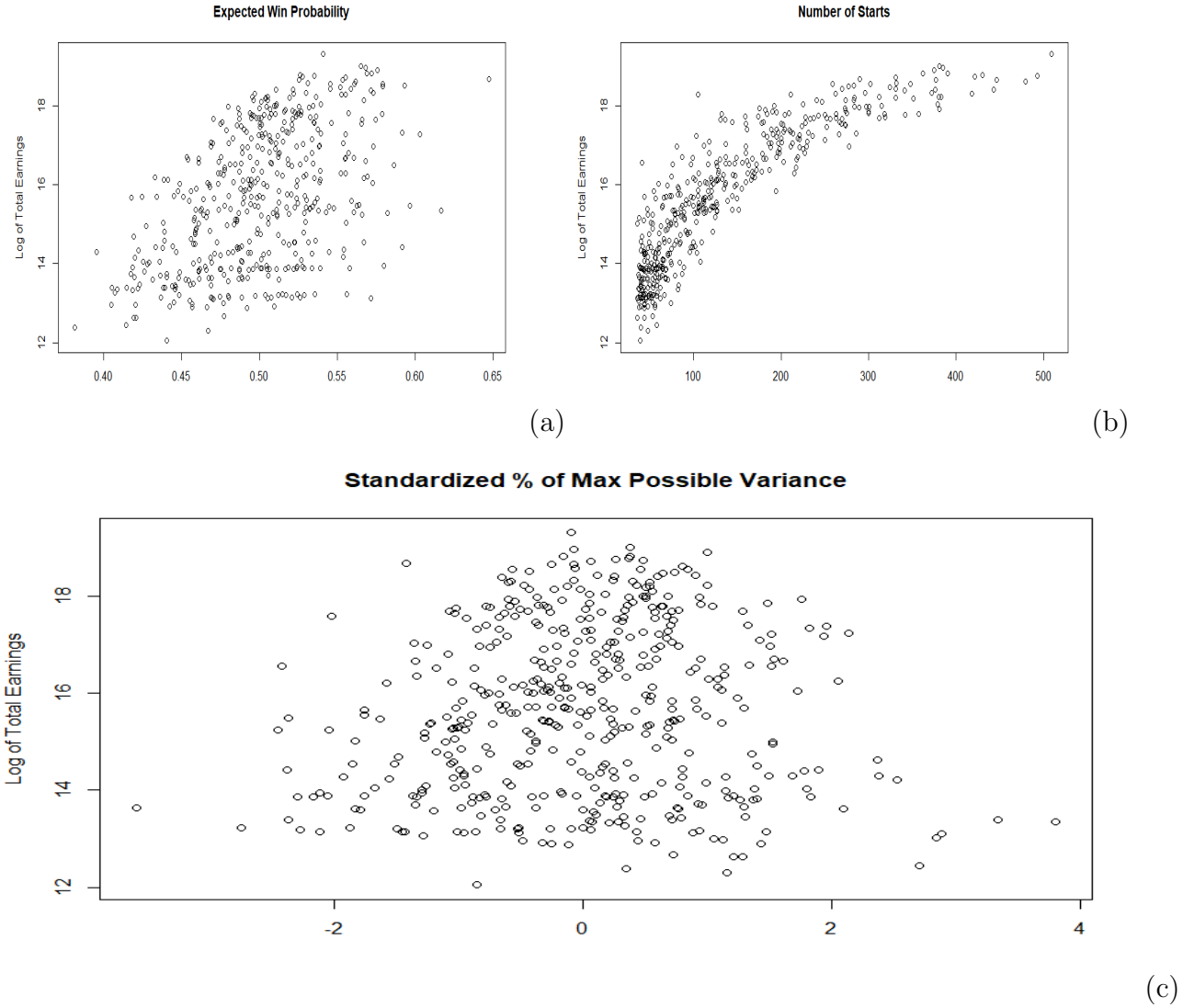

Figure 3.4: *Scatterplots of Predictor Variable Relationships to Log Total Earnings (1998-2017)*

Not considering any of the other relationships at the moment, the form displayed in Figure 3.4(c) implies that perhaps starting pitchers with average variability were actually being valued more fairly than either inconsistent or consistent pitchers during this time frame. Table 3.5 shows the results of the regression that considers all of these factors together.

Table 3.5: *Regression on Log Total Earnings (1998-2017)*

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	9.7623	0.4142	23.57	0.0000***
Avg. Expected Win Probability	5.9390	0.8670	6.85	0.0000***
Number of Starts	0.0296	0.0011	25.81	0.0000***
(Number of Starts)^2	-0.0000	0.0000	-14.59	0.0000***
Standardized % of Max Possible Variance	0.0175	0.0330	0.53	0.5971
(Standardized % of Max Possible Variance)^2	-0.0131	0.0194	-0.68	0.4991

The resulting model has an R^2 value of 0.8529, meaning that it explains roughly 85.3% of the variation in total earnings for this subset of pitchers. Average expected win probability, number of starts, and the number of starts squared are all statistically significant at the $\alpha = 0.05$ level of significance. On the other hand, even with a relatively simple model, the standardized percentage of maximum possible variance is not statistically significant at any meaningful level of alpha. Given that the model already explains over 85% of the variation in total earnings, even if some variables were omitted that would correlate with the variance and change the magnitudes of the related coefficients, it is unlikely to change the sign or lack of significance of the results.

While not statistically significant, the relationship between variance and total earnings can still be explored. If the average expected win probability and the number of starts is fixed, then total earnings can be predicted across varying levels of the standardized percentage of maximum possible variance. Figure 3.5 tells us that, regardless of the actual dollar amounts, pitchers that are 0.67 standard deviations above the group mean percentage of maximum possible variance are expected to have, on average, the maximum total earnings

among all possible standardized percentages of maximum possible variance for any fixed average expected win probability and any fixed number of starts. If anything, it appears that perhaps consistent pitchers in this subset were valued less on the market relative to average and high-variance pitchers.

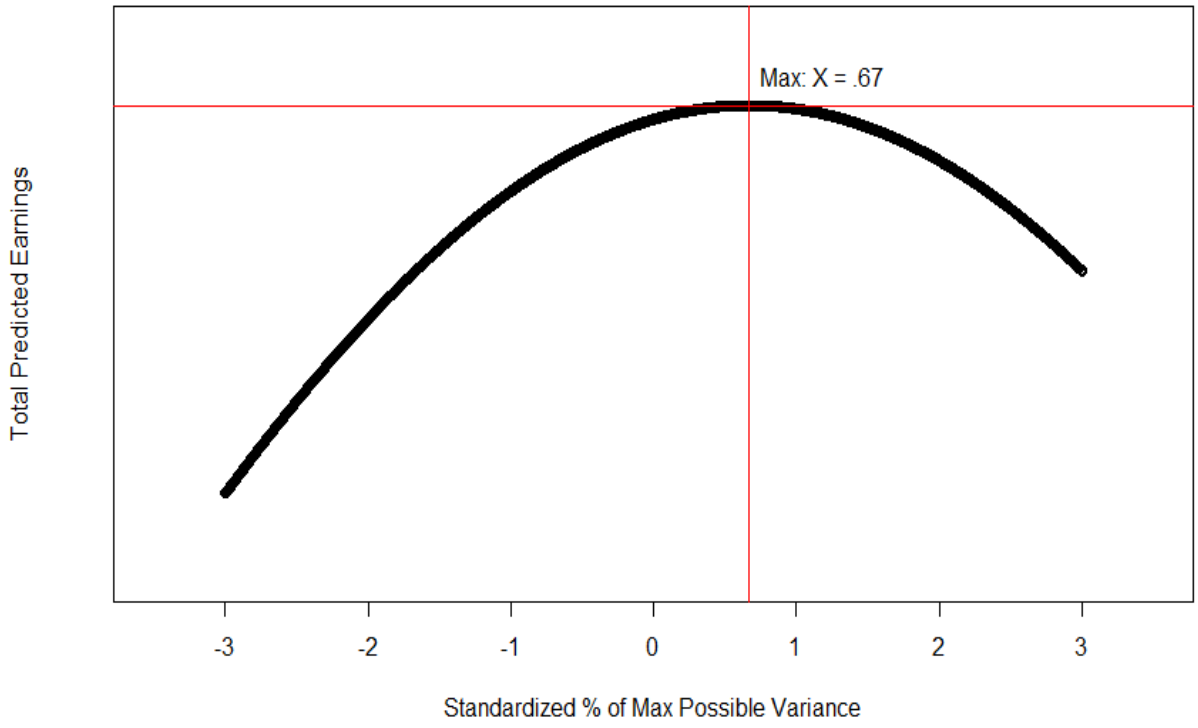

Figure 3.5: *Predicted Total Earnings Over Varying Standardized Percentages of Maximum Possible Variance*

3.4 Determining the Value of Distributional Statistics for Evaluating Pitching Matchups

Up to this point, the patterns and impact of start-to-start variation on winning has been analyzed on a global level across the full dataset. In Section 2.3, the concept of the "Probability of Superiority" was introduced. The probability of superiority was defined as the probability

that a randomly selected start from one pitcher resulted in a larger expected win probability than a randomly selected start from another pitcher. The question is whether the probability of superiority can be used as a way to compare individual pitchers and potentially make decisions regarding pitching matchups in particular games.

To investigate if the probability of superiority is indeed a useful metric in determining pitching matchups, the probability of superiority is first calculated for every game where both starters recorded at least 36 starts prior to the game in question. Each eligible game is then split into two categories. The first category is games where one of the starting pitchers had a probability of superiority over 50%, but a lower average expected win percentage in their last 36 starts than their opponent. The second category is games where the starting pitcher had a probability of superiority less than or equal to 50%, and a lower average expected win percentage in their last 36 starts than their opponent.

In both categories described above, the starting pitcher being analyzed has a lower expected win percentage than his opponent. If taking into account the probability of superiority is indeed a useful metric for evaluating pitching matchups, then it would be expected that the group that has a probability of superiority over 50% would significantly outperform the group that does not. As it turns out, there are only 199 games over the 20-year sample where a pitcher had a probability of superiority over 50%, but a lower average expected win percentage in their last 36 starts than their opponent. Even within these 199 games, the difference in average win percentage between the two starters is never more than 2.7%. Once again, in reality we see that pitchers are much more similar than they are different at the major league level. Nonetheless, in these 199 games the starter with the probability of superiority over 50% recorded the better start 50.8% of the time despite having a lower average expected win percentage than their opponent. Over this same period, there were 6,083 games where the starter had a probability of superiority less than or equal to 50%, but an average expected win percentage less than 2.7% lower in their last 36 starts than their opponent. This 2.7% cutoff is used to make sure that the two groups compare starts where the pitchers are in relatively similar standing with the only major difference being the probability of superiority. In these 6,083 games, the starter with the probability of superiority

less than or equal to 50% posted a better start 46.6% of the time.

In summary, among games where the pitcher had an average expected win percentage less than 2.7% lower than their opponent in their last 36 starts, pitchers with probabilities of superiority over 50% posted a better start in terms of expected win percentage more often than pitchers with probabilities of superiority less than or equal to 50%. The resulting 4.2% difference in favor of those pitchers with a probability of superiority over 50% can be tested more formally using a one-tailed Chi-squared test of proportions. The corresponding p-value of 0.1214 indicates that the 4.2% difference is not statistically significant at the $\alpha = 0.05$ level of significance. Therefore, there is not enough evidence to support the claim that having a probability of superiority over 50% results in the pitcher posting a better start than their opponent in matchups where the pitcher has an average expected win percentage less than 2.7% lower than their opponent in their last 36 starts.

CHAPTER 4

Conclusion

On a general level, across the entire 20-year sample, high-variance pitchers systematically overperformed the results that would be expected looking at only runs allowed and innings pitched. While passing the statistical significance threshold at the $\alpha = 0.05$ level of significance, this effect may not be practically significant given that win percentage increases by only 0.5% for every standard deviation above the group mean percentage of maximum possible variance. For example, even if a pitcher is two standard deviations above the group mean percentage of maximum possible variance, then over 100 starts they would be expected to win on average only one more game than a pitcher with the same average expected win percentage, but a percentage of maximum possible variance that is only average. The cause behind such a small effect became evident after plotting actual pitchers on the theoretical spectrum of starting pitching quality in Figure 2.1. While every coordinate within the triangle is theoretically possible, in reality starting pitchers at the major league level are much more similar than they are different.

Nonetheless, even if only marginally, the theory that high-variance pitchers outperformed similarly situated low-variance pitchers held true in practice. In order to fully understand the relationship between start-to-start variation and actual win percentage, a technique was devised to simulate four hypothetical pitchers by bootstrap sampling from different pools of starts to fix average expected win percentages and start-to-start variation, while randomizing out the other effects that were not of interest. Figure 2.3 showed how the resulting four simulated pitchers stretched towards the theoretical edges of the triangle, allowing for comparisons of pitchers with much more distinct characteristics. Testing for the difference in actual versus expected win totals between the "Inconsistently Average" and

"Consistently Average" pitchers yielded results consistent with the hypothesis that high-variance pitchers outperformed similarly situated low-variance pitchers. Figure 3.1 illustrated that, while the "Consistently Average" pitcher had a much wider range of possible outcomes, the "Inconsistently Average" pitcher systematically outperformed his expected win total.

Once it was shown that high-variance pitchers outperformed similarly situated low-variance pitchers both in theory and in practice, the next phase investigated whether start-to-start variance is a predictable trait. Using a moving variance regression, as well as an analysis of variance, it was shown that start-to-start variation between groups of 36 starts were predictive of each other. Essentially, pitchers exhibited an element of consistency in start-to-start variation between years relative to the differences in yearly start-to-start variation between each other. So while the linear effects of the start-to-start variation between the groups of starts may not have been one to one, there was still statistically significant evidence that consistency is in fact a pattern that pitchers display.

Knowing that start-to-start variation is predictive from year to year among pitchers, and there is a pattern of underprediction of actual win totals among high-variance pitchers, the analysis next focused on the impact of start-to-start variance on total earnings. Looking only at pitchers that debuted after 1998 and made at least 36 starts, a pitcher's percentage of maximum possible variance was not found to have a statistically significant effect on total earnings. Given the marginal effect found in reality regarding the underprediction of actual win totals among high-variance pitchers, it is unsurprising that there was not a significant difference in total earnings between low-variance and high-variance pitchers. In fact, if anything, Figure 3.5 showed that, while consistent pitchers were less likely to outperform their expected win total, they were perhaps not given enough credit for their wide range of possible outcomes. For a team that is confident in their offensive output, a pitcher that is consistently average may actually represent a value on the market.

Finally, an analysis of whether taking into account a pitcher's distribution of starts relative to their opponent's distribution of starts could be used by managers to decide potential pitching matchups was conducted. Using a metric called the "Probability of Superiority," it was shown that pitchers that held the edge in this distributional statistic were more likely

to post a better start in terms of runs allowed and innings pitched than their opponent even if the pitcher had a lower average expected win percentage than their opponent in their last 36 starts. Unfortunately, only 199 games in the 20 year sample involved a pitcher with a probability of superiority over 50%, but a lower average expected win percentage than their opponent in their last 36 starts. As such, the resulting 4.2% difference in favor of those pitchers with a probability of superiority over 50% was not statistically significant.

In all, the hypothesis that high-variance pitchers are more likely to outperform their expected win totals than low-variance pitchers proved true in both theory and reality. However, given how similarly major league pitchers have performed over the last 20 years in terms of start-to-start variation, these differences are arguably practically insignificant in terms of their potential value in driving financial and in-game decisions. That said, Tommy Lasorda, a two-time World Series champion as the manager of the Los Angeles Dodgers, famously observed, "No matter how good you are, you're going to lose one-third of your games. No matter how bad you are you're going to win one-third of your games. It's the other third that makes the difference." With such a small margin distinguishing "good" teams and "bad" teams, even an advantage of only a game or two can make all the difference. As such, knowing that the inclusion of start-to-start variation significantly improves our ability to predict actual win percentage, even modestly, is still a valuable finding. It is also possible that one of the reasons why pitchers at the major league level have such similar start-to-start variation is because inconsistent pitchers are so undervalued that they are not given an opportunity to pitch enough games to be analyzed, which if true would imply that savvy teams could unlock meaningful value by taking chances on pitchers who had otherwise been considered too erratic.

CHAPTER 5

Limitations and Future Work

The intent of this paper was to examine the hypothesis that high-variance pitchers are more likely to outperform their expected win totals than similarly situated low-variance pitchers on a general level. As such, this paper focused on observable patterns across large swaths of data that included multiple years and multiple pitchers. While on this global level the results were marginal, now that the theory has been shown to be true, there are potential applications on a more micro scale. Much like what was shown in the case-study work done by Wolverton and Hunter, start-to-start variation may still be the explanation for why individual pitchers seem to consistently outperform their expectations. As such, monetary and in-game decisions on an individual pitcher basis may still provide an advantage. Once theories have been proved generally, specific analyses hold more weight as being plausible factors to consider.

Although it is true that many things in baseball can be broken down into discrete events between pitchers and hitters, wins and losses are made up of so many of these events that they are still subject to a lot of noise when it comes to prediction. The models used in this paper are simply a starting point meant to be built upon and fine-tuned. One example of a factor not considered when building the metric for expected win probability was the year-to-year change in league offensive output. Over the course of baseball history, offensive production has had its ebbs and flows (the "Steroid Era," the "Dead Ball Era," and most recently the "Juiced Ball Era"). In some years, a start of 6 innings and 3 runs allowed is a lot better comparative to the rest of the league than in other years. For this paper, this effect was omitted because the goal was simply to divide wins as evenly as possible assuming all else was equal other than the inputs to RA/9.

Another area for further exploration is with regard to evaluating market value. Salary data is hard to come by, and identifying years where players were eligible for free agency, and what types of idiosyncrasies were built into contracts that gave a player more value than simply the dollar amount, makes modeling difficult. As such, the total earnings model employed in this paper is aggregate and very general. While the model only intends to make an inference about the relationship between total earnings and start-to-start variation, the absence of all variables that correlate with both factors can lead to what is called "omitted variable bias." While this paper addresses why it is believed that the conclusions would remain the same, there are many more potential factors that certainly would effect total earnings that could also possibly be correlated with start-to-start variation.

Perhaps the biggest area ripe for further exploration is how distributional statistics effect relief pitchers. Relief pitchers are the pitchers that come into a game after the starter leaves, and as such they have many more appearances, but far fewer innings, than starters. While the variation between starters in reality was too small to see much of an effect, outing-to-outing variation between relievers is likely far greater in magnitude due to their limited number of innings pitched. ERAs in particular for relievers are highly volatile as one bad outing can have a larger impact on their average since it is normalized by 9 innings. Imagine a situation where a reliever is asked to preserve a one run lead. Just as was discussed with starting pitchers, it might be best to simply use the team's best reliever in terms of ERA. However, consider one option is to use a reliever who 75% of the time gives up no runs, but the other 25% of the time he gives up two runs. That reliever's ERA may actually be less than a reliever that half of the time gives up no runs, but the other half of the time gives up one run. If the team only cares about preserving the one run lead, then it may be advantageous to consider the reliever's distribution of outings and choose the one that gives up no runs 75% of the time. On the other hand, if the team had a two run lead, the reliever that never gives up more than one run would be the smart choice.

REFERENCES

[Dobson and Barnett, 2008] Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. CRC Press, Taylor & Francis Group, third edition.

[Druschel, 2013] Druschel, H. (2013). The value and consistency of pitcher inconsistency. *FanGraphs*.

[Efron and Tibshirani, 1994] Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

[Faraway, 2016] Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press Taylor & Francis Group, second edition.

[Hunter, 2013] Hunter, M. (2013). Finding value in pitcher inconsistency. *FanGraphs*.

[Lewis, 2003] Lewis, M. M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton.

[McGraw and Wong, 1992] McGraw, K. O. and Wong, S. P. (1992). A common language effect-size statistic. *Psychological Bulletin*, (111):361–365.

[Millo, 2017] Millo, G. (2017). Robust standard error estimators for panel models: A unifying approach. *Journal of Statistical Software*, 82(3).

[Neyer, 2006] Neyer, R. (2006). Quality start still a good measure of quality. *ESPN*.

[Porter, 1984] Porter, M. (1984). The pc goes to bat. *PC Magazine*, page 209.

[Puerzer, 2002] Puerzer, R. J. (2002). From scientific baseball to sabermetrics: Professional baseball as a reflection of engineering and management in society. *NINE: A Journal of Baseball History and Culture*, 11(1):34–48.

[R Core Team, 2013] R Core Team (2013). R: A language and environment for statistical computing. `http://www.R-project.org/`.

[Wolverton, 2004] Wolverton, M. (2004). Baseball prospectus basics: The support-neutral stats. *Baseball Prospectus*.