UNIVERSITY OF CALIFORNIA

Los Angeles

A Player Based Approach to Baseball Simulation

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Statistics

by

Adam Philip Sugano

2008

The dissertation of Adam Philip Sugano is approved.

Jan de Leeuw

Rick Paik Schoenberg

Hal Stern

Mark Hansen, Committee Co-Chair

Don Morrison, Committee Co-Chair

University of California, Los Angeles

2008

To my favorite statistician and my Dad, Dr. David S. Sugano,

with love and admiration…

TABLE OF CONTENTS

1 Introduction.. 1

2 Baseball within a Markov Chain Framework.. 5

2.1 Markov Chain Properties... 6

2.2 Previous Work Modeling Baseball via Markov a Markov Chain.... 8

2.3 The Transition Matrix.. 9

2.3.1 Run Potentials.. 12

2.3.2 Transition Constraints.. 13

2.3.3 Finalized Transition Matrices... 14

2.4 Advantages of Simulation Models... 23

3 Situational Factors Affecting Transition Probabilities......................... 25

3.1 Bias and Evolvement of Baseball Statistics................................... 26

3.2 Measuring Chance vs. Ability.. 28

3.3 The Log5 Method... 31

3.3.1 Extending the Log5 Method... 32

3.3.2 Alternative Methods... 34

3.3.3 Comparing Simulated Results.. 36

3.4 Situational Affects.. 38

3.4.1 Handedness... 40

3.4.2 Home vs. Away... 40

3.4.3 Clutch Hitting ... 42

3.4.4 Streakiness ... 43

3.4.5 Momentum ... 45

3.5 Conclusions About Situational Affects in Baseball ... 46

4 Batter-Pitcher Matchups in Baseball and Small Sample Decision Making ... 47

4.1 The Batter-Pitcher Matchup: A Simple Binomial View ... 49

4.2 A Hierarchical Model for Batter-Pitcher Matchup Data ... 51

4.2.1 Data for a Single Player ... 51

4.2.2 A Probability Model for Batter-Pitcher Matchups ... 52

4.2.3 Results – Kenny Lofton ... 55

4.2.4 Results – Derek Jeter ... 56

4.2.5 Results – Multiple Players ... 60

4.3 Batter-Pitcher Data from the Pitcher's Perspective ... 61

4.3.1 Results – Tim Wakefield ... 62

4.3.2 Results – Mike Mussina ... 65

4.3.3 Results – Multiple Players ... 67

4.4 Towards a More Realistic Model ... 68

4.5 Discussion ... 69

5 Methods for Predicting Individual Performance.................................. 72

5.1 The Aging Process: A Multilevel Modeling Approach.................. 73

5.1.1 Variance Decomposition.. 75

5.1.2 The Data... 77

5.1.3 The Intra-Class Correlation.. 78

5.1.4 Group Mean Reliability.. 78

5.1.5 Determining whether τ_{00} is Significant.............................. 80

5.1.6 Adding Predictors... 81

5.2 The Aging Process: An Individual Quadratic Fit Approach........... 84

5.3 The Aging Process: A Time Weighted Approach........................... 88

5.4 Predictive Accuracy of the Methods.. 90

6 A Simulation Routine for Baseball.. 92

6.1 Strategic Decision Making in Baseball.. 92

6.1.1 The Intentional Walk... 93

6.1.2 The Sacrifice Bunt... 94

6.2 Non-Strategic Decision Making Transitions in Baseball................. 95

6.2.1 Adjusting Outcome Probabilities due to Handedness........... 96

6.2.2 Adjusting Outcome Probabilities due to Park Affects.......... 97

6.2.3 Adjusting Outcome Probabilities due to Defense................. 98

6.3 Determining Type of Batter Advance.. 101

6.4 Speed, Baserunner Destinations, and Pitcher Substitutions............. 103

6.4.1 Accounting for Double Plays.. 105

6.4.2 Bullpen Behavior... 106

6.5 Prediction Variability from Simulations and Computing Time........ 106

7 Betting Strategies and Simulation Results.. 109

7.1 Sports Gambling Markets.. 109

7.1.1 American and European Odds.. 110

7.1.2 Gambling on Baseball.. 111

7.2 Gambling Strategies.. 113

7.2.1 The Kelly Criterion – A Derivation.................................... 114

7.2.2 The Kelly Criterion – An Example..................................... 118

7.2.3 Simultaneous Kelly Betting.. 120

7.3 Preliminary Results – Training Data... 121

7.4 Final Results – 2007 MLB Season... 123

7.4.1 Portfolio Specifications.. 125

7.4.2 Over/Under Betting Line Performance.............................. 126

7.4.3 Run Line Betting Performance.. 128

7.4.4 Money Line Betting Performance...................................... 132

7.4.5 Comparing Simulation vs. Posted Odds Predictions........... 136

7.5 A Two-Way Analysis of Portfolio Returns.................................... 139

7.5.1 Tukey Test for Additivity.. 142

7.5.2 Results... 144

8 Applications for Simulation Baseball.. 147

8.1 The Economics of Major League Baseball..................................... 147

8.1.1 The Salary Cap in Major League Baseball......................... 148

8.1.2 Previous Research Estimating Player Worth....................... 151

8.1.3 The Economic Value of a Win.. 153

8.2 Salary Arbitration in Major League Baseball................................. 162

8.3 Is Alex Rodriguez worth $27.5 Million a Year?............................. 166

8.4 Conclusion... 169

References.. 171

LIST OF FIGURES

2.1 Transition Matrix for a Game of Baseball 10

2.2 Decomposition of Matrix A_0 12

2.3 Updated Version of A_0 with Run Potentials 13

2.4 Finalized Version of Transition Matrix A_0 14

2.5 Finalized Version of Transition Matrix B_0 16

2.6 Finalized Version of Transition Matrix C_0 17

2.7 Finalized Version of Transition Matrix A_1 18

2.8 Finalized Version of Transition Matrix B_1 20

2.9 Finalized Version of Transition Matrix A_2 21

2.10 Finalized Version of Transition Matrix D_0 22

2.11 Finalized Version of Transition Matrix E_1 22

2.12 Finalized Version of Transition Matrix F_2 23

3.1 OBP Predictions of Various Scenarios using the *Log5* and *W-Log5* Methods 38

4.1 Three Beta Distributions with $\frac{\alpha}{\alpha+\beta} = .300$ but Varying Degrees of Concentration around the Mean 54

4.2 Plot Showing the Estimated Batting Ability of Derek Jeter Against 18 Different Opposing Pitchers from Table 4.2 59

4.3 Contour Plots of the Marginal Posterior Density of $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ for the Wakefield Data 64

5.1 Player Performance versus Age 73

5.2 Group Mean Reliability and Group Size Measures for Batter Data........... 79

5.3 Group Mean Reliability and Group Size Measures for Pitcher Data.......... 79

5.4 Player Performance Progression as a Function of Age............................... 82

5.5 Quadratic Fit with 95% Confidence Interval Career Trajectories for a few of MLB's Longest Tenured Offensive Players.. 85

5.6 Quadratic Fit with 95% Confidence Interval Career Trajectories for a few of MLB's Longest Tenured Pitchers.. 87

5.7 OBP Model Weights Histograms for (a) Base Case (b) April Modified Case.. 90

6.1 Design Routine for Calculating OBP^*... 100

6.2 Determining the Number of Game Simulations Needed in Order to Achieve Stable Results.. 107

6.3 Computer Run Time versus Number of Simulations................................. 108

7.1 A Search for Appropriate Home Field Advantage Levels.......................... 123

7.2 Over/Under Betting Line Results... 127

7.3 Run Line Betting Results.. 128

7.4 Visual Display of Run Line Season Performance – All Betting Strategies.. 132

7.5 Money Line Betting Results.. 134

7.6 Visual Display of Money Line Season Performance – All Betting Strategies.. 136

8.1 Relationship Between Wins and Payroll for 2007 MLB Teams................. 149

8.2 Probability Estimates for Reaching the Playoffs Given Number of Regular Season Wins.. 160

8.3 Total Revenue versus Win Totals... 161

8.4 The Marginal Win Value... 161

LIST OF TABLES

3.1 Ability Distributions for Fitted Random Effects Model for each of Eight Batting Statistics.......... 29

3.2 Correlation of Batting Statistics Between the 2002 & 2003 Seasons.......... 30

3.3 Net Expected Run Values.......... 34

3.4 Posterior Mean Values of Situational Effects on Batting Average.......... 40

4.1 Kenny Lofton’s Record against Selected Pitchers.......... 52

4.2 Kenny Lofton’s Estimated Ability against Selected Pitchers.......... 56

4.3 Derek Jeter's Estimated Ability against Selected Pitchers.......... 58

4.4 Tim Wakefield’s Record and Estimated Ability against Selected Hitters.......... 63

4.5 Mike Mussina's Record and Estimated Ability against Selected Hitters.......... 66

5.1 Model Performance Summaries for Batters and Pitchers.......... 83

5.2 MAD for the Various Prediction Methods.......... 91

6.1-6.3 Sacrifice Bunt Scenarios and Corresponding Outcome Probabilities.......... 95

6.4 2007 OBP Park Factor Effects.......... 98

6.5 2007 Team Defensive Abilities.......... 100

6.6 Historical Rate of Walks and Homeruns allowed by Curt Schilling.......... 102

6.7 Historical Rate of Walks, Homeruns, Singles, Double, and Triples Made by Derek Jeter.......... 102

6.8 Expected Rate of Walks, Homeruns, Singles, Double, and Triples Made by Derek Jeter versus Curt Schilling.......... 103

6.9 Historical Base Runner Destinations by Type of Hit.......... 104

7.1 Over/Under Line Performance of Models and Betting Strategies.......... 127

7.2	Money and Run Line Season Performance	129
7.3	Run Line Performance of Models and Betting Strategies	131
7.4	Money line Performance of Models and Betting Strategies	135
7.5	Top 10 Games with The Greatest Discrepancy in Predicted Win Percentages	137
7.6	Two-Way Layout of Money Line Performance	140
7.7	Two-Way Layout of Run Line Performance	140
7.8	Interaction and Factor Effects for Run Line Data	145
7.9	Interaction and Factor Effects for Money Line Data	145
8.1	List of 2007 MLB Team Wins and Payroll	150
8.2	Local Revenues in MLB (source: Blue Ribbon Panel report)	158
8.3	Two-Tiered Model of Revenue for an Average MLB Franchise	159
8.4	Baseball Arbitration Case Settlements, 1984 – 1991	165
8.5	2008 MLB Season Predictions	167
8.6	2008 MLB Season Predictions – Rodriguez/Garciaparra Trade	168

ACKNOWELDGEMENTS

There are many people who have played a role in the completion of this project and to each of them I am tremendously grateful. To my committee members, Professors Jan de Leeuw, Rick Schoenberg, and Hal Stern, and my co-chairs, Professors Mark Hansen and Don Morrison, I owe a special thanks for their willingness and efforts to guide me through the dissertation process. I also wish to give particular recognition and my sincere appreciation to Hal Stern (UC Irvine) and Don Morrison (UCLA Anderson School of Management) who have both gone above and beyond their duties as faculty members outside of my own department and have selflessly invested in me nonetheless.

I am also forever indebted to my friend, mentor, and professor, Nicolas Christou. He was the instructor of my first statistics course at UCLA back in 2001, and over the years I have had the privilege of being his student through 7 different courses. He played a significant part in my decision to pursue a PhD. in statistics and has played an even greater role in my completion of such an arduous task. Our conversations have meant much to me over these years and I will miss having him down the hallway with his office door always open.

I would also like to acknowledge my friend and classmate, Chris Barr, who has unselfishly given me generous amounts of his time and effort in learning all aspects of computing with R, and my appreciation to Professor Vivian Lew for her prayers and continual words of encouragement.

Lastly, I would like to thank all of my friends and family who had confidence in me and my abilities, even when I did not. For my confidants, Emily, Esther, Jonathan, and Randy, each of you are more than friends to me and words cannot express how blessed I am to know you. And finally for my parents, Dave and Sharyn Sugano, who for 26 years have provided me every opportunity to make this PhD. a reality, and who have more importantly always instructed their children in the ways of the Lord. To God be all glory and honor.

VITA

1981	Born, Ann Arbor, Michigan, USA.
2002	B.S. Mathematics/Applied Science – Actuarial Science, University of California, Los Angeles
2004	M.S. Biostatistics, University of California, Los Angeles
2006	M.S. Statistics, University of California, Los Angeles
2006	C. Phil. Statistics, University of California, Los Angeles

PUBLICATIONS

Stern H., Sugano A. "Batter-Pitcher Matchups in Baseball and Small Sample Decision Making". <u>Statistical Thinking in Sports</u> Ed. Jim Albert and Ruud Koning. CRC Press. 153-165. (2007)

Stern H., Sugano A. "Small Sample Decision Making in Baseball." *Chance.* 20(4): 40-47 (2007).

ABSTRACT OF THE DISSERTATION

A Player Based Approach to

Baseball Simulation

By

Adam Philip Sugano

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2008

Professor Mark Hansen, Co-Chair

Professor Don Morrison, Co-Chair

Because of it's discrete, start-and-stop nature with finite set of possible outcomes, the game of baseball lends itself well to study via Markov chain simulations. Although this framework for baseball simulation has been widely discussed in academic literature for decades, actual and realistic implementation of this model has been sparse due to former time prohibitive computational capabilities, as well as the lack of availability to modern sabermetric baseball statistics. In this study, teams are broken down to their true component parts-individual players-and various estimates are used to predict any given player's current level of ability while also adjusting for an assortment of situational affects. The role and informational value of batter-pitcher matchup data is given

particular attention through use of a hierarchical beta-binomial model. The accuracy of this player based approach is measured by profitability of wagers versus the daily betting lines offered on individual games throughout the 2007 Major League Baseball Season. Natural applications for this method also include the cost-effectiveness of potential free-agent signings for major league teams, optimal batting lineup orderings, strategic in-game decision making, and a benchmark for teams, players, agents, and Major League Baseball in salary arbitration hearings.

Chapter 1

Introduction

Because of its discrete nature and finite set of possible outcomes, the game of baseball lends itself well to study via Markov chain simulations. Although this framework for baseball simulation has been widely discussed in academic literature for decades, actual and realistic implementation of this model has been sparse due to former time prohibitive computational capabilities, as well as the lack of availability to modern sabermetric baseball statistics. In this study, teams are broken down to their true component parts-individual players-and various estimates are used to predict any given player's current level of ability. The accuracy of this player based approach is measured by profitability of wagers versus the daily betting lines offered on individual games. Natural applications for this method also include the cost-effectiveness of potential free-agent signings for major league teams, optimal batting lineup orderings, and a benchmark for teams, players, agents, and major league baseball in salary arbitration.

Chapters 2 and 3 are mostly a review of past literature involving statistical research in baseball. Chapter 2 deals specifically with some of the various mathematical properties in the structure of a baseball game that allows it to be simulated within a Markov chain framework. A discussion of past models for baseball simulation and their limitations are assessed and mention of how the proposed player based simulation routine of this dissertation improves upon these limitations is also stated. Chapter 3 begins by discussing the bias that exist in many traditional baseball statistics and the evolvement of new sabermetric measures aimed at better reflecting a player's true value. The chapter continues by summarizing results from past work involving the magnitude of situational biases affecting a batter's ability to get on base, as well as proposing a new method aimed at predicting the probability of success in an at-bat with a batter and pitcher of given abilities.

Chapter 4 evaluates the possibility of an interactive matchup effect between individual batters and pitchers. For example, if a .300 career hitter has historically performed poorly, say 1 for 20 (.050), against a particular pitcher, how significant is this deviation from the batter's career average which is comprised of a much larger sample size (at-bats). This raises an important question about how a baseball manager should combine the limited information about a particular batter-pitcher matchup with the larger amount of data available about the individual player when making decisions. Is this batter really having a tough time against this particular pitcher? Or has he just been unlucky? In this chapter a closer look is taken at baseball batter-pitcher matchups using a hierarchical beta-binomial model. Although the results of the study may not be surprising to a statistician, it is the first formal work of its kind to address such a question within baseball and serves as a solid contribution to sports related statistical research.

Chapter 5 evaluates various methods of predicting a player's future performance based on past historical data. It addresses the relationship between performance and age and the question of how similar the aging process is across players. A multilevel model, an individual quadratic fit model, and an age independent time weighted model are all introduced and analyzed through their ability to effectively predict future player performance.

Chapter 6 outlines the simulation process at a player based level that takes factors such as speed, defense, pitching substitutions, and park effects into account and uses empirical data for strategic decision making in baseball. Although in-game strategic decision making is not a focus of this dissertation, it should nevertheless be noted that searches for optimal decisions made in baseball can be performed via the simulation framework that is being introduced.

Chapter 7 discusses the various betting lines that are available for baseball as well as various betting strategies that can be implemented by a bettor, including the notable Kelly Criterion. The performance of the player based simulation process is reported across all input methods, betting strategies, and betting lines for the 2007 MLB season. An analysis is given as to which methods perform best at beating the odds makers lines and model superiority is defined primarily by final total bankroll of a hypothetical gambler. A breakdown of games that had significant discrepancies in predicted outcomes given by the simulator and the odds makers is also reviewed and discussed.

Chapter 8 outlines some of the possible applications of a player based approach to baseball simulation within baseball's free agency and salary arbitration system. More specifically, the chapter discusses the models use for team general managers in evaluating potential trade scenarios and for determining the marginal value of individual players. The case of Major League Baseball's most highly paid player, Alex Rodriguez, and his $27.5 million dollar a year contract is also scrutinized. The chapter concludes with predictions for the 2008 MLB season.

Chapter 2

Baseball within a Markov Chain Framework

Baseball is one of the very few competitive team sports which can be analyzed as a discrete series of events with only a limited number of primary variables. This is different from the situation in most other sports. In football, for example, every play can be presented as a discrete event, but the possible combinations of offensive and defensive tactics and possible outcomes are essentially unlimited. In sports such as hockey and basketball, play is continuous. In baseball, the nine defensive players, for all practical purposes, always play the same basic positions, and we can consider the primary variables in each turn at bat to be:

1. The probability of the batter generating a particular play
2. The location of the base runners
3. The number of outs

The latter two variables are jointly referred to as the "base-out situation." When a player comes up to bat, he will always be in one of 24 possible base-out situations. These situations are determined by the 2^3 or 8 possible ways runners can occupy the three bases at any given time combined with the 3 possibilities of a team having 0, 1, or 2 outs throughout any half-inning. The half inning ends when a team reaches 3 outs.

Because of this discrete structure, baseball can reasonably be viewed as a Markov chain or Markov process, since it meets the following requirements:

1. There are a finite number of possible outcomes, or states (here defined by the location of base runners, the number of outs, and the lineup position).
2. The probability that the system will be in a given state depends only on the previous state and not how that state was reached; for example, a batter facing the situation of bases empty and no outs has the same probabilities, whether he is the first batter in the inning or follows a homerun
3. For each possible state, the probability that the system initially occupies that state is known; here, the only possible initial state is bases empty, no outs, and the leadoff batter up.

2.1 Markov Chain Properties

In mathematics, a Markov chain is a discrete-time stochastic process with the Markov property. Having the Markov property means for a given process, knowledge of the previous states is irrelevant for predicting the probability of subsequent states. Formally,

$$Pr(X_{n+1} = x \mid X_n = x_{n,} \ldots, X_1 = x_1) = Pr(X_{n+1} = x \mid X_n = x_n)$$

where the possible values of X_i form a countable set S called the state space of the chain. In this way a Markov chain is "memoryless": no given state has any causal connection with a previous state. At each point in time the system may have changed states from the state the system was in the moment before, or the system may have stayed in the same state. The changes of state are called transitions. If a sequence of states has the Markov property, then every future state is conditionally independent of every prior state. While some baseball aficionados may have trouble accepting successive at-bats (states) as truly independent, evidence to be provided later will show this assumption yet to be disproved.

Before proceeding further, a few other properties of Markov chains and definitions need to be mentioned. First, let the probability of going from state *i* to state *j* in *n* time steps be written as

$$p_{ij}^{(n)} = \Pr(X_n = j \mid X_0 = i)$$

and define the single-step transition as

$$p_{ij} = Pr(X_1 = j \mid X_0 = i)$$

Next, a state *j* is said to be accessible from state *i* (written $i \to j$) if, given that we are in state *i*, there is a non-zero probability that at some time in the future, we will be in state *j*. That is, there exists an *n* such that

$$Pr(X_n = j \mid X_0 = i) > 0$$

In the context of baseball, any state with *m* outs is not accessible from any state with *n* outs if $m < n$. Also, a state *i* is called absorbing if it is impossible to leave this state. Therefore, the state *i* is absorbing if and only if

$$p_{ii} = 1 \text{ and } p_{ij} = 0 \text{ for } i \neq j$$

In baseball, the absorbing state is whenever a team reaches 3 outs and the half inning is now over. And finally, if the state space is finite, the transition probability distribution can be represented by a matrix, called a transition matrix, with the $(i, j)^{th}$ element of the matrix equal to

$$p_{ij} = Pr(X_{n+1} = j \mid X_n = i)$$

2.2 Previous Work Modeling Baseball via a Markov Chain

Modeling baseball as a Markov process is nothing new. Baseball was first discussed as an example of a mathematical model involving dynamic programming and Markov processes in Howard (1960) and Bellman (1964). Both authors were primarily concerned with using baseball as an interesting example illustrating the usage of particular types of mathematical models. Howard, formulating a simplified computational example, went so far as to specify numerical values for the probabilities of a limited set of plays, but he assumed all players in the lineup were identical, and he issued a specific disclaimer regarding the validity of either assumptions or data. Bellman confined himself to theoretical aspects of his models.

Later work by Bukiet and Harold (1994) allowed for unique players statistics and incorporated team pitching quality into the model. The use of this player based approach is something that feasibly could only have been implemented recently with the advent of modern computing power along with the arrival and availability of detailed player statistics. Although the work by Bukiet was a major step forward in creating a more realistic view of simulating baseball through Markov chains, it still failed to capture a lot of the inherent individual nature of a baseball game. For example, while pitching is incorporated into the Bukiet model for baseball simulation, it is only done so at the team level by scaling the offensive production of each variable (walks, singles, doubles, triples, and home runs) for a team by the ratio of the on-base average of the league against the opposing team's pitchers to the mean on-base average for pitchers in the league. Also, once it is determined that a player reaches base, all players are given the same probabilities of advancing either via a walk, single, double, triple, or homerun. This again is not a very realistic view of baseball as some hitters hit more for power (homeruns) while others rely on just making contact (singles) but either way takes away from the unique abilities of each individual player. All runners on base are also given the same destinations for each particular type of hit. For example, a runner on 1^{st} base always advances to 3^{rd} on a single and a runner on 2^{nd} always scores a run on a double. This approach neglects the varying speed abilities of each player. Although less crucial to a simulation of a baseball game, the Bukiet model also does not allow for stolen bases, sacrifice situations, and double plays.

2.3 The Transition Matrix

As mentioned earlier, a baseball game is continually in one of 24 possible states until the added absorption - or 25^{th} state - is reached. The transition matrix for such a game is as follows:

	0	1	2	3	12	13	23	123	0	1	2	3	12	13	23	123	0	1	2	3	12	13	23	123	*
	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3
0,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
1,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
2,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
3,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
12,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
13,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
23,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
123,0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	A_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	B_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	C_0	D_0
0,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
1,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
2,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
3,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
12,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
13,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
23,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
123,1	0	0	0	0	0	0	0	0	A_1	A_1	A_1	A_1	A_1	A_1	A_1	A_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	E_1
0,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
1,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
2,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
3,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
12,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
13,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
23,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
123,2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A_2	A_2	A_2	A_2	A_2	A_2	A_2	A_2	F_2
*,3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 2.1: Transition Matrix for a Game of Baseball

The numbers in the first column represent the before state, or the situation a new batter inherits when coming to the plate. This situation is determined by the position of any runners on base and how many outs there are currently. For example, the entry **(12,1)** is the situation of having runners on first and second with one out. The numbers in the top row represent the after state, or the possible situations the game can transition to once play continues. These states are coded similarly with the top number representing the position of the runners on base and the bottom number representing the number of outs. The entries within the matrix are the probabilities of transitioning from one particular state to another. There are a large amount of zero entries in this matrix indicating that transitions between these specific states are not accessible. The reasoning for this is simple; outs can only be gained in baseball, never taken away. Since the above 25x25 matrix is rather cumbersome to deal with a straightforward simplification results in the following transition matrix for any player i

$$T_i = \begin{pmatrix} A_0 & B_0 & C_0 & D_0 \\ 0_{8x8} & A_1 & B_1 & E_1 \\ 0_{8x8} & 0_{8x8} & A_2 & F_2 \\ 0_{1x8} & 0_{1x8} & 0_{1x8} & 1 \end{pmatrix}$$

where:

A_0 = 8x8 matrix containing the probability of events from 0 out state to 0 out state
A_1 = 8x8 matrix containing the probability of events from 1 out state to 1 out state
A_2 = 8x8 matrix containing the probability of events from 2 out state to 2 out state
B_0 = 8x8 matrix containing the probability of events from 0 out state to 1 out state
B_1 = 8x8 matrix containing the probability of events from 1 out state to 2 out state
C_0 = 8x8 matrix containing the probability of events from 0 out state to 2 out state
D_0 = 8x1 matrix containing the probability of events from 0 out state to 3 out state
E_1 = 8x1 matrix containing the probability of events from 1 out state to 3 out state
F_2 = 8x1 matrix containing the probability of events from 2 out state to 3 out state
0_{8x8} or 0_{1x8} = Blocks of zeroes representing transitions which are not possible

As can be noticed in the above notation, each player i will come to bat with their own transition matrix filled with probabilities that take into account their specific skill set and abilities relative to the pitcher and defense they are facing. It is the accuracy of these probabilities combined with the ability to realistically simulate a game of baseball and all its intricacies that will determine how well of a simulation will take place.

Since each player's transition matrix is so large it helps to break down the possible scenarios piece by piece. The first subset of T_i to look at is A_0, the 8x8 matrix containing the probability of events from a 0 out state to another 0 out state. A closer look at A_0 reveals certain entries with zero probability of occurrence. These zero probabilities are due to the fact that only 1 runner can be added on base at a time. For example, there is no possible scenario in which the bases can initially be empty and then have two or more runners on base after just one at-bat.

$$A_0 = \begin{array}{c|cccccccc} & 0,0 & 1,0 & 2,0 & 3,0 & 12,0 & 13,0 & 23,0 & 123,0 \\ \hline 0,0 & A_{01} & A_{09} & A_{017} & A_{025} & 0 & 0 & 0 & 0 \\ 1,0 & A_{02} & A_{010} & A_{018} & A_{026} & A_{034} & A_{042} & A_{050} & 0 \\ 2,0 & A_{03} & A_{011} & A_{019} & A_{027} & A_{035} & A_{043} & A_{051} & 0 \\ 3,0 & A_{04} & A_{012} & A_{020} & A_{028} & A_{036} & A_{044} & A_{052} & 0 \\ 12,0 & A_{05} & A_{013} & A_{021} & A_{029} & A_{037} & A_{045} & A_{053} & A_{061} \\ 13,0 & A_{06} & A_{014} & A_{022} & A_{030} & A_{038} & A_{046} & A_{054} & A_{062} \\ 23,0 & A_{07} & A_{015} & A_{023} & A_{031} & A_{039} & A_{047} & A_{055} & A_{063} \\ 123,0 & A_{08} & A_{016} & A_{024} & A_{032} & A_{040} & A_{048} & A_{056} & A_{064} \end{array}$$

Figure 2.2: Decomposition of matrix A_0

2.3.1 Run Potentials

Certain transitions also result in runs being scored. If a batter comes to the plate with a **(0,0)** situation and the next immediate state is also **(0,0)**, then only one event could have occurred, a homerun, resulting in the score of the game to change by one run. Therefore, it is important to find the run potentials of each transition and to keep track of these runs throughout the game:

$$A_0 = \left(\begin{array}{c|cccccccc} & 0,0 & 1,0 & 2,0 & 3,0 & 12,0 & 13,0 & 23,0 & 123,0 \\ \hline 0,0 & A_{01}\,|\,1 & A_{09}\,|\,0 & A_{017}\,|\,0 & A_{025}\,|\,0 & 0 & 0 & 0 & 0 \\ 1,0 & A_{02}\,|\,1,2 & A_{010}\,|\,0,1 & A_{018}\,|\,0,1 & A_{026}\,|\,0,1 & A_{034}\,|\,0 & A_{042}\,|\,0 & A_{050}\,|\,0 & 0 \\ 2,0 & A_{03}\,|\,1,2 & A_{011}\,|\,1 & A_{019}\,|\,0,1 & A_{027}\,|\,0,1 & A_{035}\,|\,0 & A_{043}\,|\,0 & A_{051}\,|\,0 & 0 \\ 3,0 & A_{04}\,|\,1,2 & A_{012}\,|\,1 & A_{020}\,|\,1 & A_{028}\,|\,0,1 & 0 & A_{044}\,|\,0 & A_{052}\,|\,0 & 0 \\ 12,0 & A_{05}\,|\,2,3 & A_{013}\,|\,1,2 & A_{021}\,|\,1,2 & A_{029}\,|\,1,2 & A_{037}\,|\,0,1 & A_{045}\,|\,0,1 & A_{053}\,|\,0,1 & A_{061}\,|\,0 \\ 13,0 & A_{06}\,|\,2,3 & A_{014}\,|\,1,2 & A_{022}\,|\,1,2 & A_{030}\,|\,1,2 & A_{038}\,|\,1 & A_{046}\,|\,0,1 & A_{054}\,|\,0,1 & A_{062}\,|\,0 \\ 23,0 & A_{07}\,|\,2,3 & A_{015}\,|\,2 & A_{023}\,|\,1,2 & A_{031}\,|\,1,2 & A_{039}\,|\,1 & A_{047}\,|\,1 & A_{055}\,|\,0,1 & A_{063}\,|\,0 \\ 123,0 & A_{08}\,|\,3,4 & A_{016}\,|\,2,3 & A_{024}\,|\,2,3 & A_{032}\,|\,2,3 & A_{040}\,|\,1,2 & A_{048}\,|\,1,2 & A_{056}\,|\,1,2 & A_{064}\,|\,0,1 \end{array}\right)$$

Figure 2.3: Updated version of A_0 with Run Potentials

2.3.2 Transition Constraints

The numbers on the right of the vertical bars in Figure 2.3 represent the run potentials for each transition. It can be seen that many of these transitions have multiple possibilities for their run potentials. This is due to the fact that not all transitions are unique to one event. For example, in the course of a game the situation could transition from **(1,0)** to **(2,0)** and have one run score if the batter hits a double and the runner on first scores. However, it is also possible for the game to transition from **(1,0)** to **(2,0)** via a stolen base and have zero runs score. A more in depth look at the transition probabilities and run potentials for Figure 2.3, however, reveals some situations that are extremely unlikely to occur throughout the course of an entire season, let alone a single game. Transition A_{013} shows two possible run potentials of 1 or 2. A run potential of 2 for this transition is easy to imagine as it is not out of the question for a single to score runners from both first and second. A run potential of 1 for this transition, although possible, is very unlikely to happen and nearly impossible to model. It would essentially require a two base error that scores the runner on second while leaving the runner on first stationary. In order to create a more reasonable deterministic model for runner advancement a few restrictions will now be imposed. These restrictions are imposed on all events that are either extremely rare, extremely hard to model, or both.

1. Runners on third base cannot be picked off by the pitcher
2. Runners cannot steal home base
3. No double steals
4. No errors that advance any runners by more than one base
5. A runner can only advance via a hit, steal, walk, sacrifice, or defensive error (no base running errors by batter)
6. Runners on third advance on any type of hit

2.3.3 Finalized Transition Matrices

With these rules imposed a new transition matrix for A_0 follows:

$$A_0 = \left(\begin{array}{c|cccccccc}
 & 0,0 & 1,0 & 2,0 & 3,0 & 12,0 & 13,0 & 23,0 & 123,0 \\
\hline
0,0 & A_{01}\,|\,1 & A_{09}\,|\,0 & A_{017}\,|\,0 & A_{025}\,|\,0 & 0 & 0 & 0 & 0 \\
1,0 & A_{02}\,|\,2 & A_{10}\,|\,1 & A_{018}\,|\,0,1 & A_{026}\,|\,1 & A_{034}\,|\,0 & A_{042}\,|\,0 & A_{050}\,|\,0 & 0 \\
2,0 & A_{03}\,|\,2 & A_{011}\,|\,1 & A_{019}\,|\,1 & A_{027}\,|\,0,1 & A_{035}\,|\,0 & A_{043}\,|\,0 & A_{051}\,|\,0 & 0 \\
3,0 & A_{04}\,|\,2 & A_{012}\,|\,1 & A_{020}\,|\,1 & A_{028}\,|\,1 & 0 & A_{044}\,|\,0 & 0 & 0 \\
12,0 & A_{05}\,|\,3 & A_{013}\,|\,2 & A_{021}\,|\,2 & A_{029}\,|\,2 & A_{037}\,|\,1 & A_{045}\,|\,0,1 & A_{053}\,|\,1 & A_{061}\,|\,0 \\
13,0 & A_{06}\,|\,3 & A_{014}\,|\,2 & A_{022}\,|\,2 & A_{030}\,|\,2 & A_{038}\,|\,1 & A_{046}\,|\,1 & A_{054}\,|\,0,1 & A_{062}\,|\,0 \\
23,0 & A_{07}\,|\,3 & A_{015}\,|\,2 & A_{023}\,|\,2 & A_{031}\,|\,2 & A_{039}\,|\,1 & A_{047}\,|\,1 & A_{055}\,|\,1 & A_{063}\,|\,0 \\
123,0 & A_{08}\,|\,4 & A_{016}\,|\,3 & A_{024}\,|\,3 & A_{032}\,|\,3 & A_{040}\,|\,2 & A_{048}\,|\,2 & A_{056}\,|\,2 & A_{064}\,|\,1
\end{array}\right)$$

Figure 2.4: Finalized Version of Transition Matrix A_0

For each $A_{ij} \neq 0$ and following the previously outlined simulation constraints, the above transitions are possible via the following advancements:

A_{01}: *Homerun*	A_{09}: *Single/walk*	A_{017}: *Double*	A_{025}: *Triple*
A_{02}: *Homerun*	A_{010}: *Single*	A_{018}: *SB, double*	A_{026}: *Triple*
A_{03}: *Homerun*	A_{011}: *Single*	A_{019}: *Double*	A_{027}: *SB, triple*
A_{04}: *Homerun*	A_{012}: *Single*	A_{020}: *Double*	A_{028}: *Triple*
A_{05}: *Homerun*	A_{013}: *Single*	A_{021}: *Double*	A_{029}: *Triple*
A_{06}: *Homerun*	A_{014}: *Single*	A_{022}: *Double*	A_{030}: *Triple*
A_{07}: *Homerun*	A_{015}: *Single*	A_{023}: *Double*	A_{031}: *Triple*
A_{08}: *Homerun*	A_{016}: *Single*	A_{024}: *Double*	A_{032}: *Triple*

A_{033}: *Impossible*	A_{041}: *Impossible*	A_{049}: *Impossible*	A_{057}: *Impossible*
A_{034}: *Single/walk*	A_{042}: *Single*	A_{050}: *Double*	A_{058}: *Impossible*
A_{035}: *Single/walk*	A_{043}: *Single*	A_{051}: *Double*	A_{059}: *Impossible*
A_{036}: *Impossible*	A_{044}: *Walk*	A_{052}: *Impossible*	A_{060}: *Impossible*
A_{037}: *Single*	A_{045}: *SB, single*	A_{053}: *Double*	A_{061}: *Single/walk*
A_{038}: *Single*	A_{046}: *Single*	A_{054}: *SB, double*	A_{062}: *Walk*
A_{039}: *Single*	A_{047}: *Single*	A_{055}: *Double*	A_{063}: *Walk*
A_{040}: *Single*	A_{048}: *Single*	A_{056}: *Double*	A_{064}: *Single/walk*

With the above methodology and restrictions performed on the remaining subsets of a player's transition matrix, the following results:

$$B_0 = \left(\begin{array}{c|cccccccc} & 0,1 & 1,1 & 2,1 & 3,1 & 12,1 & 13,1 & 23,1 & 123,1 \\ \hline 0,0 & B_{01} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1,0 & B_{02} & B_{010} & B_{018} & B_{026} & 0 & 0 & 0 & 0 \\ 2,0 & B_{03} & B_{011} & B_{019} & B_{027} & 0 & 0 & 0 & 0 \\ 3,0 & B_{04} & B_{012} & B_{020} & B_{028} & 0 & 0 & 0 & 0 \\ 12,0 & B_{05} & B_{013} & B_{021} & B_{029} & B_{037} & B_{045} & B_{053} & 0 \\ 13,0 & B_{06} & B_{014} & B_{022} & B_{030} & B_{038} & B_{046} & B_{054} & 0 \\ 23,0 & B_{07} & B_{015} & B_{023} & B_{031} & B_{039} & B_{047} & B_{055} & 0 \\ 123,0 & B_{08} & B_{016} & B_{024} & B_{032} & B_{040} & B_{048} & B_{056} & B_{064} \end{array}\right)$$

$$\Downarrow$$

$$B_0 = \begin{array}{c|cccccccc}
 & 0,1 & 1,1 & 2,1 & 3,1 & 12,1 & 13,1 & 23,1 & 123,1 \\
\hline
0,0 & B_{01}\,|\,0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1,0 & B_{02}\,|\,0 & B_{010}\,|\,0 & B_{018}\,|\,0 & B_{026}\,|\,0 & 0 & 0 & 0 & 0 \\
2,0 & B_{03}\,|\,0 & B_{011}\,|\,0 & B_{019}\,|\,0 & B_{027}\,|\,0 & 0 & 0 & 0 & 0 \\
3,0 & B_{04}\,|\,1 & B_{012}\,|\,0 & 0 & B_{028}\,|\,0 & 0 & 0 & 0 & 0 \\
12,0 & 0 & B_{013}\,|\,0 & B_{021}\,|\,0,1 & B_{029}\,|\,0,1 & B_{037}\,|\,0 & B_{045}\,|\,0 & B_{053}\,|\,0 & 0 \\
13,0 & 0 & B_{014}\,|\,1 & B_{022}\,|\,1 & B_{030}\,|\,0,1 & 0 & B_{046}\,|\,0 & B_{054}\,|\,0 & 0 \\
23,0 & 0 & B_{015}\,|\,1 & B_{023}\,|\,1 & B_{031}\,|\,0,1 & 0 & 0 & B_{055}\,|\,0 & 0 \\
123,0 & 0 & B_{016}\,|\,2 & B_{024}\,|\,2 & B_{032}\,|\,2 & 0 & B_{048}\,|\,0,1 & B_{056}\,|\,0,1 & B_{064}\,|\,0
\end{array}$$

Figure 2.5: Finalized Version of Transition Matrix B_0

B_{01}: Batter out
B_{02}: CS
B_{03}: CS
B_{04}: Sacrifice
B_{05}: Impossible
B_{06}: Impossible
B_{07}: Impossible
B_{08}: Impossible

B_{09}: Impossible
B_{010}: Batter out
B_{011}: Single/walk & runner out
B_{012}: Failed Sac bunt
B_{013}: CS
B_{014}: Sacrifice
B_{015}: Single & runner out
B_{016}: Single & runner out

B_{017}: Impossible
B_{018}: Double & runner out, SB & batter out
B_{019}: Double & runner out, batter out
B_{020}: Impossible
B_{021}: PO at 1^{st}, double runner out
B_{022}: Double & runner out
B_{023}: Double & runner out
B_{024}: Double & runner out

B_{025}: Impossible
B_{026}: Triple & runner out
B_{027}: Sacrifice, SB & batter out, triple & runner out
B_{028}: Batter out
B_{029}: SB & PO at 1^{st}, triple & runner out
B_{030}: CS, triple & runner out
B_{031}: PO at 2^{nd}, sacrifice (both runners advance)
B_{032}: Triple & runner out

B_{033}: Impossible
B_{034}: Impossible
B_{035}: Impossible
B_{036}: Impossible
B_{037}: Batter out, single & runner out
B_{038}: Impossible
B_{039}: Impossible
B_{040}: Impossible

B_{041}: Impossible
B_{042}: Impossible
B_{043}: Impossible
B_{044}: Impossible
B_{045}: SB & batter out, single & runner out
B_{046}: Batter out
B_{047}: Impossible
B_{048}: PO at 2^{nd}, single & runner out

B_{049}: Impossible
B_{050}: Impossible
B_{051}: Impossible
B_{052}: Impossible
B_{053}: Double & runner out
B_{054}: SB & batter out
B_{055}: Batter out
B_{056}: PO at 1^{st}, double & runner out

B_{057}: Impossible
B_{058}: Impossible
B_{059}: Impossible
B_{060}: Impossible
B_{061}: Impossible
B_{062}: Impossible
B_{063}: Impossible

B_{064}: Batter out

$$C_0 = \left(\begin{array}{c|cccccccc}
 & 0,2 & 1,2 & 2,2 & 3,2 & 12,2 & 13,2 & 23,2 & 123,2 \\
\hline
0,0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1,0 & C_{02} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2,0 & C_{03} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3,0 & C_{04} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
12,0 & C_{05} & C_{013} & C_{021} & C_{029} & 0 & 0 & 0 & 0 \\
13,0 & C_{06} & C_{014} & C_{022} & C_{030} & 0 & 0 & 0 & 0 \\
23,0 & C_{07} & C_{015} & C_{023} & C_{031} & 0 & 0 & 0 & 0 \\
123,0 & C_{08} & C_{016} & C_{024} & C_{032} & C_{040} & C_{048} & C_{056} & 0
\end{array}\right)$$

$$\Downarrow$$

$$C_0 = \left(\begin{array}{c|cccccccc}
 & 0,2 & 1,2 & 2,2 & 3,2 & 12,2 & 13,2 & 23,2 & 123,2 \\
\hline
0,0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1,0 & C_{02} \mid 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2,0 & C_{03} \mid 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3,0 & C_{04} \mid 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
12,0 & 0 & C_{013} \mid 0 & C_{021} \mid 0 & C_{029} \mid 0 & 0 & 0 & 0 & 0 \\
13,0 & C_{06} \mid 1 & C_{014} \mid 0 & C_{022} \mid 0 & C_{030} \mid 0 & 0 & 0 & 0 & 0 \\
23,0 & C_{07} \mid 1 & C_{015} \mid 0 & C_{023} \mid 0 & C_{031} \mid 0 & 0 & 0 & 0 & 0 \\
123,0 & 0 & C_{016} \mid 1 & C_{024} \mid 1 & C_{032} \mid 1 & C_{040} \mid 0 & C_{048} \mid 0 & C_{056} \mid 0 & 0
\end{array}\right)$$

Figure 2.6: Finalized Version of Transition Matrix C_0

C_{01}: Impossible
C_{02}: CS & batter out, double play
C_{03}: CS & batter out, double play
C_{04}: Double play
C_{05}: Impossible
C_{06}: Double play
C_{07}: Double play
C_{08}: Impossible

C_{09}: Impossible
C_{010}: Impossible
C_{011}: Impossible
C_{012}: Impossible
C_{013}: CS & batter out, double play
C_{014}: Double play
C_{015}: Double play
C_{016}: Single & runners out

C_{017}: Impossible
C_{018}: Impossible
C_{019}: Impossible
C_{020}: Impossible
C_{021}: Double play
C_{022}: Double play
C_{023}: Double play
C_{024}: Double & runners out

C_{025}: Impossible
C_{026}: Impossible
C_{027}: Impossible
C_{028}: Impossible

C_{033}: Impossible
C_{034}: Impossible
C_{035}: Impossible
C_{036}: Impossible

C_{041}: Impossible
C_{042}: Impossible
C_{043}: Impossible
C_{044}: Impossible

C_{029}: *Double play*
C_{030}: *Double play*
C_{031}: *Double play*
C_{032}: *Triple & runners out*

C_{037}: *Impossible*
C_{038}: *Impossible*
C_{039}: *Impossible*
C_{040}: *Double Play*

C_{045}: *Impossible*
C_{046}: *Impossible*
C_{047}: *Impossible*
C_{048}: *Double play*

C_{049}: *Impossible*
C_{050}: *Impossible*
C_{051}: *Impossible*
C_{052}: *Impossible*
C_{053}: *Impossible*
C_{054}: *Impossible*
C_{055}: *Impossible*
C_{056}: *Double play*

C_{057}: *Impossible*
C_{058}: *Impossible*
C_{059}: *Impossible*
C_{060}: *Impossible*
C_{061}: *Impossible*
C_{062}: *Impossible*
C_{063}: *Impossible*
C_{064}: *Impossible*

$$A_1 = \left(\begin{array}{c|cccccccc}
 & 0,1 & 1,1 & 2,1 & 3,1 & 12,1 & 13,1 & 23,1 & 123,1 \\
\hline
0,1 & A_{11} & A_{19} & A_{117} & A_{125} & 0 & 0 & 0 & 0 \\
1,1 & A_{12} & A_{110} & A_{118} & A_{126} & A_{134} & A_{142} & A_{150} & 0 \\
2,1 & A_{13} & A_{111} & A_{119} & A_{127} & A_{135} & A_{143} & A_{151} & 0 \\
3,1 & A_{14} & A_{112} & A_{120} & A_{128} & 0 & A_{144} & A_{152} & 0 \\
12,1 & A_{15} & A_{113} & A_{121} & A_{129} & A_{137} & A_{145} & A_{153} & A_{161} \\
13,1 & A_{16} & A_{114} & A_{122} & A_{130} & A_{138} & A_{146} & A_{154} & A_{162} \\
23,1 & A_{17} & A_{115} & A_{123} & A_{131} & A_{139} & A_{147} & A_{155} & A_{163} \\
123,1 & A_{18} & A_{116} & A_{124} & A_{132} & A_{140} & A_{148} & A_{156} & A_{164}
\end{array}\right)$$

$$\Downarrow$$

$$A_1 = \left(\begin{array}{c|cccccccc}
 & 0,1 & 1,1 & 2,1 & 3,1 & 12,1 & 13,1 & 23,1 & 123,1 \\
\hline
0,1 & A_{11}\,|\,1 & A_{19}\,|\,0 & A_{117}\,|\,0 & A_{125}\,|\,0 & 0 & 0 & 0 & 0 \\
1,1 & A_{12}\,|\,2 & A_{110}\,|\,1 & A_{118}\,|\,0,1 & A_{126}\,|\,1 & A_{134}\,|\,0 & A_{142}\,|\,0 & A_{150}\,|\,0 & 0 \\
2,1 & A_{13}\,|\,2 & A_{111}\,|\,1 & A_{119}\,|\,1 & A_{127}\,|\,0,1 & A_{135}\,|\,0 & A_{143}\,|\,0 & A_{151}\,|\,0 & 0 \\
3,1 & A_{14}\,|\,2 & A_{112}\,|\,1 & A_{120}\,|\,1 & A_{128}\,|\,1 & 0 & A_{144}\,|\,0 & 0 & 0 \\
12,1 & A_{15}\,|\,3 & A_{113}\,|\,2 & A_{121}\,|\,2 & A_{129}\,|\,2 & A_{137}\,|\,1 & A_{145}\,|\,0,1 & A_{153}\,|\,1 & A_{161}\,|\,0 \\
13,1 & A_{16}\,|\,3 & A_{114}\,|\,2 & A_{122}\,|\,2 & A_{130}\,|\,2 & A_{138}\,|\,1 & A_{146}\,|\,1 & A_{154}\,|\,0,1 & A_{162}\,|\,0 \\
23,1 & A_{17}\,|\,3 & A_{115}\,|\,2 & A_{123}\,|\,2 & A_{131}\,|\,2 & A_{139}\,|\,1 & A_{147}\,|\,1 & A_{155}\,|\,1 & A_{163}\,|\,0 \\
123,1 & A_{18}\,|\,4 & A_{116}\,|\,3 & A_{124}\,|\,3 & A_{132}\,|\,3 & A_{140}\,|\,2 & A_{148}\,|\,2 & A_{156}\,|\,2 & A_{164}\,|\,1
\end{array}\right)$$

Figure 2.7: Finalized Version of Transition Matrix A_1

A_{11}: *Homerun*	A_{19}: *Single/walk*	A_{117}: *Double*	A_{125}: *Triple*
A_{12}: *Homerun*	A_{110}: *Single*	A_{118}: *SB, Double*	A_{126}: *Triple*
A_{13}: *Homerun*	A_{111}: *Single*	A_{119}: *Double*	A_{127}: *SB, Triple*
A_{14}: *Homerun*	A_{112}: *Single*	A_{120}: *Double*	A_{128}: *Triple*
A_{15}: *Homerun*	A_{113}: *Single*	A_{121}: *Double*	A_{129}: *Triple*
A_{16}: *Homerun*	A_{114}: *Single*	A_{122}: *Double*	A_{130}: *Triple*
A_{17}: *Homerun*	A_{115}: *Single*	A_{123}: *Double*	A_{131}: *Triple*
A_{18}: *Homerun*	A_{116}: *Single*	A_{124}: *Double*	A_{132}: *Triple*

A_{133}: *Impossible*	A_{141}: *Impossible*	A_{149}: *Impossible*	A_{157}: *Impossible*
A_{134}: *Single/walk*	A_{142}: *Single*	A_{150}: *Double*	A_{158}: *Impossible*
A_{135}: *Single/walk*	A_{143}: *Single*	A_{151}: *Double*	A_{159}: *Impossible*
A_{136}: *Impossible*	A_{144}: *Walk*	A_{152}: *Impossible*	A_{160}: *Impossible*
A_{137}: *Single*	A_{145}: *SB, single*	A_{153}: *Double*	A_{161}: *Single/walk*
A_{138}: *Single*	A_{146}: *Single*	A_{154}: *SB, double*	A_{162}: *Walk*
A_{139}: *Single*	A_{147}: *Single*	A_{155}: *Double*	A_{163}: *Walk*
A_{140}: *Single*	A_{148}: *Single*	A_{156}: *Double*	A_{164}: *Single/walk*

$$B_1 = \left(\begin{array}{c|cccccccc}
 & 0,2 & 1,2 & 2,2 & 3,2 & 12,2 & 13,2 & 23,2 & 123,2 \\
\hline
0,1 & B_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1,1 & B_{12} & B_{110} & B_{118} & B_{126} & 0 & 0 & 0 & 0 \\
2,1 & B_{13} & B_{111} & B_{119} & B_{127} & 0 & 0 & 0 & 0 \\
3,1 & B_{14} & B_{112} & B_{120} & B_{128} & 0 & 0 & 0 & 0 \\
12,1 & B_{15} & B_{113} & B_{121} & B_{129} & B_{137} & B_{145} & B_{153} & 0 \\
13,1 & B_{16} & B_{114} & B_{122} & B_{130} & B_{138} & B_{146} & B_{154} & 0 \\
23,1 & B_{17} & B_{115} & B_{123} & B_{131} & B_{139} & B_{147} & B_{155} & 0 \\
123,1 & B_{18} & B_{116} & B_{124} & B_{132} & B_{140} & B_{148} & B_{156} & B_{164}
\end{array}\right)$$

$$\Downarrow$$

$$B_1 = \left(\begin{array}{c|cccccccc} & 0,2 & 1,2 & 2,2 & 3,2 & 12,2 & 13,2 & 23,2 & 123,2 \\ \hline 0,1 & B_{11}\,|\,0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1,1 & B_{12}\,|\,0 & B_{110}\,|\,0 & B_{118}\,|\,0 & B_{126}\,|\,0 & 0 & 0 & 0 & 0 \\ 2,1 & B_{13}\,|\,0 & B_{111}\,|\,0 & B_{119}\,|\,0 & B_{127}\,|\,0 & 0 & 0 & 0 & 0 \\ 3,1 & B_{14}\,|\,1 & B_{112}\,|\,0 & 0 & B_{128}\,|\,0 & 0 & 0 & 0 & 0 \\ 12,1 & 0 & B_{113}\,|\,0,1 & B_{121}\,|\,0,1 & B_{129}\,|\,1 & B_{137}\,|\,0 & B_{145}\,|\,0 & B_{153}\,|\,0 & 0 \\ 13,1 & 0 & B_{114}\,|\,1 & B_{122}\,|\,1 & B_{130}\,|\,0,1 & 0 & B_{146}\,|\,0 & B_{154}\,|\,0 & 0 \\ 23,1 & 0 & B_{115}\,|\,1 & B_{123}\,|\,1 & B_{131}\,|\,0,1 & 0 & 0 & B_{155}\,|\,0 & 0 \\ 123,1 & 0 & B_{116}\,|\,2 & B_{124}\,|\,2 & B_{132}\,|\,2 & B_{140}\,|\,1 & B_{148}\,|\,0,1 & B_{156}\,|\,0,1 & B_{164}\,|\,0 \end{array}\right)$$

Figure 2.8: Finalized Version of Transition Matrix B_1

B_{11}: Batter out
B_{12}: CS
B_{13}: CS
B_{14}: Sacrifice
B_{15}: Impossible
B_{16}: Impossible
B_{17}: Impossible
B_{18}: Impossible

B_{19}: Impossible
B_{110}: Batter out
B_{111}: Single & runner out/CS & walk
B_{112}: Failed Sac bunt
B_{113}: CS, single & runner out
B_{114}: Sacrifice
B_{115}: Single & runner out
B_{116}: Single & runner out

B_{117}: Impossible
B_{118}: double & runner out
B_{119}: Batter out
B_{120}: Impossible
B_{121}: double & runner out
B_{122}: Double & runner out
B_{123}: Sacrifice/double & runner out
B_{124}: Double & runner out

B_{125}: Impossible
B_{126}: Triple & runner out
B_{127}: Sacrifice/triple & runner out
B_{128}: Batter out
B_{129}: Triple & runner out
B_{130}: CS, triple & runner out
B_{131}: PO at 2^{nd}, triple & runner out
B_{132}: Triple & runner out

B_{133}: Impossible
B_{134}: Impossible
B_{135}: Impossible
B_{136}: Impossible
B_{137}: Batter out/single & runner out
B_{138}: Impossible
B_{139}: Impossible
B_{140}: Sacrifice/single & runner out

B_{141}: Impossible
B_{142}: Impossible
B_{143}: Impossible
B_{144}: Impossible
B_{145}: Sacrifice/single & runner out
B_{146}: Batter out
B_{147}: Impossible
B_{148}: PO at 2^{nd}, Sacrifice (2 runners advance)

B_{149}: Impossible
B_{150}: Impossible
B_{151}: Impossible
B_{152}: Impossible
B_{153}: Sacrifice/double & runner out
B_{154}: SB & batter out
B_{155}: Batter out
B_{156}: PO at 1^{st}, double & runner out

B_{157}: Impossible
B_{158}: Impossible
B_{159}: Impossible
B_{160}: Impossible
B_{161}: Impossible
B_{162}: Impossible
B_{163}: Impossible

B_{164}: Batter out

$$A_2 = \left(\begin{array}{c|cccccccc} & 0,2 & 1,2 & 2,2 & 3,2 & 12,2 & 13,2 & 23,2 & 123,2 \\ \hline 0,2 & A_{21} & A_{29} & A_{217} & A_{225} & 0 & 0 & 0 & 0 \\ 1,2 & A_{22} & A_{210} & A_{218} & A_{226} & A_{234} & A_{242} & A_{250} & 0 \\ 2,2 & A_{23} & A_{211} & A_{219} & A_{227} & A_{235} & A_{243} & A_{251} & 0 \\ 3,2 & A_{24} & A_{212} & A_{220} & A_{228} & 0 & A_{244} & A_{252} & 0 \\ 12,2 & A_{25} & A_{213} & A_{221} & A_{229} & A_{237} & A_{245} & A_{253} & A_{261} \\ 13,2 & A_{26} & A_{214} & A_{222} & A_{230} & A_{238} & A_{246} & A_{254} & A_{262} \\ 23,2 & A_{27} & A_{215} & A_{223} & A_{231} & A_{239} & A_{247} & A_{255} & A_{263} \\ 123,2 & A_{28} & A_{216} & A_{224} & A_{232} & A_{240} & A_{248} & A_{256} & A_{264} \end{array} \right)$$

$$\Downarrow$$

$$A_2 = \left(\begin{array}{c|cccccccc} & 0,2 & 1,2 & 2,2 & 3,2 & 12,2 & 13,2 & 23,2 & 123,2 \\ \hline 0,2 & A_{21}\,|\,1 & A_{29}\,|\,0 & A_{217}\,|\,0 & A_{225}\,|\,0 & 0 & 0 & 0 & 0 \\ 1,2 & A_{22}\,|\,2 & A_{210}\,|\,1 & A_{218}\,|\,0,1 & A_{226}\,|\,1 & A_{234}\,|\,0 & A_{242}\,|\,0 & A_{250}\,|\,0 & 0 \\ 2,2 & A_{23}\,|\,2 & A_{211}\,|\,1 & A_{219}\,|\,1 & A_{227}\,|\,0,1 & A_{235}\,|\,0 & A_{243}\,|\,0 & A_{251}\,|\,0 & 0 \\ 3,2 & A_{24}\,|\,2 & A_{212}\,|\,1 & A_{220}\,|\,1 & A_{228}\,|\,1 & 0 & A_{244}\,|\,0 & 0 & 0 \\ 12,2 & A_{25}\,|\,3 & A_{213}\,|\,2 & A_{221}\,|\,2 & A_{229}\,|\,2 & A_{237}\,|\,1 & A_{245}\,|\,0,1 & A_{253}\,|\,1 & A_{261}\,|\,0 \\ 13,2 & A_{26}\,|\,3 & A_{214}\,|\,2 & A_{222}\,|\,2 & A_{230}\,|\,2 & A_{238}\,|\,1 & A_{246}\,|\,1 & A_{254}\,|\,0,1 & A_{262}\,|\,0 \\ 23,2 & A_{27}\,|\,3 & A_{215}\,|\,2 & A_{223}\,|\,2 & A_{231}\,|\,2 & A_{239}\,|\,1 & A_{247}\,|\,1 & A_{255}\,|\,1 & A_{263}\,|\,0 \\ 123,2 & A_{28}\,|\,4 & A_{216}\,|\,3 & A_{224}\,|\,3 & A_{232}\,|\,3 & A_{240}\,|\,2 & A_{248}\,|\,2 & A_{256}\,|\,2 & A_{264}\,|\,1 \end{array} \right)$$

Figure 2.9: Finalized Version of Transition Matrix A_2

A_{21}: Homerun	*A_{29}: Single/walk*	*A_{217}: Double*	*A_{225}: Triple*
A_{22}: Homerun	*A_{210}: Single*	*A_{218}: SB, double*	*A_{226}: Triple*
A_{23}: Homerun	*A_{211}: Single*	*A_{219}: Double*	*A_{227}: SB, Triple*
A_{24}: Homerun	*A_{212}: Single*	*A_{220}: Triple*	*A_{228}: Triple*
A_{25}: Homerun	*A_{213}: Single*	*A_{221}: Double*	*A_{229}: Triple*
A_{26}: Homerun	*A_{214}: Single*	*A_{222}: Double*	*A_{230}: Triple*
A_{27}: Homerun	*A_{215}: Single*	*A_{223}: Double*	*A_{231}: Triple*
A_{28}: Homerun	*A_{216}: Single*	*A_{224}: Double*	*A_{232}: Triple*
A_{233}: Impossible	*A_{241}: Impossible*	*A_{249}: Impossible*	*A_{257}: Impossible*
A_{234}: Single/walk	*A_{242}: Single*	*A_{250}: Double*	*A_{258}: Impossible*
A_{235}: Single/walk	*A_{243}: Single*	*A_{251}: Double*	*A_{259}: Impossible*
A_{236}: Impossible	*A_{244}: Walk*	*A_{252}: Impossible*	*A_{260}: Impossible*
A_{237}: Single	*A_{245}: SB, single*	*A_{253}: Double*	*A_{261}: Single/walk*

A_{238}: Single	*A_{246}: Single*	*A_{254}: SB, double*	*A_{262}: Walk*
A_{239}: Single	*A_{247}: Single*	*A_{255}: Double*	*A_{263}: Walk*
A_{240}: Single	*A_{248}: Single*	*A_{256}: Double*	*A_{264}: Single/walk*

The final three matrices, D_0, E_1, and F_2, are all transitions that result in three outs and the end of the half inning. Although it is technically possible for runs to score on a play that results in 3 outs these plays are few and far between and nearly impossible to model. Because of this, all plays that result in the end of an inning will have zero run potentials.

$$D_0 = \left(\begin{array}{c|c} & *,3 \\ \hline 0,0 & 0 \\ 1,0 & 0 \\ 2,0 & 0 \\ 3,0 & 0 \\ 12,0 & D_{05} \mid 0 \\ 13,0 & D_{06} \mid 0 \\ 23,0 & D_{07} \mid 0 \\ 123,0 & D_{08} \mid 0 \end{array}\right)$$

Figure 2.10: Finalized Version of Transition Matrix D_0

$$E_1 = \left(\begin{array}{c|c} & *,3 \\ \hline 0,1 & 0 \\ 1,1 & E_{12} \mid 0 \\ 2,1 & E_{13} \mid 0 \\ 3,1 & E_{14} \mid 0 \\ 12,1 & E_{15} \mid 0 \\ 13,1 & E_{16} \mid 0 \\ 23,1 & E_{17} \mid 0 \\ 123,1 & E_{18} \mid 0 \end{array}\right)$$

Figure 2.11: Finalized Version of Transition Matrix E_1

$$F_2 = \begin{array}{c|c} & *,3 \\ \hline 0,2 & F_{21} \mid 0 \\ 1,2 & F_{22} \mid 0 \\ 2,2 & F_{23} \mid 0 \\ 3,2 & F_{24} \mid 0 \\ 12,2 & F_{25} \mid 0 \\ 13,2 & F_{26} \mid 0 \\ 23,2 & F_{27} \mid 0 \\ 123,2 & F_{28} \mid 0 \end{array}$$

Figure 2.12: Finalized Version of Transition Matrix F_2

It is through this Markov Chain framework that simulations of half-innings, complete games, and entire seasons can now be performed. Before modern computing, the size of the necessary transition matrices along with the number of repetitions required to complete even a single 9 inning game of baseball would have been limiting factors in this process.

2.4 Advantages of Simulation Models

Simulation models, since they generally require a number of replications of a given statistical experiment, provide statistical estimates of the numerical quantities of interest; they do not provide solutions. On the other hand, when a given analytical model is solved for a certain set of input data values, the solution will always be the same for that same set of input values. In many situations, the primary reason for using simulation models is that development of a solvable analytical model requires simplifying assumptions which are unrealistic. The advantage of a simulation model is that it is readily understandable by a non-mathematician and that it allows determination of the variance of the statistical quantities of interest.

While baseball is an exceedingly complicated statistical process, it enjoys a relatively simple, repetitive pattern for the basic performances of nine offensive batsmen against the opposing pitcher supported by eight defensive players. Thus, a sophisticated computer model of the game may be designed in which all possible interactions of play are precisely simulated. Chance determines each successive event in that every play is selected by means of a random-number generator. The computer, in effect, functions only as a scorekeeper; it summarizes all data for the amount of games one chooses to simulate. By this so-called 'Monte Carlo' method, it is possible to estimate the average number of wins a given team will accumulate in a season and estimate the probability with which one team will beat another team in a single game matchup.

However, just because baseball enjoys a structure that can be modeled as a Markov process does not mean that accurate predictions will automatically result. The preciseness of any Markov model is heavily contingent upon the probabilities contained in the transition matrix. The next step in the simulation process is to fine tune these probabilities to best reflect empirical results from years of baseball data.

Chapter 3

Situational Factors Affecting Transition Probabilities

The probabilities contained in the transition matrix for a game of baseball define the chance of moving from one base-out state to another. Except for the case of a stolen base, all transitions require actions by both the pitcher and the opposing batter. Assuming that it is the abilities of these two players that best determine the likelihood of the next transition, it is natural to consider which currently available baseball statistics best reflect a player's (batter or pitcher) true ability and not just chance variation or luck. In addition to finding the most stable measures of player ability, it is also important to search for particular game situations which tend to offer players a consistent edge over their opponents and to update modeling accordingly.

As the focus of this research is aimed at creating a player based approach to baseball simulation, it is important to find those statistics that not only give the best estimates of player ability but also statistics that are directly interpretable. Interpretability of statistical measures is important because it allows for a straightforward and natural implementation of these variables into a simulation routine. A black box model with strong predictions but whose inputs have no clear relationship with the outcome has only solved part of the problem. However, with interpretable inputs and a reasonably deterministic simulation model, successful predictions can more easily be attributed back to certain key inputs. With so much data recorded on every game, player, and team in

Major League Baseball, the focus of Chapters 3 and 4 is as much an attempt to dismiss irrelevant information as it is a search for meaningful prediction measures.

3.1 Bias and Evolvement of Baseball Statistics

The Batting Average (*AVG*) is the oldest and most widely used standardized estimator of offensive performance. It is the rate at which hits (*H*) are produced per at-bat (*AB*) and although it does not directly measure run production, every year each league crowns a so-called "Batting Champion" on the basis of this statistic. Critics of using a player's *AVG* as the gold standard of offensive efficiency argue that hits should be weighted according to their value. Obviously a home run is a more valuable hit than a single, but *AVG* counts all hits equally. Thus, a statistic was constructed that weights each hit by the number of bases reached by the batter (single, double, triple, homerun), and the corresponding rate is called the slugging percentage (*SLG*). Neither of these estimators included walks (*BB*) in their calculations. This situation is remedied by the on-base percentage (*OBP*), which is the rate of getting on base through hits, walks, or hit by pitch (*HBP*).

$$AVG = \frac{h}{AB}$$

$$SLG = \frac{(1B) + (2 \times 2B) + (3 \times 3B) + (4 \times HR)}{AB}$$

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

All of the above statistics are also used to quantify pitching ability by adding the suffix "against" and all of these pitcher statistics suffer from the same drawbacks as the ones just discussed for batters. It should be noted, however, that whereas higher values of *AVG*, *SLG*, and *OBP* indicate stronger offensive performance, lower values of *BAA* (batting average against), *SLGA*, and *OBPA* indicate a stronger pitching performance.

A statistic that is unique to pitchers and has long been viewed as the best estimator of pitching ability is the earned run average (*ERA*). *ERA* is defined as the amount of earned runs (*ER*) given up by a pitcher per nine innings pitched (*IP*)

$$ERA = 9 \times \frac{ER}{IP}$$

An earned run is any run for which the pitcher is held accountable (i.e., the run scored as a result of normal pitching, and not due to a fielding error or passed ball). Similar to the batting average for batters, the earned run average statistic for pitchers has some built in biases as well. First, a pitcher is only charged with the number of runners that reached base while he was pitching. So, if a pitching change occurs, the new pitcher inherits any runners that are on base at the time, and if they later score, those runs are charged (earned or unearned) to the prior pitcher. Earned run averages are also generally lower among closers than starting pitchers and strictly using *ERA* as a comparison across different types of pitchers (starters, relievers, or closers) seems unfair. Closers only pitch if their team is tied or winning and along with relievers, rarely pitch for more than 1 or 2 innings. Starting pitchers must exert themselves for a much longer stretch of the game and have the added disadvantage of giving batters a chance to "learn" their tendencies from one at-bat to another.

3.2 Measuring Chance vs. Ability

Whether it is for batters or pitchers, all statistics are imperfect measurements of player ability. Batter A may have a higher slugging percentage than Batter B, but Batter B could have a better batting average than Batter A, who is the better offensive player? Comparing more statistics between the two players may provide a clearer answer but how to quantify the magnitude of this difference leads back to the problem of interpretability. Instead, searching for those statistics that best balance the importance of accurate measured ability with a straightforward interpretability easily implemented into a simulation routine becomes the best solution.

In Albert (2001) and Albert (2006) an answer to the first part of this problem is addressed. Random effects models were used to analyze the ability to luck ratios of various batter and pitcher statistics. The random effects model utilized can be described succinctly as follows: assume N players with different abilities and assume these abilities $p_1, \ldots, p_n$ come from a probability distribution $g(p)$. Think of these abilities as the true probabilities of a successful (unsuccessful) outcome for a batter (pitcher). Once the probabilities are known, the actual numbers of successes $y_1, \ldots, y_n$ have binomial distributions, where $y_i \sim Binomial(n_i, p_i)$. Because the only statistics analyzed were those that vary between 0 and 1 it is convenient to assume the ability distribution has the beta form: $g(p) \propto p^{k\eta-1}(1-p)^{k(1-\eta)-1}$, $0 < p < 1$, where η represents the mean or average ability of all players and k gives some indication of the spread of the abilities. Small values of k indicate there is a wide range of abilities among players; consequently, a statistic with a relatively small k is a better measure of true ability. After fitting the random effects model the estimates of the quantities k and η for each of eight hitting measures are given using data for all players in the 2003 baseball season with at least 100 at-bats.

	Fitted Ability Distribution				
Statistic	K	5^{th} Percentile	Median	95^{th} Percentile	Order
SO Rate	45	0.094	0.173	0.278	More Ability
HR rate	70	0.012	0.039	0.088	
BB rate	81	0.042	0.083	0.143	
BB rate (without Bonds)	85	0.040	0.079	0.136	
OBP	209	0.278	0.330	0.384	
OBP (without Bonds)	230	0.280	0.330	0.382	
AVG	486	0.235	0.267	0.301	
Double + Triples Rate	495	0.055	0.072	0.093	More Luck
Singles Rate	530	0.186	0.215	0.245	

Table 3.1: Ability distributions for fitted random effects model for each of eight batting statistics.

Comparable results are also found after finding simple year to year correlations for each of the above batting statistics.

Batting Statistic	Year-to-Year Correlation
Singles Rate	.301
Doubles + Triples Rate	.303
Batting Average	.362
OBP	.607
HR Rate	.682
BB Rate	.788
SO Rate	.819

Table 3.2: Correlation of batting statistics between the 2002 & 2003 seasons

For pitchers, the results mirror the results on batting data reported above. Pitchers have varying probabilities to strike out batters, allow runs and walk batters. In contrast, pitchers' in-play batting average and batting average against are seemingly much more influenced by chance variation than anything else.

Of all the statistics analyzed, only *AVG* and *OBP* give a quantity that is useful in a simulation that progresses at-bat by at-bat. From the results of Albert's study it is clear *OBP* is the statistic which provides a more stable and more accurate measure of offensive ability. However, *OBP* still has the drawback of treating all hits as equal. From a subject matter perspective, just knowing the probability a batter gets on base is not enough. Some batters hit for power (homeruns) while some hit for average (singles). Making use of the frequency with which a player gets on base via a walk, single, double, triple, or homerun is critical information. And although an ability distribution for *SLG* cannot be fit like it can for the other statistics above (*SLG* varies between 0 and 4), the advantageous makeup of the statistic can still be utilized in a simulation routine by simulating at-bats at two different levels. The first level would be computing the chance a batter gets on base against a certain pitcher via the two players' *OBP* and *OBPA* respectively. If it is decided that the batter gets on base, the second level of computing which type of advance the batter made is now needed. Before moving forward to the details of the second level though, a closer look needs to be given to the details of the first level for computing an expected *OBP* between any given batter-pitcher matchup.

3.3 The Log5 Method

Consider the following situation: if a .310 hitter faces a pitcher with a batting average against (*BAA*) of .290 what should the resulting expected batting average be? At first glance it may appear that the result should be the average of the two, i.e. .300. Upon reflection, however, this solution is flawed. Assuming a .260 league batting average against, the .290 pitcher is worse than average. Therefore, the batter should hit for a higher average against this pitcher than his overall average. This reasoning seems logical but it cannot quantify just how much better one can expect the batter in this situation to perform. To remedy this situation, James (1981) devised the following "Log5" formula that he used to predict the probability for a hitter with batting average *BA* of getting a hit against a pitcher with a batting average against *PA*, in a league that has a league average batting average against of *LA*

$$ExpAvg = \frac{\frac{BA * PA}{LA}}{\frac{BA * PA}{LA} + \frac{(1 - BA)(1 - PA)}{1 - LA}} \tag{3.1}$$

Although this formula was originally intentioned to predict the probability of a batter generating a hit, it does not seem far fetched to assume that this same model can be used to predict the probability of a batter getting on base by adjusting the above inputs to be a batter's *OBP*, a pitcher's *OBPA*, and a league average *OBPA*.

3.3.1 Extending the Log5 Method

A characteristic that is implicit in this model, however, is the equal weights given to the three above inputs. Since the coefficients are all 1, each input is assumed to be just as important as the others in predicting an expected average. Although this very well could be the case, it would be wise to see if the model coefficients can be tweaked in such a way to give better estimates. In essence, this fitting would be seeking to answer the question of who is more responsible for the outcome of an at-bat, the batter or the pitcher. In order to do so, the form of (3.1) needs to be made more amenable to such analysis. After some manipulation, it is shown that (3.1) can be expressed equivalently in a logit form.

$$ExpAvg = \frac{\frac{BA * PA}{LA}}{\frac{BA * PA}{LA} + \frac{(1 - BA)(1 - PA)}{1 - LA}} = \frac{1}{1 + \left(\frac{1 - BA}{BA}\right)\left(\frac{1 - PA}{PA}\right)\left(\frac{LA}{1 - LA}\right)} = \frac{1}{1 + e^{\log\left[\left(\frac{1-BA}{BA}\right)\left(\frac{1-PA}{PA}\right)\left(\frac{LA}{1-LA}\right)\right]}} = 1 - p$$

$$\Rightarrow \frac{p}{1-p} = e^{\alpha} \Rightarrow p = (1-p)e^{\alpha} = e^{\alpha} - pe^{\alpha} \Rightarrow (1+e^{\alpha})p = e^{\alpha} \Rightarrow p = \frac{e^{\alpha}}{1+e^{\alpha}}$$

$$1 - p = \frac{1}{1+e^{\alpha}}$$

where

$$logit(p) = \alpha_C + \alpha_B * logit(BA) + \alpha_P * logit(PA) + \alpha_L * logit(1\text{-}LA) \tag{3.2}$$

In this form, it is now possible to find estimates for α_C, α_B, α_P, and α_L via logistic regression. To accommodate this, data from 8900 plate appearances with roughly an even amount taken from both leagues over the 2001-2006 seasons were recorded. The *OBP* of the batter, *OBPA* of the pitcher, and league average *OBPA* are the explanatory variables with the outcome variable defined as whether the batter failed or succeeded in reaching base in his at-bat (0 or 1). A model with no intercept term was used and the regression coefficients were all found to be significant at the $\alpha = .05$ level. The values of the regression coefficient estimates are as follows: $\alpha_B = .8617$, $\alpha_P = .5693$, $\alpha_L = -.4400$. Notice that there is no real need to constrain $\alpha_B + \alpha_P + \alpha_L = 1$ because the coefficients sum to .9910 themselves. The results from this analysis are interesting in that they imply about a 60/40 split in batter versus pitcher ability in determining the outcome of an at-bat. Meaning the skill of the batter plays a greater role within each at-bat than the skill of the pitcher. These results are also encouraging when viewed in comparison with results found in a study using completely different techniques to answer the same question.

3.3.2 An Alternative Approach to Analyzing the Batter-Pitcher Subgame

Baseball researchers have long used the concept of Net Expected Run Value (NERV) to analyze a player's performance and a manager's strategic decisions. NERV, typically represented in a 3X8 table, is the run expectation for the remainder of a half inning given the number of outs and the configuration of men on base.

		Runners							
		None On	1^{st}	2^{nd}	3^{rd}	1^{st}, 2^{nd}	1^{st}, 3^{rd}	2^{nd}, 3^{rd}	Bases loaded
Number of Outs	0	.51	.85	1.11	1.3	1.39	1.62	1.76	2.15
	1	.27	.51	.68	.94	.86	1.11	1.32	1.39
	2	.10	.23	.31	.38	.42	.48	.52	.65

Table 3.3: Expected runs scored in remainder of inning from each of 24 possible runner/out states; data collected from 1987 MLB season.

The value of a given plate appearance, for example, can be calculated as the change in NERV caused by the plate appearance plus any runs scored on the play. In previous work, the total value of the change in NERV has been assigned to the batter or the pitcher (Lindsay, 1963) and then the total value of all of the batters plate appearances or pitcher's batters faced was equal to their total value. For example, with no runners on base and no outs, the expected run value is approximately 0.51. If the batter in this situation hits a homerun, he scores one run and the expected run value, when the next batter comes to plate is still 0.51. In this case, the NERV is simply 1 because there has been no change in the expected run value and the batter has scored 1 run. In past analyses, researchers have focused either on batters or pitchers and would have credited the focus of their analysis with 1 point of NERV. Many plate appearances involve only the batter and the pitcher.

Strikeouts, homeruns and walks are all principally affected only by the batter and pitcher. Other plays such as any ball put into play, involves multiple defensive players and runners if there are men on base. The change in the run expectancy then is not due entirely to the outcome of the duel between the pitcher and the batter, but rather, that duel is a sub game within the game. For example, a batter may hit a line drive that would ordinarily result in a single; in this case, it would be fair to say that the batter got the better of the pitcher. If however, a fielder makes a great defensive play on the batted ball and creates an out instead of a man on first, the fielder deserves credit for the change in NERV that results, not the pitcher who lost the sub game against the batter. The quantitative problem then becomes splitting the change in NERV in such a way that a batter who put a ball in play that is hard to field gets positive credit, even if a defensive player makes an unlikely play on the ball and the change in NERV for the entire play is negative. Additionally, as past analysis has used NERV to focus simply on pitchers or batters, the results of the analysis are not comparative across pitchers and batters. There is no way, for example, to determine using a NERV analysis whether a great pitcher or a great batter has more total effect on a team over the course of a season. Splitting NERV between batters and pitchers objectively, allows for the direct comparison of total NERV between pitchers and batters to determine which will have a greater impact on a team. Alamar (2006) develops a system that estimates NERV while controlling for many game states not previously considered. For example, different parks have different dimensions that change the run expectancy. Coors Field in Colorado, for example, is known as a hitter's park. It is therefore important to adjust NERV for the park where the game being analyzed takes place. The system then isolates the pitcher/batter sub game by estimating the expected outcome of any ball in play given the hit type, trajectory, distance and speed of the hit. Using the expected outcome of the ball in play, the effect that the hitter and pitcher have on that expected outcome is estimated using measures of skill for the players involved, independent of other defensive players. Finally, the results show that in an at bat between an average batter and an average pitcher, the batter should accrue 62% of the resulting expected NERV change and the pitcher should accrue 38% of the expected NERV change. Intuitively this may be true because the pitcher has significantly more time to plan his action than the batter, who must assess a pitch in a split second. As the batter seems to have a more difficult task, the skill of the batter has a greater effect on the outcome and thus receives more credit for the outcome of the plate appearance. These results are quite similar to results found from the previous logistic regression analysis in which the relative weights assigned to the batter/pitcher sub game are $\frac{\alpha_B}{\alpha_B + \alpha_P} = .602$ and $\frac{\alpha_P}{\alpha_P + \alpha_B} = .398$ respectively.

3.3.3 Comparing Simulated Results

We now have two different methods to estimate the probability of getting on base for a batter with ability *OBP*, against a pitcher with ability *OBPA*. The *Log5* and Weighted Log5 (*W-Log5*) methods provide similar estimates across average ability levels (.330 for both *OBP* and *OBPA*) of batters and pitchers as can be seen from Figure 3.2. However, as the ability of either the batter or pitcher deviates from their respective norms (extremely poor or extremely good players) their estimates start to diverge. The *Log5* method appears to give a much wider range of estimated *OBP* compared to more homogenous predictions given by the *W-Log5* method. Because there are many more occurrences of “average” players facing each other than matchups involving exceptional players (bad or good) the practical significance of the difference in prediction methods is not directly clear. Comparing results after an entire season of game simulations though, may better reveal the magnitude of improved accuracy gained from the *W-Log5* method.

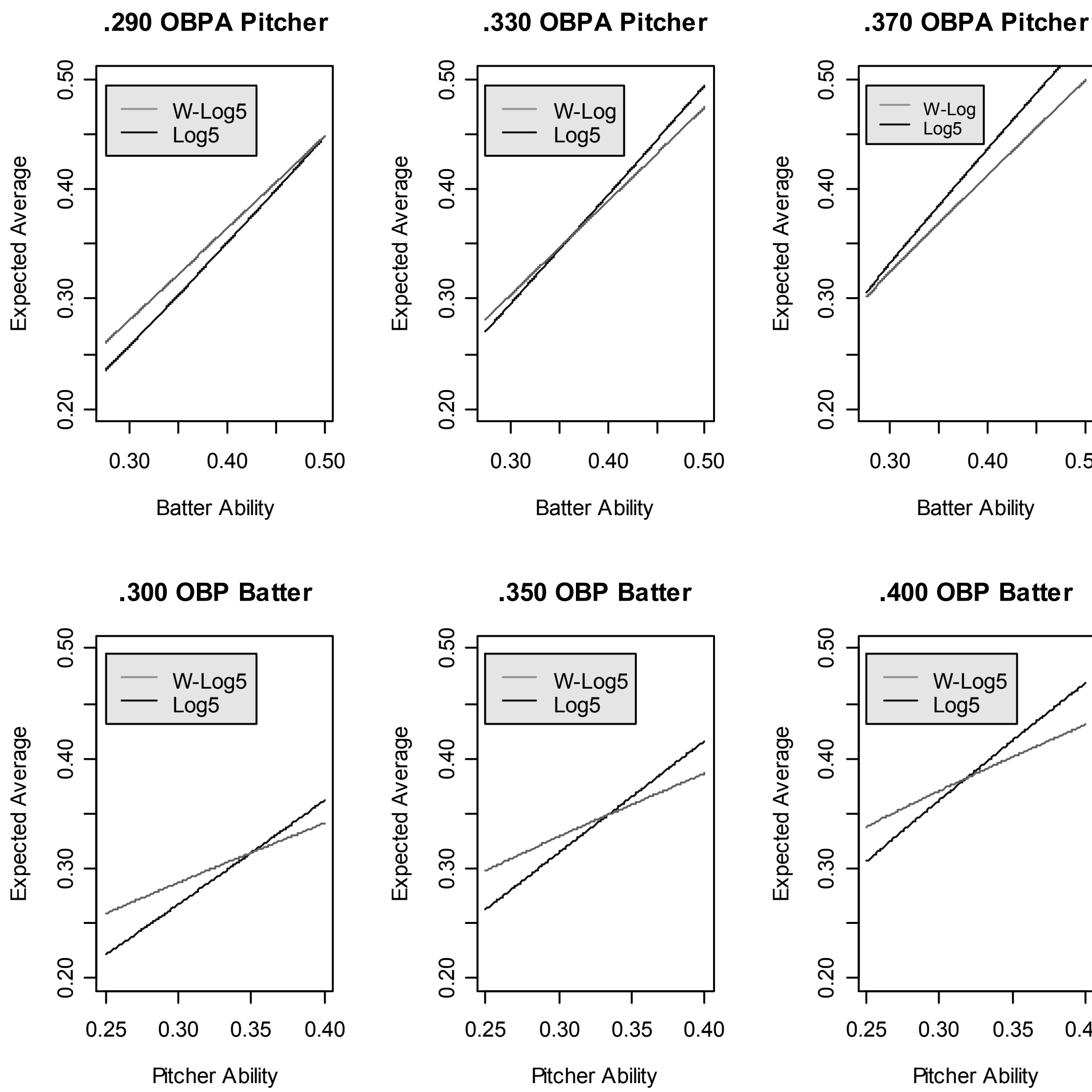

Figure 3.1: OBP predictions of various scenarios using the *Log5* and *W-Log5* methods

3.4 Situational Affects

Like most all sports, baseball is a game in which action takes place in various settings. A follower of baseball often may hear how baseball players perform differently dependent on their current situation, such as playing at home or on the road, on grass or on turf, in clutch situations, and ahead or behind in the count. With such an abundance of situational talk such as this, fans may get the misleading impression that much of the variability in players' hitting performance can be explained by one or more of these situational variables. For example, an announcer may state that a particular player was likely to strike out because he was behind in the count and was facing a left-handed pitcher. In baseball one can now investigate the effect of various situations, as hitting data is recorded in very fine detail.

Albert (1994) performed an analysis that looks at the hitting performance of major league regulars during the 1992 baseball season to see which situational variables are "real" in the sense that they explain a significant amount of the variation in hitting of the group of players. Bayesian hierarchical models are used in measuring the size of a particular situational effect and in identifying players whose hitting performance is very different in a particular situation. For a particular situation it is of interest to look for a general pattern across all players. Baseball aficionados believe that most hitters perform better in various situations (James, 1981). In particular, managers often make decisions under the following assumptions: batters hit better against pitchers throwing from the opposite side, batters hit better at home, batters hit better during day games, batters hit better on artificial turf than on grass. The table below displays some results from the Albert analysis and shows that batters on average hit 20 points higher when facing a pitcher of the opposite arm, 11 points higher when facing a groundball pitcher, and 8 points higher when batting at home. Being ahead in the count versus behind is also shown to be significant but this information is not useful in a simulation that advances only at-bat by at-bat and not pitch by pitch.

			One unit = .001			
SITUATION	E(μ_α)	E(σ_α)	Q_1	Median	Q_3	$Q_3 - Q_1$
Grass - Turf	-.002	.107	-17	-3	15	32
Scoring – None on/out	.000	.108	-13	0	17	30
Day – Night	.004	.105	-13	2	16	29
Pre/A.S. – Post/A.S.	.007	.101	-9	3	17	26
Home – Away	.016	.103	-8	8	21	29
Gball - Fball	.023	.109	-7	11	24	31
Opp - Same	.048	.106	5	20	32	27
Ahead – 2 Strikes	.320	.110	104	123	142	38

Table 3.4: Posterior means of parameters of population distribution of the situation effects and summary statistics of the posterior means of the batting average differences across all players.

3.4.1. Handedness

Conventional baseball wisdom states that batter and pitcher handedness is believed to be an important factor in the outcome of an at-bat because of the trajectory of the ball. For a right handed (left handed) batter, it is easier to see the ball as it is released from a left handed (right handed) pitcher. This slight change in the angle of delivery allows the batter a split second more to pick up the trajectory of the ball and hence increase his chances of making solid contact with the pitch. As noted by Albert (1994) and James (1981), this handedness effect appears to be a significant factor in determining the outcome of an at-bat and should be accounted for when simulating.

3.4.2. Home vs. Away

Another seemingly important factor that should be incorporated into the modeling process is home field advantage. Three factors appear to impart an advantage to the home team: a physiological factor due to travel fatigue that affects the visiting team, a psychological factor due to encouragement of the home team and intimidation of the visiting teams by the home crowd and a tactical factor due to familiarity of the home team with the venue. In a study by Stefani (2007), three sports (rugby union at 25.1%, soccer at 21.7%, and NBA basketball at 21.0%) had the highest regular season home advantage measured by the fraction of home wins minus the fraction of home losses. The continuous nature of these sports is consistent with greater player fatigue than most other sports and logically greater home advantage for the more rested home team. Three sports were at the middle of the range of home advantages: Australian Rules football at 18.8%, NFL football at 17.5%, and US college football at 16.6%. These three sports provided much more stoppage time and player substitutions with resulting reduced excitation of the three factors. The lowest regular season home advantages were for NHL hockey at 9.7% and for MLB baseball at 7.5%. These sports also had the least excitation of the physiological, psychological, and tactical factors. For all of these sports, a home field advantage factor, *h*, was computed as follows:

$$h = \frac{1}{N}\sum_{k}(w_i - l_i)$$

where w_i is the number of home wins for a team i, l_i is the number of home losses for team i, and N is the total number of home games played for teams $1 \ldots k$. The sample of baseball data consisted of 30,347 games taken over 13 seasons. Although small compared to other sports in this study, the home effect for baseball still appears real and should also be accounted for when modeling.

3.4.3. Clutch Hitting

In sports there is also a general interest in the ability of athletes to perform well in "clutch" or important situations. In baseball, there will be particular times during a game when a batter has opportunities to produce runs, and it is very desirable for him to do well in these clutch situations. For example, consider a single; the value of a single, from the viewpoint of scoring runs, varies greatly depending on the runner's situation. A single when the bases are loaded is much more valuable to the team than a single with the bases empty. Although baseball players have different levels of performance in clutch situations, most researchers have found little evidence for players to have a higher level of ability in these important clutch situations. Examples of these studies include Thorn & Palmer (1985) and Skoog (1987). Albert (2001) used 1987 NL data to look for the existence of clutch hitting ability. He first defined the value of a plate appearance as the difference in run potentials before and after the plate appearance. For all players, he computed the average value of a plate appearance for all situations with different runners on base and number of outs. He fit a random effects model and found some situations where batters tend to be more or less successful but found little evidence that players had an ability to perform better in particular situations. In a recent analysis, Silver (2006) revisits this clutch ability question with a regression model for hitters that combines the number of runs he produces, the number of wins he created for his team, and his opportunities to get hits in clutch situations. He develops a statistic called "clutch" that measures the results of a player's clutch hitting over and above what would be predicted by his batting statistics, opportunities, and run-scoring environment. He makes the interesting conclusion that only10% of clutch-hitting performance can be attributed by skill. Albert (2007) revisits his work of (2001) and fits a random effects model with a bias component that is used to represent hitting data in "clutch" and "nonclutch" situations. All of these studies walk away with one of two possible conclusions: clutch hitting does not exist, or, clutch hitting exists but its size is quite small. Because of the amount of studies that have found little or no evidence of clutch hitting in baseball, no effect for clutch hitting will be incorporated into the simulation model.

3.4.4. Streakiness

Another aspect of most sports that is often spoken of as if it were a proven reality is the idea of streakiness. All baseball fans know that players go through periods where they never seem to make an out and other periods where nothing will get them to base. There is no doubt that "hot streaks" and "cold streaks" do occur. The question explored in an analysis by Albright (1993), however, is whether these streaks occur more (or less) frequently than would be predicted by a probabilistic model of randomness. In this study, the author examined the records of many regular MLB players through four seasons, 1987-1990, and used several statistical methods to check for streakiness. These include standard methods such as the runs test, as well as a more complex logistic regression model with several explanatory variables. The runs test classifies each at-bat as a success or failure, and then counts the season-long number of runs. Fewer runs than would be expected under a model of randomness then provide an indication of streakiness. In contrast, more runs than expected indicates stability, where a batter tends to do better after failures and worse after successes. In his analysis Albright lets N_S, N_F, N, and R be the number of successes, the number of failures, the number of at-bats, and the number of runs for a given player in a given season. The runs test is then based on the statistic

$$Z = \frac{[R - E(R)]}{\sigma_R}$$

where

$$E(R) = \frac{N + 2N_S N_F}{N} \text{ and } \sigma_R = \sqrt{\frac{2N_S N_F (2N_S N_F - N)}{N^2(N-1)}}$$

Under randomness Z is approximately $N(0,1)$. The drawback to the preceding method is that it uses only the time series of successes and failures in individual at-bats. As stated earlier, the amount of streakiness could certainly be affected by situational variables. Therefore, a probabilistic model is also used by Albright that incorporates these variables, as follows. Let X_n be 1 or 0, depending on whether a player is or is not successful on at bat n, and let $p_n = P(X_n=1)$. Next, define a logistic regression model of the form

$$\ln\left(\frac{p_n}{1-p_n}\right) = \alpha + \beta Y_n + \sum_{k=1}^{K} \gamma_k Z_{kn}$$

where Y_n is related to the batter's recent history of success and Z_{kn} is an explanatory variable relating to the situational variables in effect during at-bat n. For example, Z_{1n} might be 1 or 0, depending on the handedness of the pitcher, Z_{2n} might be the pitcher's *ERA*, and Z_{3n} is 1 or 0 depending on whether any runners are in scoring position. In contrast, Y_n, indicates the player's success in recent at-bats. 8 different possibilities were analyzed. The first method gave Y_n as the number of successes in the most recent k at-bats, for k=1,2,3,6,10, and 20. In addition, Y_n was also treated as an exponentially weighted sum of the batter's last 20 *X's*, specifically $Y_n = \sum_{i=1}^{20} \delta^{i-1} X_{n-i}$ for δ = .80 and δ = .95. The purpose here is to allow the probability of success to depend on the last 20 at-bats, but to give more weight to the more recent at-bats.

Based on all of these methods, there is no doubt that a certain number of players exhibited streakiness in certain years. But Albright also finds the evidence suggests that the behavior of all players examined, taken as a whole, does not differ significantly from what would be expected under a model of randomness. Furthermore, none of the players examined in the study exhibited unusually streaky behavior over the entire four-year period. Similar to the story of clutch hitting in baseball, if streakiness does still happen to exist, its effect is very small and should not be included in a simulation of a game.

3.4.5. Momentum

Another artifact within most sporting events that is similar to streakiness and as equally accepted as fact by most observers is the idea of momentum. The nature of baseball is that it is characterized by intermittent periods of action, separated by periods of inactivity. To a fan of the game, it often appears that this results in a sense of momentum, in the sense that the results of previous plays seem to have a relationship to ones that follow. This is in direct contrast to the assumption that baseball may be viewed as a Markov process. The existence of such a momentum effect can be explored using play-by-play data. In work by Sela (2007) multinomial logistic and negative binomial regression models are used to explore the question of whether momentum effects appear in MLB games. Their findings are that the evidence for such effects is weak at best, implying little chance of making meaningful improvements in the prediction of future events using models that include such effects.

3.5 Conclusion

Much statistical research has already been done with applications for baseball. Sorting through these findings and reporting the relevant information has been the focal point of this chapter. Since the batter/pitcher matchup is the means through which any baseball game advances, it is essential to simulate these matchups particularly well in order to produce realistic and accurate results. And although it is fairly clear which statistics currently available best represent true ability rather than luck, and which situational variables are significant (handedness and home-field) in determining the outcome of an at-bat, some questions still remain. Specifically, is there a tendency for some batters (pitchers) to perform significantly better or worse against certain pitchers (batters). In other words, is there evidence of a significant batter/pitcher interaction effect in baseball? Currently, no such formal work on the subject has been performed and Chapter 4 will analyze specifically this issue.

Chapter 4

Batter-Pitcher Matchups in Baseball and Small Sample Decision Making

In the book *Three Nights in August: Strategy, Heartbreak and Joy Inside the Mind of a Manager* author Buzz Bissinger chronicles a three-game Cubs/Cardinals series played in August 2003 from Cardinals manager Tony LaRussa's perspective. A passage from this book that is of interest to both statisticians and baseball fan's alike relates how LaRussa uses individual batter-pitcher matchup data to influence his decisions on his game day lineups. His basic philosophy is explained in the following passage:

> La Russa pays special attention to the individual matchups, an essential ingredient of his approach to managing … The term *bench player* doesn't really apply to the Cardinals, because LaRussa so frequently plugs utility players into the lineup based on little opportunities he unearths by sifting through the results of their previous experience with players on the opposing team. These individual matchups are so integral to his strategy that he copies them onto 5-by-7-inch preprinted cards that managers normally use to make out the game's lineup. With ritualistic precision, he folds the cards down the middle 10 minutes before game time and then slips them into the back pocket of his uniform. During a game, he pulls them out continually, almost like worry beads, peering at them as if in search of evidence that everything is fine, that he is doing *exactly what he needs to be doing*. More practically, he refers to them when deciding who to bring on in relief or who may be the best matchup to pinch-hit.

Bissinger notes that LaRussa knows that matchups are not fool proof but still has reason for his beliefs. After all, there is definite baseball logic that would support considering such data. Pitchers all release the ball from different angles, some of which may make it easier for certain hitters to pick the ball's trajectory, pitchers (and batters) also have particular tendencies (fastball pitchers, location pitchers, fastball hitters, etc.) which may match up well with the strength of the current opponent. In all of his years managing baseball games, Larussa has observed that, “There are some hitters who, never mind their mediocre batting averages, simply tag the living crap out of some pitchers. Conversely, there are pitchers, despite soggy ERAs, who simply do well against particular high-stroke hitters.”

The question that now needs to be answered is whether or not these observed matchup effects are statistically significant. Because the sample size of most all matchups found among MLB players is necessarily small compared to a player’s overall amount of career plate appearances, how much weight or consideration should be given to a few seemingly significant matchups? A quick study in basic statistics reveals that extremely poor or good performances against certain opponents may just be due to bad or good luck and nothing more. For example, it was reported that Los Angeles Dodger player Kenny Lofton was rested by Los Angeles Dodger manager Grady Little during the team's August 29, 2006 game against the Cincinnati Reds because of Lofton's record against the scheduled Reds pitcher, Eric Milton. Lofton, whose season-long batting average was .308 at the time and whose career batting average was .299, had just one hit in 19 career attempts against Milton. Although Lofton’s .056 career average versus Milton is drastically lower than his lifetime batting average, there is still about a $1/130 \approx .009$ chance that a hitter of Lofton's ability would have one hit in 19 attempts just by chance. Lofton's been around a long time and opposed many pitchers. Is he really having a tough time with Milton? Or has he just been unlucky? In what follows we look more closely at baseball pitcher-hitter matchups using a hierarchical beta-binomial model.

4.1 The Batter-Pitcher Matchup: A Simple Binomial View

The simplest statistical approach to matchup data is to view the number of hits Y in a sequence of n trials between a particular batter and pitcher as a binomial random variable with probability of success p. One can infer p from the observed data or in the present case assess whether a particular data set is consistent with a particular hypothesis about p. If we consider the Lofton-Milton matchup and assess whether the observed data ($Y = 1$ in $n = 19$ trials) is consistent with $p = .300$ (Lofton's approximate ability level), then we find $\Pr(Y = 1) = .009$ and $\Pr(Y \leq 1) = .01$. These data suggest that a 1-for-19 run is quite unusual for a .300 hitter.

This previous calculation, however, omits an important consideration, namely the multiplicity of possible matchups. It is unusual to find 1 success in 19 pre-specified trials for an event with success probability .300, but it is not necessarily unusual to find 1 success in 19 trials if we were to consider many different sets of 19 trials before settling on the one to look at. In baseball terms we need to account for the fact that just by chance Lofton will have a higher batting average against some pitchers and a lower batting average against others. One way to learn about this is by simulation. During the course of Lofton's career he has faced many different pitchers. If we take the 100 pitchers with whom Lofton has the most career matchups and simulate Lofton's performance assuming constant probability of success .300, then we find it is not at all unusual to have one pitcher against whom Lofton's performance is 1-for-19 or worse; in fact, this happened in 20% of the simulated careers. These simulation results are consistent with other findings as well. On-line baseball writer Dan Fox gathered data on more than 30,000 individual pitcher-batter matchups (Fox, 2005). His conclusion was that there is nothing in the data that one would not expect to see just by chance.

The simple binomial calculation and simulation study suggest that there may be much less to batter-pitcher matchups than baseball fans and baseball managers believe. All of the observed data appear to be consistent with a simple binomial model. Such a conclusion however is almost certainly inappropriate. The failure to reject a null model does not mean the null model is correct. Most basic statistics courses emphasize that our ability to reject a null model or null hypothesis depends on the true size of the effect under study and the sample size. It seems likely here that those two factors are working against us - the variability of batter's ability from pitcher-to-pitcher is likely not very big and the sample sizes are generally fairly small. In fact, we'd argue that one can be fairly certain that the binomial analysis is not powerful enough to detect failures of the simple model. The reason for feeling this way is that there are some well-studied baseball phenomenons that tell us the simple binomial model is wrong. First, there is many years worth of data across many players that show batters perform better against pitchers with the opposite dominant hand (as discussed in Chapter 3); in other words, left-handed batters have a higher average against right-handed pitchers than they do against left-handed pitchers (and vice versa for right-handed batters). This simple fact means a left-handed batter like Lofton should have a higher average against some pitchers (right-handed ones) than others. The second key fact is that some pitchers are better than other pitchers. Thus batters perform worse against someone like Johan Santana than against someone like Byung-Hyun Kim. Thus we'd expect at least some variation in success probability due to the quality of the opposing pitcher. Other factors may also be important including the site of the game (some stadiums are easier to hit in than others).

The two arguments presented so far, the simple data analysis arguing in favor of the null binomial model and other information arguing against it, cause us to consider a hierarchical model that allows for a compromise. In essence, the hierarchical model that we introduce below allows the data to determine the degree to which the binomial model with constant probability holds. The approach is closely related to the ideas of shrinkage estimation described for example in Lehmann and Casella (2003), and demonstrated in the baseball context by Efron and Morris (1997).

4.2 A Hierarchical Model for Batter-Pitcher Matchup Data

4.2.1 Data for a single player

To make the discussion concrete Table 4.1 provides data for a number of batter-pitcher matchups involving Kenny Lofton. During his 15 year career (through July 23, 2006), Lofton had faced 465 pitchers at least five times, 235 pitchers at least 10 times, and 85 pitchers at least 20 times. The outcomes from a small number of those matchups are provided in the table along with his aggregate record, 2283 hits in 7630 attempts for a batting proportion, known as the batting average, of .299. Data were collected from the website www.sports.yahoo.com. The table shows extremely high and low batting averages based on a variety of sample sizes. As noted previously in Chapter 3, the batting average may not be the most reliable measure of hitting performance. We consider it here because it is still widely reported and more importantly because we could easily obtain the data we needed

Pitcher	At-bats	Hits	Average
J.C. Romero	9	6	.667
S. Lewis	5	3	.600
B. Tomko	20	11	.550
T. Hoffman	6	3	.500
K. Tapani	45	22	.489
A. Cook	9	4	.444
J. Abbott	34	14	.412
A.J. Burnett	15	6	.400
K. Rogers	43	17	.395
A. Harang	6	2	.333
K. Appier	49	15	.306
R. Clemens	62	14	.226
C. Zambrano	9	2	.222
N. Ryan	10	2	.200
E. Hanson	41	7	.171
E. Milton	19	1	.056
M. Prior	7	0	.000
Total	7630	2283	.299

Table 4.1: Kenny Lofton's record against selected pitchers.

4.2.2 A probability model for batter-pitcher matchups

The probability model specification begins again with the simple binomial model for the outcomes in a single batter-pitcher matchup. Let n_i denote the number of batting attempts for Kenny Lofton against pitcher i and let y_i denote the number of hits (successful outcomes). Assuming the at-bats against a pitcher can be treated as independent, identically distributed trials leads us to assume that $y_i \sim \text{Binom}(n_i, p_i)$ with p_i

the probability of success for Lofton against the ith pitcher. One can question the i.i.d. assumption; we revisit that issue later.

To address the possibility that Lofton's ability varies across opposing pitchers, we model the p_i's as draws from a Beta(α, β) probability distribution. The beta distribution is commonly used as a probability model for parameters (like the batting average) that are known to lie in the interval from 0 to 1. The mean of the beta distribution is $\mu = \frac{\alpha}{\alpha + \beta}$ and the variance is $\nu = \left(\frac{\alpha}{\alpha + \beta}\right)\left(\frac{\beta}{\alpha + \beta}\right)\left(\frac{1}{\alpha + \beta + 1}\right)$. The variance is proportional to the usual Bernoulli variance $(\mu(1-\mu))$ with the additional term $\phi = \frac{1}{\alpha + \beta + 1}$ reflecting the degree to which the beta distribution is concentrated around the mean. Large values of α, β or small values of ϕ correspond to distributions that are highly concentrated around the mean. Figure 4.1 illustrates the beta distribution for three different choices of (α, β) having the same mean (0.3) but varying degrees of concentration around that mean.

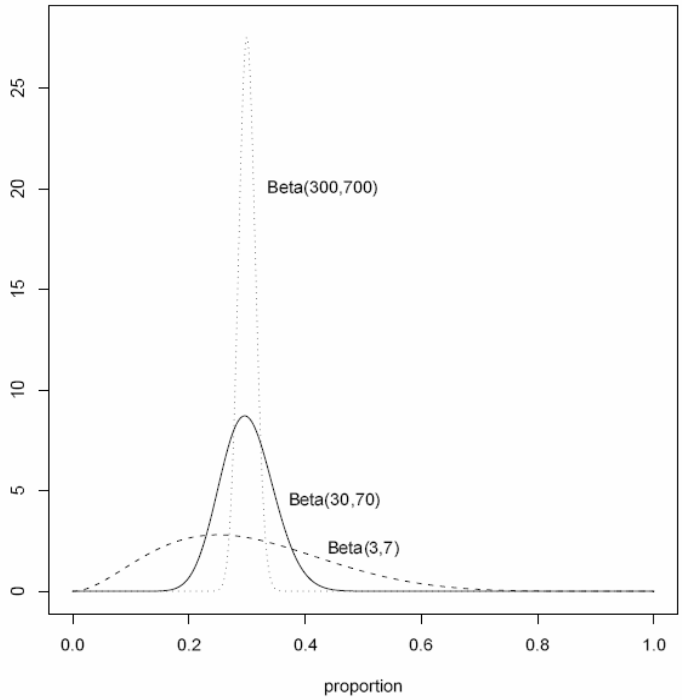

Figure 4.1: Three beta distributions with $\frac{\alpha}{\alpha+\beta} = .300$ but varying degrees of concentration around the mean

The family of beta distributions for the batter-pitcher matchup success probabilities provides a flexible family that can accommodate data that appears to be consistent with constant success probability ($\phi = 0$) or highly variable success probability (larger ϕ). This joint probability model (binomial distributions for the outcomes and a beta distribution for the binomial probabilities) is known as the beta-binomial hierarchical model. It has been applied elsewhere with sports data (Morrison and Kalwani, 1993). To complete the probability model we require a prior probability distribution on α, β. We select this prior distribution to reflect great uncertainty about the beta parameters so that the inference about the beta distribution will be determined to as large an extent as possible by the data. Following the discussion in Gelman, Carlin, Stern, and Rubin

(2003) we use a flat prior distribution on the parameters μ and $(\alpha+\beta)^{-1/2}$. Though this is not a proper probability distribution it does lead to a proper posterior distribution.

The Bayesian approach to inference requires that we calculate the posterior distribution of the parameters (α, β), and of the individual matchup parameters p_i if that is desired. The posterior distribution is

$$p(\alpha,\beta,\{p_i, i=1,\ldots,I\} \mid \{Y_i, i=1,\ldots,I\}) \propto p(\alpha,\beta)\prod_{i=1}^{I} p_i^{y_i}(1-p_i)^{n-y_i}\prod_{i=1}^{I}\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p_i^{\alpha-1}(1-p_i)^{\beta-1}\right)$$

with $p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$ for the vague prior distribution. The results provided here are obtained via posterior simulations from the posterior distribution. A simulation algorithm is used to simulate from the posterior distribution and 1000 random draws from the posterior distribution are used to estimate the posterior mean and posterior median, and to provide a posterior interval giving plausible values for the parameters.

4.2.3 Results – Kenny Lofton

In applying this model to the data for Lofton's batter-pitcher matchups (all 465 matchups, not just the ones in Table 4.1) we find that the estimated posterior median of $\mu = \dfrac{\alpha}{\alpha+\beta}$ is .299 and a 95% posterior interval for μ is (.290, .311). This is what we would expect since Lofton's career average of .299 based on more than 7600 at-bats. This speaks to Lofton's overall ability but then one can also obtain the posterior distribution for his estimated ability in particular matchups. Table 4.2) presents results for the matchups introduced in Table 4.1). The information from the initial table is repeated along with the posterior median for Lofton's average in matchups with a particular pitcher. The most important thing to notice is that Lofton's estimated ability is fairly constant. The highest estimated ability for Lofton is .348 against Kevin Tapani against whom Lofton has 22 hits in 45 attempts. The lowest estimated ability is .266 against the aforementioned Eric Milton against whom Lofton has 1 hits in 19 attempts.

Pitcher	At-bats	Hits	Observed Average	Estimated Average
J.C. Romero	9	6	.667	.326
S. Lewis	5	3	.600	.312
B. Tomko	20	11	.550	.337
T. Hoffman	6	3	.500	.309
K. Tapani	45	22	.489	.348
A. Cook	9	4	.444	.311
J. Abbott	34	14	.412	.323
A.J. Burnett	15	6	.400	.310
K. Rogers	43	17	.395	.326
A. Harang	6	2	.333	.301
K. Appier	49	15	.306	.299
R. Clemens	62	14	.226	.273
C. Zambrano	9	2	.222	.294
N. Ryan	10	2	.200	.295
E. Hanson	41	7	.171	.269
E. Milton	19	1	.056	.266
M. Prior	7	0	.000	.284
Total	7630	2283	.299	

Table 4.2: Kenny Lofton's estimated ability against selected pitchers.

4.2.4 Results - Derek Jeter

The same approach can be used in fitting matchup data for the Yankees popular shortstop Derek Jeter. Although a veteran, Jeter has less MLB experience than Lofton, facing only 382 different pitchers a total of 5 times or more, 217 pitchers 10 times or more, and 90 pitchers 20 times or more for a career total of 6530 at bats. In applying the beta-binomial model to the data for Jeter's batter-pitcher matchups we find that the estimated posterior median of $\mu = \frac{\alpha}{\alpha + \beta}$ is .318 and a 95% posterior interval for μ is (.310, .327). This is what we would expect since Jeter's career average is .316 based on his more than 6000 at-bats. Just like for Lofton, Table 4.3 presents results for a subset of Jeter's career matchup data but also includes a 95% posterior interval for Jeter's average in matchups with a particular pitcher. The most important thing to notice is that Jeter's estimated ability is nearly constant. The highest estimated ability for Jeter is .326 against Hideo Nomo against whom Jeter has 12 hits in 20 attempts. The lowest estimated ability is .311 (just .005 below Jeter's career average of .316) against Brad Radke against whom Jeter has 8 hits in 41 attempts.

A graphical display of the results in Table 4.3 is provided in Figure 4.2. The plus signs indicate the posterior median and the dashed lines indicate the posterior intervals for the players. These are plotted against the observed average.

Pitcher	At-bats	Hits	Observed Average	Estimated Average	Posterior 95% Interval
R. Mendoza	6	5	.833	.322	(.282, .394)
H. Nomo	20	12	.600	.326	(.289, .407)
A.J. Burnett	5	3	.600	.320	(.275, .381)
E. Milton	28	14	.500	.324	(.292, .397)
D. Cone	8	4	.500	.320	(.278, .381)
R. Lopez	45	21	.467	.326	(.292, .401)
K. Escobar	39	16	.410	.322	(.281, .386)
J. Wetteland	5	2	.400	.318	(.275, .375)
T. Wakefield	81	26	.321	.318	(.279, .364)
P. Martinez	83	21	.253	.312	(.254, .347)
K. Benson	8	2	.250	.317	(.264, .368)
T. Hudson	24	6	.250	.315	(.260, .362)
J. Smoltz	5	1	.200	.317	(.263, .366)
F. Garcia	25	5	.200	.314	(.253, .355)
B. Radke	41	8	.195	.311	(.247, .347)
J. Julio	13	0	.000	.312	(.243, .350)
D. Kolb	5	0	.000	.316	(.258, .363)
Total	6530	2061	.316		

Table 4.3: Derek Jeter's estimated ability against selected pitchers.

The graph shows clearly that the observed heterogeneity in matchup averages is discounted heavily in the estimated abilities. The estimates are often said to shrink away from the observed success probabilities in a given matchup towards the overall success probability for Jeter.

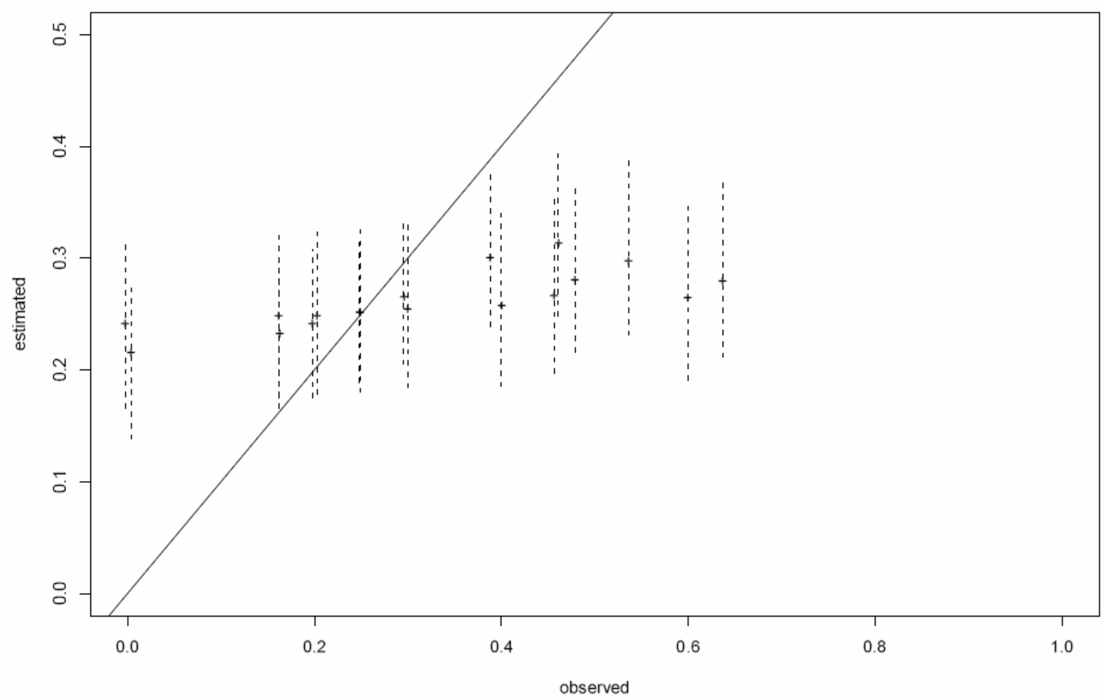

Figure 4.2: Plot showing the estimated batting ability of Derek Jeter against 18 different opposing pitchers from Table 4.2. Horizontal axis shows observed success probability. Plus signs indicated median estimate of ability and dashed lines indicate 95% posterior intervals for the ability. The 45 degree is provided for reference.

The amount of shrinkage depends on two factors. First, it depends on how much heterogeneity is evident in the data and second on the amount of data. The estimates for all of the high and low average matchups are close to Jeter's lifetime average. This indicates that there is not much evidence for heterogeneous performance. The effect of sample size can be seen in that the relative order of the estimated abilities does not match the order of the observed averages. The first matchup in the table which corresponds to a batting average of .833 is discounted heavily because it depends on only 6 attempts. Another feature of the graph that is striking is that the posterior intervals are extremely wide. Even with 83 attempts (the most for any matchup involving Jeter) there remains great uncertainty about the true ability of Jeter versus pitcher Pedro Martinez. (It should be noted that the posterior interval is narrower than it would be in a simple binomial analysis because the posterior interval combines information from the 83 at-bats with information from the beta distribution of Jeter's averages against other pitchers).

4.2.5 Results - multiple players

It is natural to wonder whether the pattern that we find above pertains to other hitters as well. We repeated the analysis for 230 hitters. These are batters with at least 350 plate appearances (indicating they were significant players) in 2006. The career data are analyzed for each player; the sample includes both young players and experienced players. A key parameter is $\phi = \dfrac{1}{\alpha + \beta + 1}$ which indicates the estimated heterogeneity in a batter's ability across the population of pitchers. Figure 4.1 provides one way to calibrate the values of ϕ. The Beta(3,7) distribution has ϕ = .091; the Beta(30,70) distribution has $\phi \approx .01$; the Beta(300,700) distribution has ϕ = .001. The smaller the value the more concentrated the distribution and the less variable a batter's ability is across pitchers. For Derek Jeter the posterior median of ϕ is .002. (Note this value too is an estimate with uncertainty attached but we are confident that it is less than .01.) This value is quite small and reflects the fact that Jeter's ability varies a bit but not very much across the population of pitchers.

Over 230 major league batters we found the median value to be about .005. Thus Jeter is more consistent than the typical players in the sense that his ability does not appear to vary much across pitchers. The values in this group of batters ranged from .0006 to .11. High values generally correspond to players with less information (young players) for whom performance to date suggests great variability. There is typically a great deal of uncertainty about the value of ϕ for such players.

Though ϕ is one useful way to characterize performance, it is not immediately obvious how ϕ translates into batting averages. Recall that Jeter (ϕ = .002) had estimated abilities that ranged from about .311 to .326 across the 382 pitchers he had faced at least 5 times. To provide some context we note that Kenny Lofton has estimated value of ϕ equal to 0.008, indicating that Lofton may be a more variable batter with respect to different pitchers. Though small, Lofton's ϕ value is considerably larger than Jeter's value. It is reflected in Lofton's estimated ability across pitchers. Lofton's estimated ability varies from .265 (against Eric Milton - the pitcher in the introduction) to .340. That is much more variability than we see in Jeter's performance.

4.3 Batter-Pitcher Data from the Pitcher's Perspective

Most often one hears about batter-pitcher matchup data from the batter's perspective. As in the introduction to this chapter a player may sit out a game when the opposing pitcher is a bad matchup or alternatively may be removed from a game at a key juncture if there is a bad matchup. On the other hand most pitchers are put in the game to face more than one batter and therefore managers are unlikely to make a change because of one matchup. However, in the later stages of a game when the starting pitcher begins to fatigue, a manager usually chooses to go to his bullpen of relief pitchers to finish the game. If a true matchup effect does exist (as assumed by Tony LaRussa) than it would be wise to bring in the reliever with the best matchup numbers against the batter's he would be facing. There is nothing in the current model that is particular to the batters so the exact same model can be used to analyze batter-pitcher matchups from the pitchers' perspective. Is there any evidence that pitchers do better against some hitters than others?

4.3.1 Results – Tim Wakefield

To demonstrate the analysis for a pitcher, consider the record of Derek Jeter's longtime Boston Red Sox nemesis Tim Wakefield. Wakefield has been pitching for 14 years and through July 23, 2006 had faced 576 batters 5 or more times, 312 batters 10 or more times, and 137 batters 20 or more times for a total of 8,363 batters faced. The first three columns of Table 4.4 provide Wakefield's results against a selection of hitters along with his career total. As with Jeter we focus only on batting average here though the analysis could be repeated for other measures of pitching effectiveness. Wakefield's career batting average against (BAA) is very good, although his on-base percentage against is considerably higher due to his propensity to walk batters. He allows batters a success proportion of .255 which is below the average for the entire league. There is considerable variability in the outcomes that we see ranging from Aaron Rowand who has 11 hits in 17 attempts to Edgar Martinez who has only 1 hit in 19 attempts against Wakefield. However, the hierarchical beta-binomial model suggests a more moderate variability about Wakefield's performance across the pool of batters.

Pitcher	At-bats	Hits	Observed Average	Estimated Average	Posterior 95% Interval
A. Rowand	17	11	.647	.278	(.227, .373)
J. Pierre	5	3	.600	.259	(.206, .331)
C. Garcia	13	5	.385	.259	(.201, .332)
M. Scutaro	8	3	.375	.256	(.197, .329)
O. Vizquel	53	19	.358	.271	(.227, .333)
M. McGwire	24	5	.208	.250	(.189, .315)
R. Belliard	5	1	.200	.253	(.190, .319
J. Giambi	79	14	.177	.236	(.172, .284)
E. Martinez	19	1	.056	.241	(.170, .297)
TOTAL	8363	2130	.255		

Table 4.4: Tim Wakefield's record and estimated ability against selected hitters.

The right hand side of Table 4.4 summarizes the posterior distribution of the probabilities of getting a hit for each player in the table when facing Tim Wakefield. The results show similarities and differences relative to the analysis of the Jeter data. Once again the estimated abilities are concentrated much more closely around Wakefield's career average of .255 than the observed averages. The amount by which an individual estimate changes relative to the observed average depends on the sample size. Thus Jason Giambi's lack of success in 79 attempts is reflected in a much lower estimated success probability (.177) than the estimate for Belliard (.253) who has one hit in only 5 attempts thus far. There is also considerable uncertainty about the true ability of each batter based on the relatively small sample sizes.

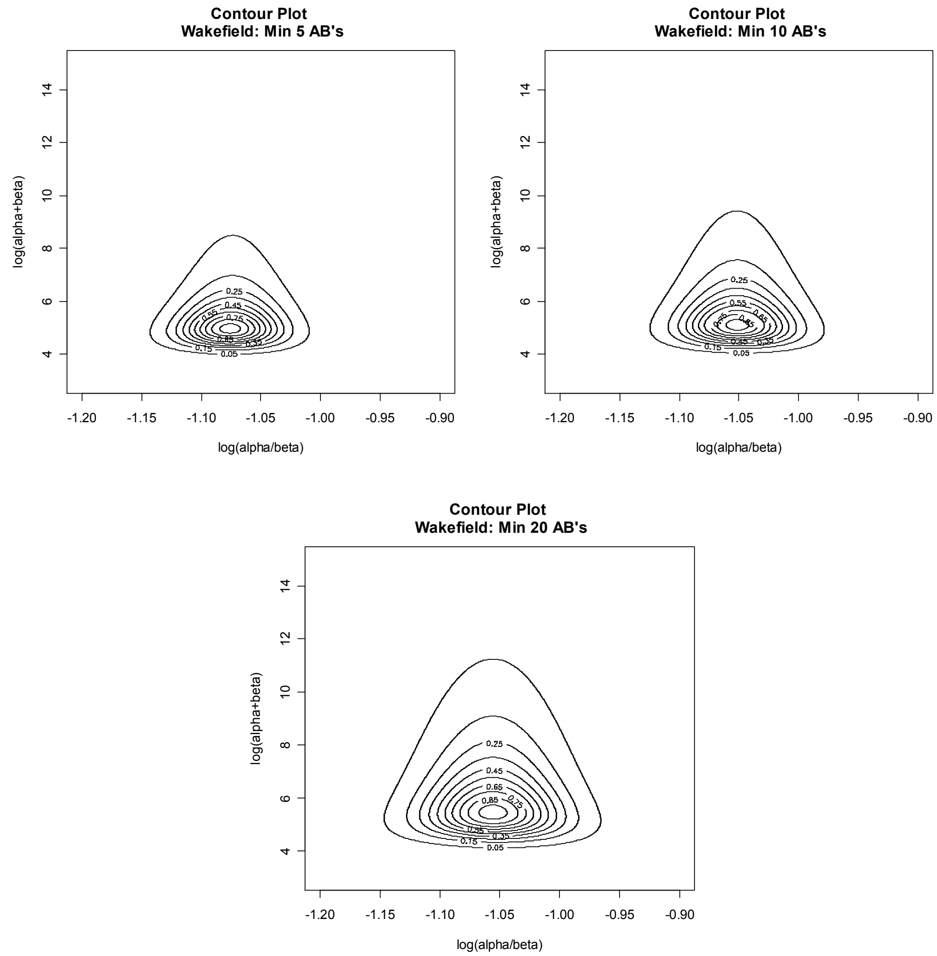

Figure 4.3: Contour plots of the marginal posterior density of ($\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)$) for the Wakefield data. Contour lines are at 0.05, 0.15, …, 0.95 times the density at the mode.

Although moderate, the reduced variability in the posterior distribution that results from an increase in batter-pitcher matchup at bats can be seen through a careful graphical inspection of Figure 4.3 which provides contour plots when the number of matchups are equal to or greater than 5, 10, and 20. Corresponding changes in ϕ also result as the parameter decreases in value (.0049, .0041, .0023) as matchup at bats increase from 5, to

10, to 20. As explained earlier, this is to be expected as more at bats provide more information which in turn decreases variability.

4.3.2 Results - Mike Mussina

To demonstrate the analysis for another pitcher, consider the record of Derek Jeter's teammate Mike Mussina. Mussina has been pitching for 16 years and through July 23, 2006 had faced 576 batters 5 or more times, 390 batters 10 or more times, and 202 batters 20 or more times. The first three columns of Table 4.5 provide Mussina's results against a selection of hitters along with his career total. Similar to Wakefield, Mussina's career record is excellent allowing batters’ hits only one quarter of the time. There is again considerable variability in the outcomes observed ranging from Mueller who has 0 hits in 23 attempts to Hidalgo who has 7 hits in 11 attempts.

Pitcher	At-bats	Hits	Observed Average	Estimated Average	Posterior 95% Interval
R. Hidalgo	11	7	.636	.278	(.211, .368)
A. Cintron	5	3	.600	.263	(.191, .348)
B. Roberts	26	14	.538	.296	(.231, .389)
R. Ibanez	21	10	.476	.279	(.216, .367)
F. Catalanotto	56	26	.464	.312	(.248, .397)
R. White	11	5	.455	.265	(.197, .352)
M. Huff	5	2	.400	.256	(.185, .340)
F. Thomas	78	30	.385	.299	(.239, .374)
P. Burrell	10	3	.300	.253	(.184, .335)
J. Canseco	61	18	.295	.264	(.205, .332)
B.J. Surhoff	40	10	.250	.250	(.189, .319)
A. Soriano	8	2	.250	.250	(.180, .331)
H. Baines	35	7	.200	.240	(.175, .308)
T. Hafner	10	2	.200	.247	(.178, .324)
C. Fielder	42	7	.167	.231	(.166, .296)
S. Posednick	6	1	.167	.247	(.175, .325)
B. Mueller	23	0	.000	.214	(.138, .279)
J. Kent	6	0	.000	.240	(.166, .316)
TOTAL	11954	2992	.250		

Table 4.5: Mike Mussina's record and estimated ability against selected hitters.

As was the case with Wakefield's data, the results from Table 4.5 show similarities and differences relative to the analysis of both the Lofton and Jeter data. Just like the past analyses, the estimated abilities are concentrated much more closely around Mussina's career average .250 than the observed averages. And again, the amount by which an individual estimate changes relative to the observed average depends on the sample size. There is again a sizeable uncertainty about the true ability of each batter based on the relatively small sample sizes of most matchups compared to the overall career totals. Notice also that the posterior intervals are quite wide.

The key difference relative to the data from Jeter is that there is more variation in the estimated abilities. They range from .214 (Mueller) to .312 (Catalanotto) which is a much bigger range than was evident for Jeter or other hitters. Part of the explanation lies in the selected individuals. Jeter is among the more consistent hitters - there is little variability in his success probability across pitchers (recall that ϕ which measures this consistency was .002 for Jeter). Mussina's performance is above the median in terms of heterogeneity of performance. The estimate of ϕ for Mussina is .008 while the estimate of ϕ for Wakefield is .005 which means that Mussina's estimated abilities will be more spread out than Wakefield's. Interestingly, the estimated abilities for Mussina are more spread out than those for Kenny Lofton, a batter with similar estimate for ϕ.

4.3.3 Results - multiple players

The analysis performed on the Wakefield and Mussina data was replicated on 158 other pitchers. These are all pitchers who pitched more than 85 innings during 2006. This includes both 2006 rookies and more experienced pitchers (like Mussina and Wakefield). The distribution of ϕ (our measure of consistency or concentration) is quite similar to what was found in the population of hitters. The values range from a low of .0007 (which suggests almost no variability across batters) to .03 (which suggests considerable variability). Successful pitchers are found at both ends of the distribution. As remarked above, Mussina is slightly above the median for pitchers (which is .005), while Wakefield is just about the norm in terms of pitcher matchup variability.

One feature that is currently unexplained in our data is that it appears the distribution of individual batter-pitcher matchup ability estimates tends to be more spread out for pitchers than for hitters, even given the same value of ϕ. This is a bit paradoxical but may reflect the distribution of sample sizes.

4.4 Towards a More Realistic Model

The analyses done here provide one natural check of the methodology. Derek Jeter had batted against Mike Mussina 33 times before they became teammates. Jeter has 12 hits for an average of .363. This is above Jeter's typical average (.316) and substantially higher than what Mussina usually allows (.250). Based solely on the analysis of Jeter's record one finds that an estimate of Jeter's ability against Mussina is .319 with a 95% posterior interval of (.279, .372). Based on the analysis of Mussina' record one finds that an estimate of Jeter's ability against Mussina is .271 with a 95% posterior interval of (.210, .353). The point estimates differ considerably but the posterior intervals are sufficiently wide that differences of this magnitude are not terribly unusual. Though focusing on an individual hitter or pitcher seemed like a natural first step in addressing the question of batter-pitcher matchups it is disheartening though to have two estimates that differ by so much. It is natural to wonder whether a unified modeling approach can provide better inferences in this setting. Work in this direction is ongoing but some preliminary ideas were already presented in Chapter 3 through the Log 5 and Weighted Log 5 approaches. Of course the question about whether matchups are important is really the question about whether an interaction term is needed in either of the aforementioned models. One positive feature of a more sophisticated modeling approach is that it naturally extends to allow consideration of other factors known to effect batting success like the ballpark and the dominant-handedness of the players. Such factors have been ignored in our study.

4.5 Discussion

The main result from this batter-pitcher matchup study is not particularly surprising to statisticians. The evidence is that baseball fans and baseball professionals may be placing too much an emphasis on the chance outcomes found in small batter-pitcher matchup samples. A comprehensive analysis suggests that there is in fact much less variation in batting performance across different pitchers than would be suggested by looking at the results of small samples. For Derek Jeter our results suggest that his performance against a pitcher rarely alters his probability of success by more than about .010 from his lifetime average of .316. Because the information contained in a player’s career numbers are much more accurate than those found from matchup data, when a batter has had success (or lack of success) against a particular pitcher in 5, 10 or even 20 or 50 at-bats we would be wise not to put too much stock in this data. At the same time, however, there is also evidence that some batters are more heterogeneous in their batting than others. Jeter is among the most consistent. For Kenny Lofton, the batter whose experience starts the Chapter, there is in fact evidence of more variability. In particular, Lofton's estimated success rate against Milton (the starting pitcher when Lofton was rested for a game) is estimated as being .033 lower than his lifetime average .299.

Though still a modest difference this might be large enough to justify trying a different hitter.

Before closing though it is important to recognize that there are significant issues, both statistical and non-statistical, that limit the inferences we should draw from the results presented here. On the statistical side we have worked primarily with data from a small number of batters and pitchers. Though summary results are provided for a range of other hitters and pitchers these have not been explored in as much detail. Also, as discussed in Chapter 3, a player's OBP (OBPA for pitchers) is shown to be a better measure of ability than the traditional batting average. Analyzing data based on this statistic may provide stronger evidence for a matchup effect. Perhaps more importantly, we would be wise to recall the words of the statistician George Box who said, "All models are wrong, some models are useful." The simple hierarchical model that we have used here is clearly not quite right. It ignores some of the factors, site of game and dominant hand information are two, which are known to impact the probability of a hit. When asking whether there are "matchup effects" we likely want to know whether there is anything beyond the known determinants of batting success that is affecting a particular matchup. The current analysis does not really address this question because it ignores some factors that we know are relevant. However, taking account of such factors is likely to make matchup effects appear even less prevalent than shown already.

Another caveat before criticizing the work of baseball decision makers is that non-statistical information is likely to impact a decision as well. In the situation that started the article, we are not privy to all of the information that Grady Little had nearby as he built his lineup for the August 29 game against Cincinnati. Perhaps Kenny Lofton was slightly injured or perhaps there was a player returning from the injured list whom he wanted to evaluate. The challenge of the manager is to balance quantitative information about the ability and expected performance of a player with factors like these that are less statistical.

Despite the statistical and non-statistical caveats, the results here argue strongly against drawing conclusions based on the limited information available about the matchup between a particular batter and a particular pitcher. A manager should avoid reacting to small sample variation which may lead him to rest a superior player at an inopportune time.

Chapter 5

Methods for Predicting Individual Performance

This chapter will discuss the topic of predicting future performance by an individual player. We begin with the assumption that past performance can be a reliable predictor for future performance, an assumption that has been validated quite reliably within baseball (Schall, 2000). Consideration will also be given to: the effect of age on performance, the plausibility of assuming all players age in a similar manner, and the value of early season statistics in predicting performance over the remainder of a season. These topics are addressed through the development of a multilevel model with age used as the grouping variable, a more individual and flexible quadratic fit model, and an age independent time weighted model. The chapter will conclude with a quick look at the ability of each model in predicting player performance for the 2007 season.

Factors that possibly affect a baseball player's offensive ability (where offensive ability will be defined as a player’s *OBP*) are experience, age, talent, injuries, and coaching. Experience should improve *OBP* numbers, but past some age, a player's physical coordination should decline, tending to reduce offensive ability. Talent is the inherent ability of a player, which of course is unobservable. The extent of injuries and the level of coaching received are difficult to measure. The effects of experience and age can be combined to form a relationship between the *OBP* and age of an individual player, as shown in Figure 1. As a young player gets older, the added experience will improve his *OBP* until a certain point when physical capabilities begin to decline and this aging process now outweighs the effects of additional experience. This occurs at the apex of the curve in Figure 1. A player's peak *OBP* level, represented by the distance from the apex to the x-axis on the figure, is affected by the player's talent. This model of a player's lifetime batting profile can be represented by

$$OBP_i(t) = \alpha_0 + \alpha_1 AGE_i(t) + \alpha_2[AGE_i(t)]^2 + U_i(t)$$

$$i = 1,2,\ldots, N;\ t = 1,2,\ldots, T$$

where $OBP_i(t)$ is the on-base percentage of player i in year t; $AGE_i(t)$ is the age of player i in year t, and $U_i(t)$ is the random disturbance with an unconditional mean of zero.

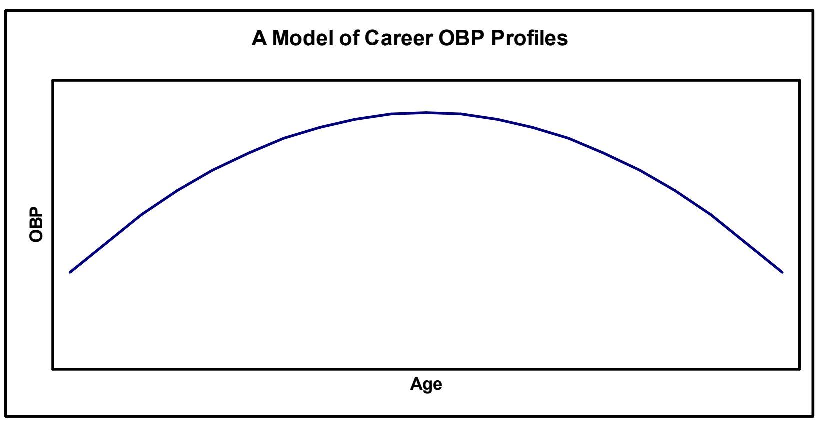

Figure 5.1: A player's peak age (apex) can be defined as the point where the effect of additional experience is exactly offset by a loss of physical ability. The perpendicular distance of the curve to the x-axis can be interpreted as a function of the individual player's current offensive ability.

5.1 The Aging Process: A Multilevel Modeling Approach

One of the main purposes of this section is to address the question of how much the performance variation among players in general is due to variation within an age year and how much is accounted for by variation among ages. A player's ability is thought to generally start at a relatively low level, increase until a particular peak age, and then deteriorate gradually until retirement. The question concerns whether ages "homogenize" players over their careers with respect to the performance measure in question.

The problem can be restated in terms of whether players of the same age are more alike when compared to players across ages, and whether there is some type of age effect that accounts for that resemblance. In the context of statistical modeling, should appreciably large portions of variance lie at the level of individual differences within age, then interest would focus on player characteristics that might account for that variance. Similarly, should appreciably large portions of performance variance remain among ages, then interest would focus on explanatory variables at this level.

This section will address the decomposition of variance in player performance via the method of multilevel modeling. One can think of multilevel models as ordinary regression models that have additional variance terms for handling group membership effects. The key to understanding MLM in this context is understanding how group membership (age) can lead to additional sources of variance in ones model.

The first variance term that distinguishes a multilevel model from a regression model is a term that allows groups to differ in their mean values (intercepts) on the dependent variable. When this variance term, τ_{00}, is non-zero, it suggests that groups differ on the dependent variable. One predicts group-mean differences with group-level variables. These are variables that differ across groups, but do not differ within-groups. Group-level variables are often called "level-2" variables.

The second variance term that distinguishes a multilevel model from a typical regression model is the term that allows slopes between independent and dependent variables to differ across groups, τ_{11}. Single-level regression models generally assume that the relationship between the independent and dependent variable is constant across groups. If slopes randomly vary, one can attempt to explain this slope variation as a function of group differences within multilevel modeling.

A third variance term is common to both multilevel models and regression models. This variance term, σ^2, reflects the degree to which an individual score differs from its predicted value within a specific group. One can think of σ^2 as an estimate of within-group variance.

5.1.1 Steps in multilevel modeling: Variance Decomposition

Because multilevel modeling involves predicting variance at different levels, one typically begins a multilevel analysis by determining the levels at which significant variation exists. In the case of the two-level model, one generally assumes that there is significant variation in σ^2 – that is, one assumes that within-group variation is present. One does not necessarily assume, however, that there will be significant intercept variation, τ_{00}, or between-group slope variation, τ_{11}. Therefore, one typically begins by examining intercept variability. If τ_{00} does not differ by more than chance levels, there may be little reason to use random coefficient modeling since simple OLS modeling will suffice. Note that if slopes randomly vary even if intercepts do not, there may still be reason to estimate random coefficient models.

In step 1 of a multilevel analysis, one explores the group-level properties of the outcome variable to determine three things: first, how much of the variance in the outcome can be explained by group membership. Second, whether the group means of the outcome variable are reliable. By convention, one would like the group mean reliability to be around .70 because this indicates that groups can be reliably differentiated (Bliese, 2000). Third, one wants to know whether the variance of the intercept τ_{00} is significantly larger than zero.

The initial model that is used to decompose the individual-level variance into "within-age" and "among-ages" components is equivalent to an unconditional model equivalent to the one-way random effects analysis-of-variance (ANOVA) model. The "Level-1" model describes the performance for player *i* who is nested in age *j* and can be written as

$$Y_{ij} = \beta_{0j} + r_{ij} \tag{5.1}$$

where, for example, Y_{ij} could be Derek Jeter's current *OBP* at his current age of 33, β_{0j} can be interpreted as the unadjusted true OBP mean for all players 33 years old, and r_{ij} is the level-1 random error term reflecting variation among Jeter and all other players 33 years old. The error term, r_{ij} is assumed to be normally distributed and homogenous within ages, $r_{ij} \sim N(0, \sigma^2)$. The component that accounts for the age effect (the so-called "Level-2" model) can be written as

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \tag{5.2}$$

where γ_{00} is the average performance across all of the ages, and μ_{0j} is the level-2 random component reflecting variation in average *OBP* around the grand mean of all ages. It is also assumed $\mu_{0j} \sim N(0, \tau^2)$ with τ^2 being the variance of the true means β_{0j} around the grand mean γ_{00}. (Raudenbush and Bryk, 2002).

Inserting equation (5.1) into equation (5.2) yields the one-way random effects ANOVA model

$$Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij} \tag{5.3}$$

so that the performance of player *i* of age *j* is a function of a grand mean performance across all ages (γ_{00}), a component related to the effect of being a particular age (μ_{0j}) and a component related to that player's individual ability (r_{ij}).

5.1.2 The Data

The data for this analysis consist of both batters and pitchers. The top 387 players in at-bats from the 2006 season were used which resulted in 2800 player-year observations. Meaning, for each player, every year of his career is treated as a separate observation with current *OBP*, career average *OBP* to date, and age serving as the variables that were recorded. For pitchers, data from the 425 most used pitchers from 2006 were recorded for a total of 2239 player-years. For both sets of players, current

OBP (*OBPA* for pitchers) will serve as the outcome variable, with the previous season *OBP* (*OBPA*) and age as potential explanatory variables at the individual level.

5.1.3 The Intra-Class Correlation

The statistic that we will use to obtain the variance in the individual performance accounted for by differences among ages is the *intraclass correlation*, written as

$$\rho_{IC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$$

where $\sigma^2 = var(r_{ij})$ represents the component of variation among players and $\tau_{00} = var(\mu_{0j})$ represents the component of variation accounted for by ages. Offensive players had a $\rho_{IC} = 0.05741239$ and pitchers had a $\rho_{IC} = 0.02652585$. These results indicate that there is little variation at the age level for both groups of players, but a large amount of variation at the individual level.

5.1.4 Group-Mean Reliability

When exploring the properties of the outcome variable, it can also be of interest to examine the reliability of the group mean. The reliability of group means often affects one's ability to detect emergent phenomena. In other words, a prerequisite for detecting changes at the aggregate level is to have reliable group means. By convention, one strives to have group mean reliability estimates around .70. Group mean reliability estimates are a function of the Intra-Class Correlation and group size (Bliese, 2000).

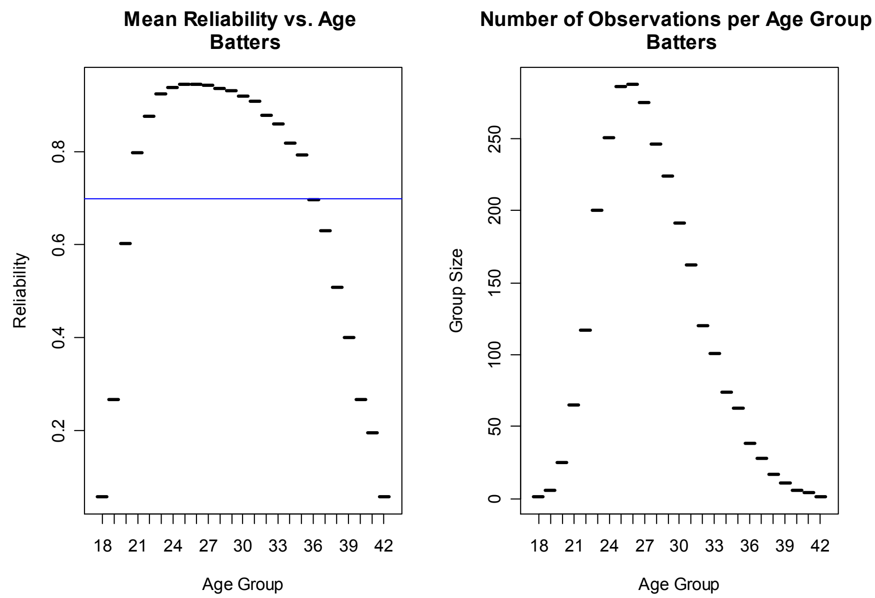

Figure 5.2: Group Mean Reliability and Group Size Measures for Batter data.

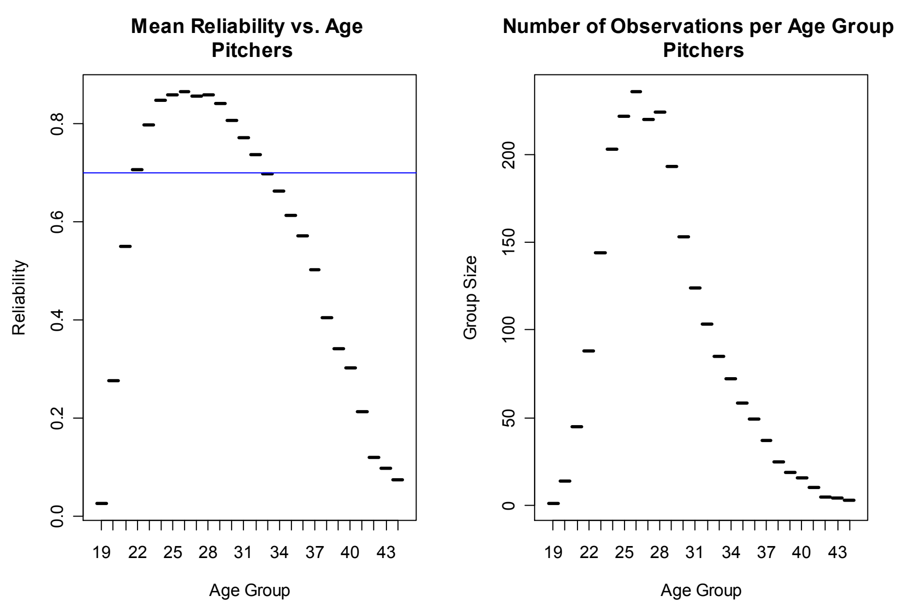

Figure 5.3: Group Mean Reliability and Group Size Measures for Pitcher data.

As can be seen from the above histograms, the group-mean reliability scores for both sets of players vary primarily according to sample size. Both batters and pitchers have acceptable scores beginning in the early 20's age range but their reliability weakens in the early 30's for pitchers and the mid 30's for batters. Although certain levels of age lack significant power to detect change at the aggregate level, the overall weighted group-mean reliability scores are 0.8996683 and 0.775374 for batters and pitchers respectively. These results are above the acceptable level of .70 and indicate that overall trends in the outcome variable can be viewed with a fair amount of confidence.

5.1.5 Determining whether τ_{00} is significant

Before continuing further in this analysis, it would be interesting to know whether the intercept variance (τ_{00}) estimates of .0000464923 and 0.0001727730 for pitchers and batters respectively are significantly different from zero. To do this, we can compare the -2 log likelihood values between a model with a random intercept, and a model without a random intercept. For both groups, the results suggest ($p < .0001$) that there is significant intercept variation.

In summary, it can be concluded that there is significant variation in terms of *OBP* across ages for both pitchers and batters. We also estimate that 5.74% and 2.65% of the variation in batters and pitchers *OBP* (*OBPA*) is a function of a player's age. Thus, a model that allows for random variation in *OBP* among ages is better than a model that does not allow for this random variation.

5.1.6 **Adding Predictors**

The next logical step in the analysis would be to add predictor variables into the equation. Since it is assumed that past performance is the best predictor of future performance, each batter's previous year (pitchers) *OBP (OBPA*) will be used as a predictor of current *OBP (OBPA).* The form of such a model is

$$\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}(OBP_{ij}) + r_{ij} \\
\beta_{0j} &= \gamma_{00} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10}
\end{aligned}$$

The first row states that current *OBP* is a function of each age group's intercept plus a component that reflects the linear effect of previous year *OBP* plus some random error. The second line states that each groups' intercept is a function of some common intercept and random between-group error. The third line states that the slope between current individual *OBP* and previous year *OBP* is fixed – it is not allowed to randomly vary across groups. Stated another way, we assume that the relationship between previous year *OBP* and current *OBP* is identical in each group.

When we combine the three rows into a single equation, we get an equation that looks like a common regression equation with an extra error term (μ_{0j}). This error term indicates that the estimated current OBP intercepts (i.e., means) can randomly differ across groups. The combined model is

$$Y_{ij} = \gamma_{00} + \gamma_{10}(OBP_{ij}) + \mu_{0j} + r_{ij}$$

This constant slope model with previous year *OBP* performs poorly with respect to its ability to explain away added within-group variance σ^2. This should come as no surprise as the relationship between current *OBP* and previous *OBP* changes over the course of a player's career. Figure 5.4 illustrates this progression. The color changes represent the sample size for each age with lighter colors indicating more observations. For batters, a typical career appears to increase in performance until a peak year around age 27 or 28 and then decrease steadily. For pitchers (where lower *OBPA* indicates better performance), players seem to peak at a slightly lower age of 26 or 27 and then decrease in performance, but with a much greater variability than the average batter.

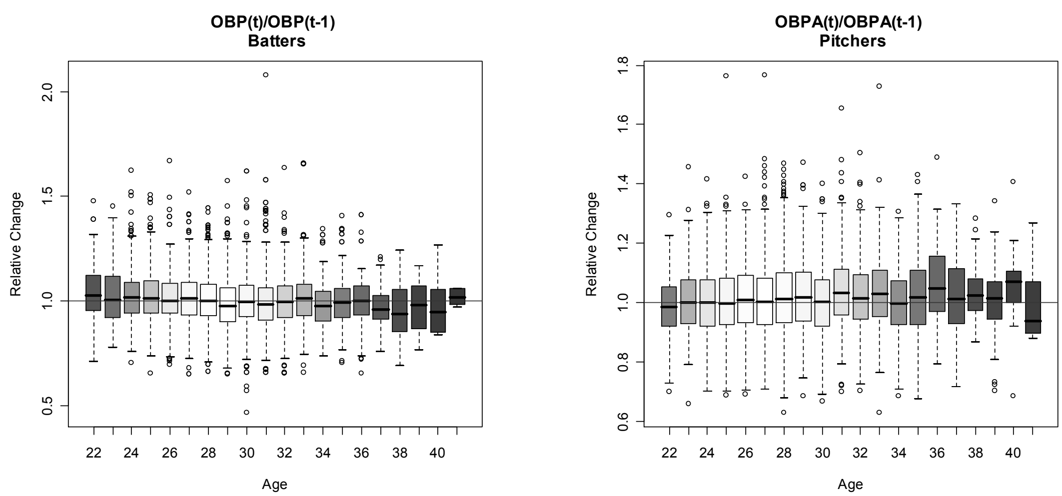

Figure 5.4: Player Performance Progression as a function of age

To account for this variability among age groups with respect to predicting current *OBP*, another term can be added to the model which allows for variation in slopes.

$$
\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}(OBP_{ij}) + r_{ij} \\
\beta_{0j} &= \gamma_{00} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \mu_{1j}
\end{aligned}
$$

The last line of the model includes the error term μ_{1j} and indicates that the relationship between current *OBP* and previous year *OBP* is permitted to randomly vary across age. The variance term associated with μ_{1j} is τ_{11} and the significance of τ_{11} is verified after using a log likelihood ratio test between the two models (constant vs. variable slope). As shown in Table 5.1, a greater portion of the variability among batters and pitchers is now explained by age.

	Batters	Pitchers
Random Intercept	$\rho_{IC} = 0.0574$	$\rho_{IC} = 0.0265$
Random Intercept + Constant Slope	$\rho_{IC} = 0.0597$	$\rho_{IC} = 0.0324$
Random Intercept + Random Slope	$\rho_{IC} = 0.1204$	$\rho_{IC} = 0.05858$

Table 5.1: Model Performance Summaries for Batters and Pitchers

It appears, however, that age does not have the same effect on batters as it does for pitchers. These numbers suggest that age plays a greater role in determining the performance of a batter than of a pitcher. This claim can also be seen from the boxplots where both the IQR and range of values across age for pitchers are more varied than for batters.

The idea that age affects player ability is nothing new. Because of various physical processes in the human body, an athlete's body cannot perform at a peak level for an extended amount of time. The multilevel model used in this analysis confirms the claim that ability varies as a function of age. A drawback to this analysis, however, is it assumes all players within a particular age group, age at the same rate. Because the validity of such an assumption seems questionable, the search for a more flexible, player based approach to the performance versus age relationship should also be explored.

5.2 The Aging Process: An Individual Quadratic Fit Approach

Because factors like experience, nagging injuries, physical demand from playing a certain position, and natural variability among players are certain to exist and affect everyone differently, aging functions unique to each individual player may produce more accurate estimates of future performance. As an example, the career performance trajectories (as measured by *OBP* and *OBPA* for batters and pitchers) of a few of Major League Baseball's longest tenured current batters and pitchers are given in Figure 5.5. For a batter, the rise and fall of a typical career is expected to look a lot that of Craig Biggio's; a convex quadratic with a maximum in the late 20's and sloping gradually downward in either direction from its peak. After fitting quadratic curves for a number of baseball's longest careers though, one will observe batters with widely varying trajectories. A few of baseball's most notable modern day hitters are given as an example.

The career of Barry Bonds is perhaps most striking as his performance seems to constantly be improving way after the normal peak age of 27-28. Ken Griffey Jr.'s career does not show much variability of performance over a relatively large age range, and similar to Barry Bonds, Frank Thomas' performance appears to get more variable with age.

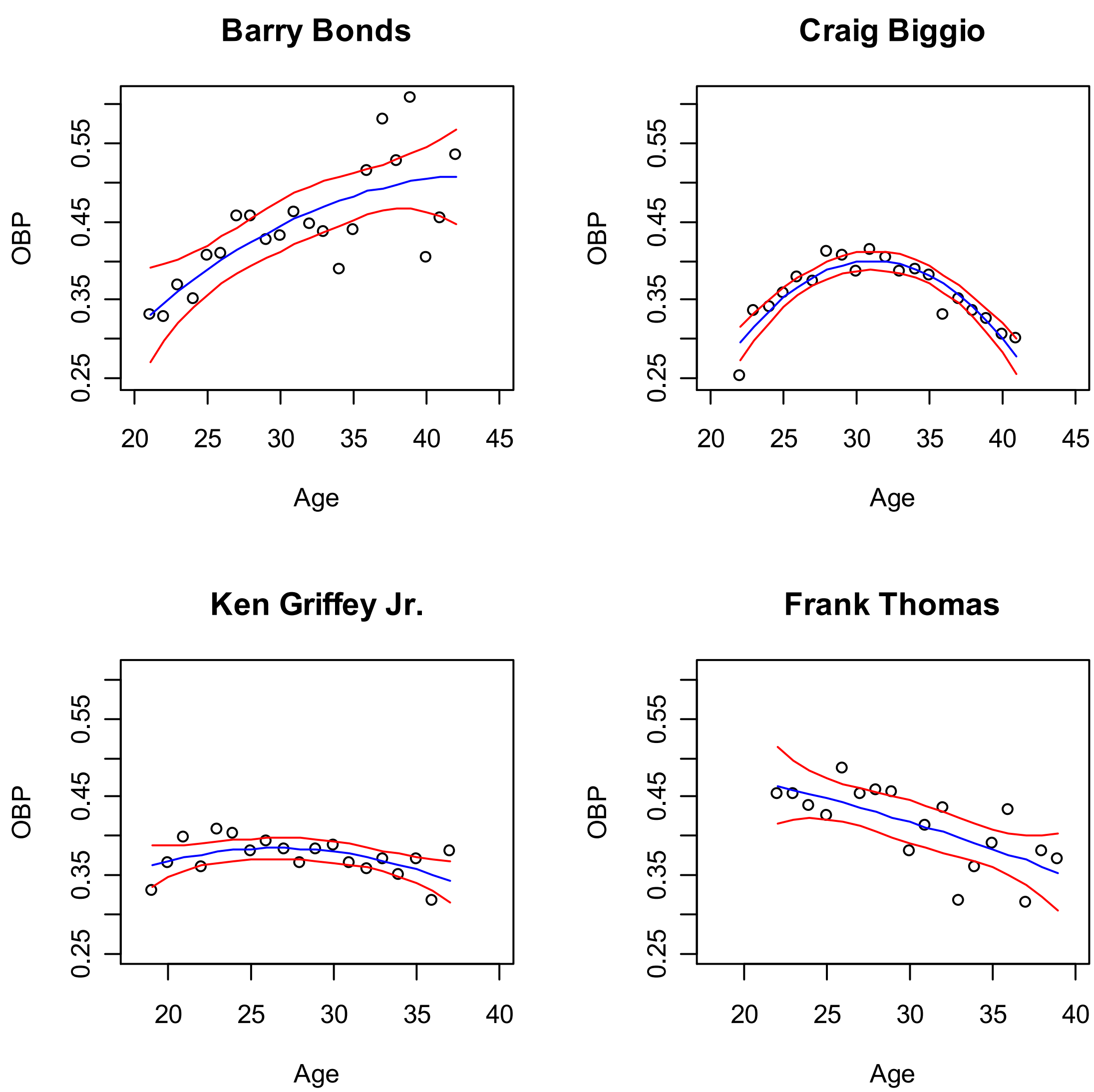

Figure 5.5: Quadratic fit with 95% confidence interval career trajectories for a few of MLB's longest tenured offensive players

A similar story can be told for pitchers. Although the multilevel model in section (5.1.6) and from observations of the careers of various pitchers suggest that pitchers do appear more variable in performance than batters, one would still suspect a relatively concave form to fit the trajectories of pitching careers well. The career of Greg Maddux appears to fit this form as his performance steadily improves into his late 20's and early 30's and then slowly declines after this mark. Randy Johnson's curve is also fairly concave but with a later career peak compared to most pitchers and more variable year to year than Greg Maddux. Roger Clemens career is actually convex in form and Tim Wakefield's trajectory is rather linear in nature and improving with age.

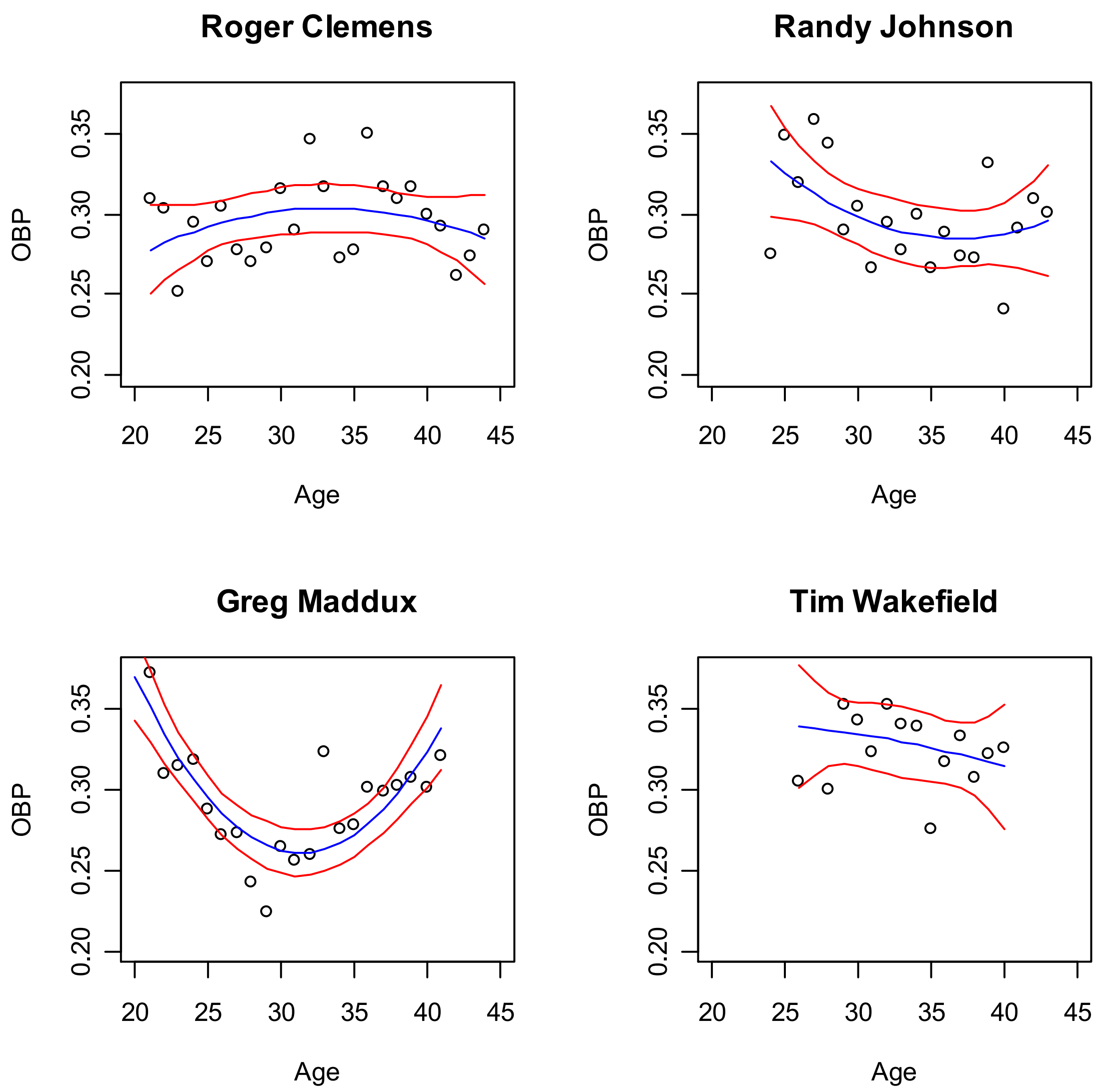

Figure 5.6: Quadratic fit with 95% confidence interval career trajectories for a few of MLB's longest tenured pitchers

Because of this strong variability within the careers of both batters and pitchers, a "one curve fits all" approach to modeling career trajectories seems to be unreasonable. Instead, a more flexible approach that utilizes quadratic polynomials to model the relationship between performance and age over a player's career may be more advantageous. Third degree polynomials were also fit to a sample of careers for both batters and pitchers but they failed to produce significantly better predictions. Also, from a subject matter perspective, it is hard to think of reasons why age should be related to performance in a cubic form.

5.3 The Aging Process: A Time Weighted Approach

A final method for predicting future performance is to search for optimal coefficients with which to discount previous year's statistics. This approach ignores the informational value that age may play in predicting performance and instead relies solely on the past *n* years of previous data. A common and practical value used for *n* in similar studies is 3 (Tango, 2006). To get initial estimates for these time weights a multiple linear regression is run on 783 batter observations, all with at least 200 plate appearances in 4 consecutive seasons. The form of this model is as follows:

$$OBP_n^k = w_{n-1}OBP_{n-1}^k + w_{n-2}OBP_{n-1}^k + w_{n-3}OBP_{n-3}^k + \varepsilon^k$$

where OBP_n^k is the *OBP* for player *k* in the current season, *n*. The fitted values for the time weights from this approach are: w_{n-1} = 0.4802, w_{n-2} = 0.3782, and w_{n-3} = 0.1378. As one would expect, seasons closer in time to the season we are trying to predict offer better information than seasons farther away.

Another approach that can be used to calculate optimal time weights is within a Bayesian MCMC routine (Allen, 2005). The appeal of the MCMC method as an optimization tool is both its robustness in locating the global minima of the phase space as well as its ability to naturally generate confidence intervals on the optimized parameters. The form of this model is as follows:

$$OBP_n^k = \frac{w_{n-1}PA_{n-1}^k OBP_{n-1}^k + w_{n-2}PA_{n-2}^K OBP_{n-2}^k + w_{n-3}PA_{n-3}^k OBP_{n-3}^k + 600M_{\mathrm{Re}gress}}{w_{n-1}PA_{n-1}^k \; w_{n-2}PA_{n-2}^K + w_{n-3}PA_{n-3}^k \; + 600}$$

where again OBP_n^k is the *OBP* for player *k* in the current season *n*, PA_i^k is the number of plate appearances player *k* had in year *n*, and w_n are the weights for which we are trying to solve. There is also an added regression component to the model to account for the fact that extreme performances one year tend to be non-repeatable. It is shown that the best value to regress to is the league average for *OBP*.

In addition to determining time weights from the past three seasons of data, Allen also measures the informational value that early season statistics can have in predicting year end performance. He defines early season statistics as those collected after a full month of play (end of April). Based on a dataset consisting of 2183 observations collected over the 1987-2003 seasons, Figure 5.7 displays the histograms of weights for both cases under study. There is clearly increasing weights for more recent seasons, indicating again that more recent data is indeed more relevant for prediction purposes. The April histogram is shifted to higher values than even the previous year's data, indicating that the information from the present year is more valuable than past data. After standardizing the median weights from each distribution so that they sum to 1, they are in fair agreement with the weights from the multiple regression estimates: $w_{n-1} = 0.5087$, $w_{n-2} = 0.3158$, and $w_{n-3} = 0.1755$. And for the case with the April information included, the resulting weights are $w_{n-1} = 0.4124$, $w_{n-2} = 0.2990$, $w_{n-3} = 0.1855$, and $w_{n-4} = 0.1031$.

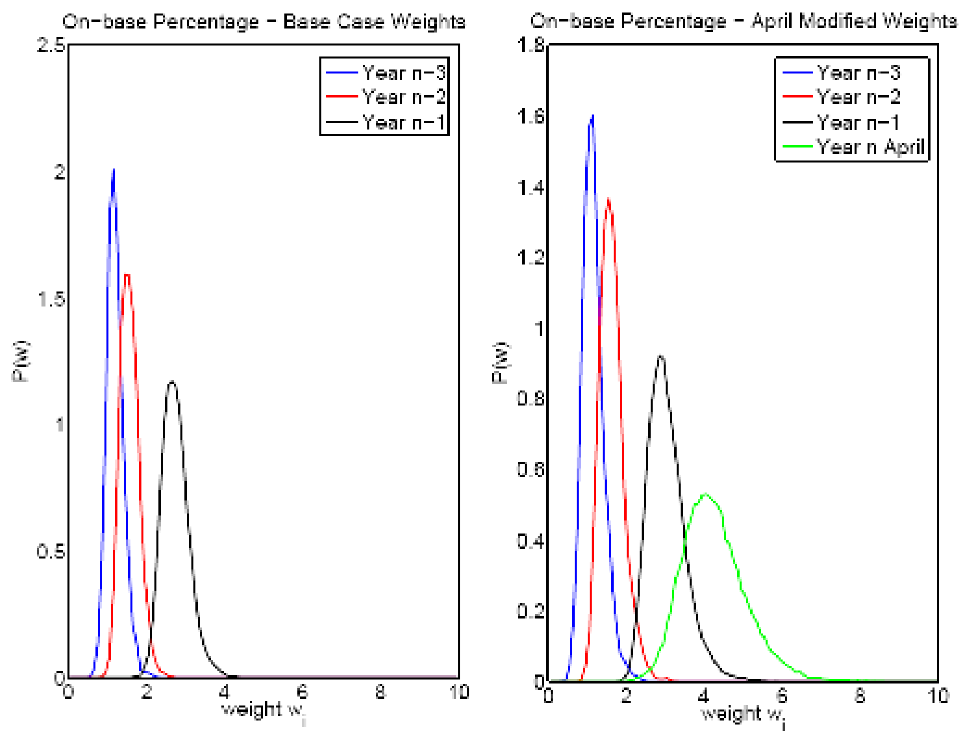

Figure 5.7: On-base Percentage model weights histograms for (a) Base case (b) April Modified Case

5.4 Predictive Accuracy of Methods

As a means of testing the accuracy of the methods discussed in this chapter, the absolute deviations of the observed versus predicted values are computed for 240 batters. These batters represent the most frequent everyday offensive starters for each of the 30 major league teams who had at least 2 years of major league experience prior to the 2007 season. The observed values are the end of the 2007 season *OBP* and the predicted values are computed using *OBP* data from the 2004, 2005, and 2006 seasons. Table (5.2) shows the mean absolute deviations when using each of the three methods (predictions using the time weighted + April statistics model not computed).

	Mean Absolute Deviation
Multilevel Model	0.0267
Quadratic Fit	0.0339
Time Weighted	0.0209

Table 5.2: MAD for the various prediction methods

These results indicate that the time weighted model provides the best predictions for future performance. The average prediction error using this method to predict OBP is almost 21 points, the average prediction errors using the multilevel and quadratic fit models are 26.7 and 33.9 points respectively. The practical significance of these differences will be explored in later chapters.

Chapter 6

A Simulation Routine for Baseball

Any simulation of a complete 9 inning baseball game is essentially a sum of a series of half inning simulations. Each half inning state begins with no men on base and 0 outs and the half inning continues until the 3 out state is reached. The probabilities of transitioning from state to state are heavily dependent upon the ability of the current batter and pitcher, situational information, and to a lesser degree, defense and park effects. Like most all sports, baseball also has its share of strategy that can be implemented throughout the game. While some of these strategies are rather difficult to implement in a simulation of baseball (defensive shifts, lineup substitutions, etc.) there are some situations in which calling upon a certain strategy can be fairly predictable. What follows is a step by step description and justification for a baseball game simulation that will be implemented and whose results will be analyzed.

6.1 Strategic Decision Making Transitions in Baseball

Special situations requiring strategic decisions occur in baseball when certain combinations of the number of outs, the location of runners on base, the score and point in the game, and the ability of the batter justify a predictably strategic move by either the batting or pitching team. Whether it is the batting or pitching team, both sides usually implement these strategic decisions in a game that is close and in the later innings. Their efforts are an attempt to maximize (offense) or minimize (defense) the possibility of scoring at least 1 run. Two of the most notable and widely accepted strategic decisions to be made in games that are close and late are the intentional walk, and the sacrifice bunt. The stolen base attempt is also a common form of strategy but its use varies significantly from team to team (Lewis, 2003). Because the frequency with which a stolen base is attempted is highly dependent on the philosophy of each team, its use cannot be seen as universal and will not be a possibility in the game simulations. It should be noted however, that allowing stolen base attempts within this player based simulation routine is entirely possible.

6.1.1 The Intentional Walk

The purpose of an intentional walk is to bypass a good hitter in order to face a batter that the defensive team feels they have a better chance at getting out, or to set up a double play ball by putting a runner on first base. The danger of issuing an intentional walk is that an extra runner is now on base for the following hitter. Both common baseball knowledge and a collection of empirical data by Tango (2006) reveal that teams are most likely to issue an intentional walk when first base is open, the current batter's OBP is above .385, it is the 7^{th} inning or later and the pitching team is down by at most 3 runs or up by as much as 1 run. Because a more relevant measure to use in this situation is OBP^* instead of just the batter's OBP, intentional walks will be issued in the simulation whenever all of the aforementioned situations occur and the predicted success of the batter (OBP^*) is greater than .385.

6.1.1 The Sacrifice Bunt

In a sacrifice bunt attempt, the batter puts the ball into play with the intention of advancing a baserunner, in exchange for the batter being thrown out. The sacrifice bunt is most often used to advance a runner from first to second base, although the runner may also be advanced from second to third base, or from third to home. The sacrifice bunt is frequently utilized in close, low-scoring games, and is usually performed by weaker hitters, especially by pitchers in games played in National League parks. A sacrifice bunt is not counted as an at-bat. Only the three most common situations for issuing a sacrifice bunt will be included in the simulation. Tables 6.1-6.3 combine empirical data collected by Tango which contains information regarding the various outcome probabilities of each situation. Tango also provides information regarding the frequency with which batters of differing abilities are asked to sacrifice. These numbers suggest that a normal distribution with a mean OBP of 0.300 and a standard deviation of 0.050 fit this data well. Meaning, for a batter with an OBP^* of 0.300, there is a 50% chance he will be asked to sacrifice when batting in a sacrifice situation. For a batter with a OBP^* of 0.250, this probability jumps to .8413.

Situation 1 – Inning > 7, Runner on 1st, 0 outs, Batting team tied or losing by 1 run		
Failed Sacrifice – 2 outs 13.5%	Successful Sacrifice – 1 out 72.9%	Infield Hit – 0 outs 13.6%

Table 6.1

Situation 2 – Inning > 7, Runner on 1st, 1 out, Batting team tied or losing by 1 run		
Failed Sacrifice – 2 outs 17.8%	Successful Sacrifice – 1 out 71.6%	Infield Hit – 0 outs 10.6%

Table 6.2

Situation 3 – Inning > 7, Runners on 1st and 2nd , 0 outs, Batting team tied or losing by 1 run		
Failed Sacrifice – 2 outs 15.9%	Successful Sacrifice – 1 out 67%	Infield Hit – 0 outs 17.1%

Table 6.3

Tables 6.1-6.3: Sacrifice bunt scenarios and corresponding outcome probabilities

As seen from these tables, a sacrifice attempt can result in one of three possible outcomes. The sacrifice attempt can cause both the runner and batter to be thrown out and increase the number of outs by 2, the sacrifice attempt can be successful and advance the lead runner while the batter is out, or, the batter may outrun the throw which results in an infield single, runners advance and no outs are accrued. When this part of the simulation is called, a random number x will be drawn from a Uniform ~ (0,1) distribution, and using the above probabilities the value of x will determine the outcome of the sacrifice.

6.2 Non-Strategic Decision Making Transitions in Baseball

For the majority of at-bats in a baseball game, intentional walks and sacrifice attempts are not an issue. Instead, transitions in the game are determined primarily between batter and pitcher. In Chapter 3, two separate methods (*Log5* and *Weighted Log5)* were introduced that attempt to predict the probability of a batter getting on base. In an effort to measure the relative accuracy of their estimates, both the *Log5* and *W-Log5* methods will be used in separate simulations of the same game predicting the percent of time a batter gets on base. The inputs for batter and pitcher ability given to these methods will be adjusted for handedness and will come from the three different performance prediction models discussed in Chapter 5. A final adjustment will take into account the quality of defense the offense is hitting against and the park in which the game is being played. This results in 6 different simulation models which will differ only in their estimates for OBP^* - the rate at which a batter of predicted ability gets on base against a pitcher of estimated ability.

6.2.1 Adjusting Outcome Probabilities due to Handedness

As discussed in Chapter 3, batter and pitcher handedness is a significant factor affecting the outcome of an at-bat. Because the size of this effect is different from player to player, the average of the batter's overall predicted *OBP* and his predicted *OBP* when hitting against a pitcher of a certain handedness will be calculated. For an example, consider the following matchup: Derek Jeter (right handed) vs. Curt Schilling (right handed) in Boston. Using the time weighted method, Jeter has an estimated current *OBP* of .409 and an estimated *OBP* of .397 when hitting against right handed pitchers. Jeter's *OBP* that will be fed into either the *Log5* or *W-Log5* process will be his predicted ability against right handed pitchers of .397. Curt Schilling has an estimated *OBPA* of .318 in a league with an average *OBPA* of .334. Using the *Log5* method, Jeter's estimated probability of getting on base at this point in the calculation is .380.

6.2.2 Adjusting Outcome Probabilities due to Park Effects

An interesting aspect of baseball compared to most other sports is that there is no standard field of play. Each ballpark is unique in its dimensions, shape, and weather conditions. Because of this, the frequencies with which runs are scored vary across parks. In order to isolate this ball park effect from the overall ability of the home team, statistics called *park factors* have been created. *Park factors* compare the rates of statistics at home versus the rate of statistics on the road by the following method:

$$PF = \frac{Home_{Rate} + Home_{Rate\_Against}}{n_{Home}} \Bigg/ \frac{Road_{Rate} + Road_{Rate\_Against}}{n_{Away}}$$

where n_{Home} and n_{Away} are the number of games played at each location. Rates higher than 1.000 favor the hitter, while rates below 1.000 favor the pitcher. Since *OBP* is the rate used throughout the simulation process, each MLB team's *OBP* park factor was computed using the above method and used to adjust the predicted OBP^*. However, because there is built in dependence between a player's statistics and his home *PF,* some of the *PF* effect is already accounted for in his statistics. In an attempt not to double count park effects, each *PF* will be average with 1, essentially cutting the effect in half. Continuing with the Jeter vs. Schilling example, Boston's Fenway Park has an adjusted PF_{OBP} of 1.047, indicating it is a hitter's ballpark. After multiplying this by the current OBP^* prediction of .380, Jeter's new expected *OBP* in Fenway Park against Curt Schilling is .398. A complete list of *PF* effects is given in Table 6.4 below.

Red Sox	1.047	Indians	1.015	Nationals	0.981
Rockies	1.045	Cardinals	1.013	Mariners	0.974
Giants	1.041	Dodgers	1.013	Twins	0.974
Orioles	1.040	Rangers	1.012	Devil Rays	0.972
Cubs	1.039	Yankees	1.011	Brewers	0.971
Angels	1.037	Tigers	1.002	Mets	0.966
Diamondbacks	1.025	Astros	0.998	Braves	0.962
Royals	1.021	Pirates	0.997	Blue Jays	0.957
White Sox	1.017	Phillies	0.991	Athletics	0.949
Marlins	1.016	Reds	0.985	Padres	0.939

Table 6.4: 2007 OBP Park Factor Effects

6.2.3 Adjusting Outcome Probabilities due to Defense

Defense is a notoriously difficult aspect of baseball to try and quantify. 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." A more objective measure for defensive ability was created in 1989 with the advent of the zone rating system. Zone rating subdivides the field of play into specific areas, or zones, for which each fielder has his assigned defensive territory. The location of every batted ball is tracked and the zone rating statistic then measures all the balls hit in these areas where a fielder can reasonably be expected to record an out and counts the percentage of outs actually made.

Defense in baseball is also unique because not all positions are created equal. Some positions, like the shortstop and catcher, see a disproportionately large of amount of balls in play that require greater skill to field. Other positions, like first base and left field, are where teams tend to place their weaker defensive players. In his *1983 Baseball Abstract*, Bill James attempts to quantify the relative importance of each position from observation and from the tendencies of players to change positions later in their careers as their defense skill erodes. What results is his defensive spectrum which moves from left to right in order of increasing difficulty and contains estimates of each positions relative importance to a team's defensive unit:

1B (6%) – LF (8%) – RF (9%) – 3B (11%) – CF (13%) – 2B (16%) – SS (18%) – C (19%)

Combining both sources of information, a player's zone rating and his usual defensive position, a weighted average for a team's overall zone rating can be produced for a given defensive lineup. Because the proper interpretation and magnitude of the resulting information within a simulation routine for baseball is unclear, only the order and not the magnitude of a team's defensive efficiency is recorded. The size of this effect is assumed to take on roughly the same range and magnitude as park factors. Table 6.5 shows that the 2007 Boston Red Sox had a very good defense. The numbers in the table can be interpreted as the percent of hits allowed by a defense. For example, for every 100 would be hits collected off of Blue Jays pitchers, the Blue Jays defense only allowed 96.5 of these hits to come to fruition. On the flip side of the things, for every 100 hits collected off of Devil Ray pitchers, a porous defense actually allowed 103.5 hits to be made. Continuing with the Jeter vs. Schilling example, because of Boston's strong defense, Jeter's updated and now final OBP^* in this situation against Curt Schilling is 0.384.

Blue Jays	0.9650	Tigers	0.9900	Rangers	1.0125
Red Sox	0.9675	Giants	0.9925	White Sox	1.0150
Cubs	0.9700	Athletics	0.9950	Royals	1.0175
Mets	0.9725	Yankees	0.9975	Angels	1.0200
Padres	0.9750	Twins	1.0000	Brewers	1.0225
Nationals	0.9775	Indians	1.0000	Reds	1.0250
Braves	0.9800	Astros	1.0025	Mariners	1.0275
Rockies	0.9825	Orioles	1.0050	Pirates	1.0300
Diamondbacks	0.9850	Phillies	1.0075	Marlins	1.0325
Cardinals	0.9875	Dodgers	1.0100	Devil Rays	1.0350

Table 6.5: 2007 Team Defensive Abilities

As previously mentioned, this process for estimating OBP^* will be performed across the 6 different combinations of input and calculation methods. A complete design framework is shown below in Figure 6.1.

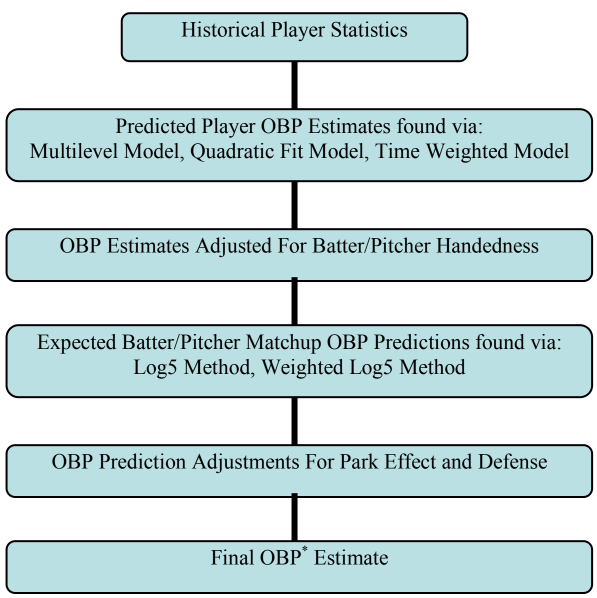

Figure 6.1: Design routine for calculating OBP^*

6.3 Determining Type of Batter Advance

After OBP^* has been calculated for each particular matchup, a random number, x, is generated from a Uniform ~ (0,1) distribution. The value of x compared to OBP^* determines the next possible set of transitions:

$$x \leq OBP^*, \text{ batter gets on base}$$

$$x \geq OBP^*, \text{ batter is out}$$

If it is decided that the batter gets on base, a criterion must now be given for which type of advancement occurs (walk, single, double, triple, homerun). Research by McCracken (2001) shows that both pitcher and batter tendencies are responsible for issuing walks and homeruns, but the frequency with which a single, double, or triple is hit is primarily a function of batter ability. In order to allow for batter and pitcher ability to determine the frequency of walks and homeruns, the *Log5* method will again be utilized. This time, however, the batter inputs will be:

$$P(walk / on\_base) = \frac{BB}{BB + h}, \quad P(HR / on\_base) = \frac{HR}{BB + h}$$

The pitcher inputs will take on the same form but will replace walks, homeruns and hits, with walks allowed, homeruns allowed and hits allowed. The 2007 major league average percent of walks and homeruns allowed, given that the batter got on base, were 23.2%

and 8.5% respectively. Once P_{BB} and P_{HR} are calculated, there is only a (1 - P_{BB} - P_{HR})% chance remaining for either a single, double, or triple. This remaining percentage is apportioned according to the historical frequency of singles, doubles, and triples hit by the batter. Tables 6.6 - 6.8 continue with the Jeter vs. Schilling example and show the individual percent of advancement by each player and also the expected percentages for this particular matchup.

Curt Schilling	
BB	12.6%
HR	10.7%

Table 6.6: Historical frequency of walks and homeruns allowed by Curt Schilling given some type of batter advance

Derek Jeter	
BB	28.2%
HR	4.7%
1B	52.5%
2B	13.1%
3B	1.5%

Table 6.7: Historical frequency of walks, homeruns, singles, double, and triples made by Derek Jeter given some type of advance

Derek Jeter vs. Curt Schilling	
BB	15.8%
HR	6.0%
1B	61.2%
2B	15.3%
3B	1.7%

Table 6.8: Expected frequency of walks, homeruns, singles, double, and triples made by Derek Jeter versus Curt Schilling.

After computing these matchup advancement percentages, we can again generate a random variable x to determine the type of advance made by the batter:

$$x \leq BB\% \rightarrow walk$$

$$BB\% \leq x \leq BB\% + HR\% \rightarrow homerun$$

$$BB\% + HR\% \leq x \leq BB\% + HR\% + 1B\% \rightarrow single$$

$$BB\% + HR\% + 1B\% \leq x \leq BB\% + HR\% + 1B\% + 2B\% \rightarrow double$$

$$BB\% + HR\% + 1B\% + 2B\% \leq x \leq BB\% + HR\% + 1B\% + 2B\% + 3B\% \rightarrow triple$$

6.4 Speed, Base Runner Destinations, and Pitcher Substitutions

All previous attempts at creating a simulation for baseball have not included individual speed rankings. All players were assumed to have the same base running ability and hence advanced equally on the base paths. This is a simplification of an important part of the game and leaves available information regarding an aspect of player ability out of the equation. A fairly robust method for assessing a player's speed was proposed by Bill James (1987). His Speed Score method is an average of a player's six following factors:

stolen base percentage, stolen base attempts, triples, runs scored, grounded into double plays, and defensive position. All of the above factors are standardized on a 0-10 scale and then averaged to arrive at a player's speed score. For interpretability within the simulation, player speed scores are transformed into percentiles. Limits are placed on these percentiles so that the slowest 15% of all players are assigned to the 15th percentile while the fastest 15% are assigned to the 85th percentile. These speed percentiles are then used to determine double play probabilities and runner destination probabilities when needed in the simulation.

If it is decided that a batter advances via either a single, double, or triple, and there is a runner on base, a model is needed to determine the runner's destination. Table 6.9 summarizes data gathered in Levitt (1999) for various hit/runner situations.

Situation	Runner Destination				# of hits
	Out	Second	Third	Home	
Single – Runner on 1st	2.1%	65.2%	31.3%	1.4%	31132
Single – Runner on 2nd	3.6%	1.2%	29.9%	65.3%	18399
Single – Runner on 3rd	0.1%		0.9%	99.0%	10133
Double – Runner on 1st	3.1%		53.6%	43.3%	6997
Double – Runner on 2nd	0.1%		1.5%	98.4%	4481

Table 6.9: Historical base runner destinations by type of hit

Because all of the above situations contain information from such a large number of hits, it is assumed that the destination percentages reflect those of a base runner with average speed. A model is now needed that gives players with greater speed a better chance of advancing further on a hit.

To accomplish this, it is easier to simplify the above information and assume that all runners are equally likely to be out on any type of hit/runner situation, runners on 1st can only advance to either 2nd or 3rd on a single, and runners on 2nd either move to 3rd or advance home on a single. If a runner's advance is then not an out, each situation above has only two possible outcomes and an odds model using player speed percentiles is used to determine runner destination. First, define $a = \left(\frac{d2}{d1}\right) * \left(\frac{speed}{1 - speed}\right)$ where *d2* is the average probability of reaching the further advance, *d1* is the average probability of reaching the closer advance, and *speed* is the player's percentile in the speed distribution. Next, let $p(d2) = \frac{a}{a+1}$ and $p(d1) = 1 - p(d2)$. For example, if a runner is on 1st, a single is hit, and the runner is not out, assume he has a 65% chance of advancing to 2nd and a 35% of advancing to 3rd (see Table 6.9). If this runner is of average speed (50th percentile) than his runner advancement probabilities remain the same. If the runner is in the 85th percentile for speed then the calculations just described allot him a 25% chance of moving to 2nd and a 75% chance of moving to 3rd.

6.4.1 Accounting for Double Plays

If a batter does not get on base during a plate appearance, *outs* increase by at least 1. Triple plays are not made possible in this simulation but double plays are possible when *outs* < *2* and there is a runner on 1st. The potential for a double play with runners on other bases is not allowed because of its rarity of occurrence. The ability of a batter to avoid a double play is due primarily to his speed. For 2007, the distribution of double play percentages is nearly normal with the average batter hitting into a double play in 13% of his double play opportunities with a standard deviation of about 7%. This information is combined with the speed distribution to determine any particular batter's chances of hitting into a double play when the situation allows it.

6.4.2 Pitching Substitutions

The flexibility of this player based simulation allows for pitching substitutions to be made within game. Starting pitchers are left in the game until their historical average number of pitches per start value is reached. For most all starting pitchers in the league, this statistic varies between 85-110 pitches per start. Batters also have a similar statistic, the number of pitches seen per plate appearance, which is added to a pitcher's pitch count each time they face a new batter. For most batters, this statistic varies between 3.5 – 4.5 pitches per plate appearance. Once a pitcher's pitch count is reached, one of the three most heavily used relievers for that team is randomly assigned to enter the game and pitch. For each game, only one middle relief pitcher has the opportunity to play. Closers for both teams only enter into the game if their team is tied or winning once the ninth inning begins. If the game goes into extra innings, both closers continue to pitch until there is a winner and the game is over.

6.5 Prediction Variability from Simulations and Computing Time

To find a suitable value for the number of times a game should be played until fairly stable results can be reached, repeated samples of the same games were simulated for various numbers of games played. The average percent difference among groups of games was recorded for each iteration value and across available betting lines. Figure (6.2) shows that after about 2,000 game simulations, results begin to stabilize with a 1 percent average difference among other 200 simulations of the same game. Graph (6.3) also displays the average run time taken for each amount of game simulations on a 2.00 GHz CPU with 1.99GB or RAM. Roughly 100 games can be simulated every 3.5 seconds.

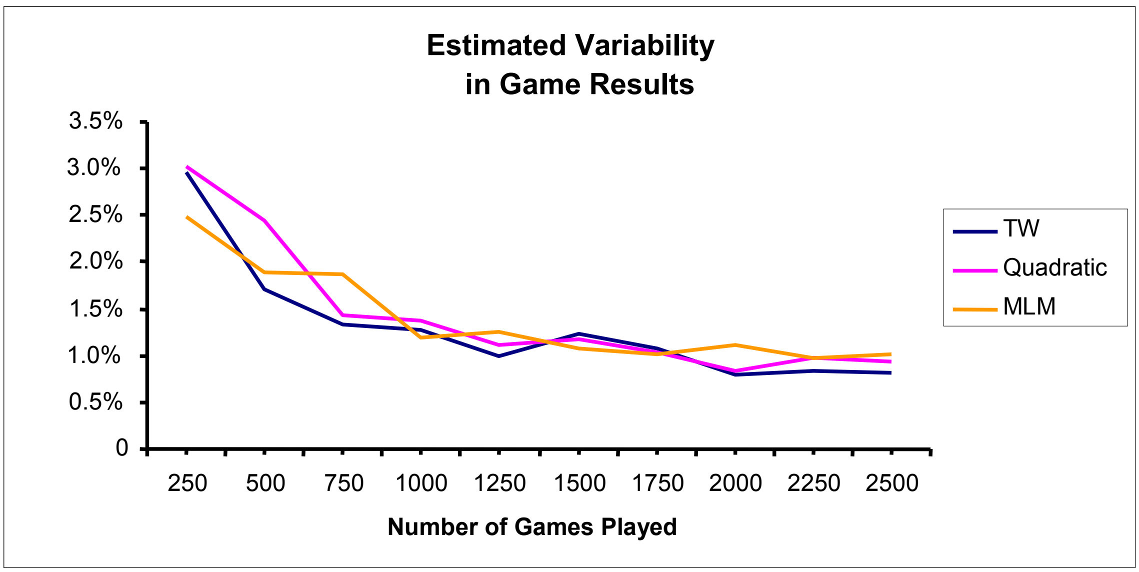

Figure 6.2: Determining the number of game simulations needed in order to achieve stable results

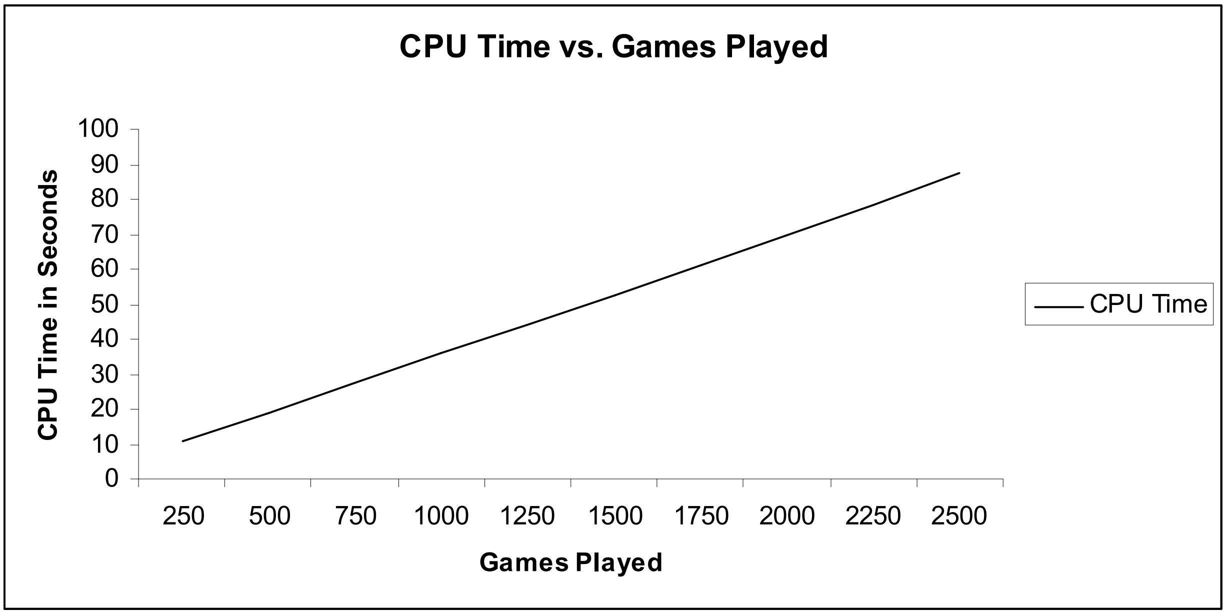

Figure 6.3: Computer run time versus number of simulations

Chapter 7

Betting Strategies and Simulation Results

7.1 Sports Gambling Markets

Many people assume that the long run outcome of sports betting is no different from that of other casino games. That is, gamblers lose money in the end because the odds favor the "house." However, the effective odds for sports betting (and horse racing) are a direct result of human decisions and can therefore potentially exhibit consistent error. Bookmakers set betting lines and the pari-mutuel pools are set by the cumulative monies bet in the pools by individual bettors. It is this fact that suggests opportunities for the gambler since the bookmaker is trying to assess public opinion so as to balance the bets. A gambler then simply needs to (a) have more insight on reality than the rest of the gambling public, and (b) win often enough to overcome the "vigorish" or house take. Unlike most mechanical games, it is evident that the possibility exists to develop winning strategies when gambling on the outcome of sporting events. Stern (1998), however, suggests that this may be a difficult task as bookmakers are good at setting lines in the sense that actual winning percentages closely resemble the probabilities implied by the lines. The major effort of academic research in gambling markets has been to determine if consistent, economically significant profits occur (often labeled as inefficiencies).

7.1.1 American and European Odds

There are many types of wagers that can be placed on sporting events. Maybe the most common is a wager placed against the point spread. For example, consider a contest between a strong team (Team A) and a weak team (Team B). Whereas popular sentiment may overwhelmingly favor Team A to win, there is typically no consensus on the magnitude of the victory. To facilitate interest in wagering on such a match, a posted line may appear as

$$\begin{array}{lll} \text{Team A} & -L & -110 \\ \text{Team B} & +L & -110 \end{array} \tag{7.1}$$

The line (7.1) is based on American odds and stipulates that a wager of \$110 placed on Team A returns the original \$110 plus an additional \$100 if Team A wins by more than L points. Alternatively, a wager of \$110 placed on Team B returns the original \$110 plus an additional \$100 if Team B wins or if Team B loses by less than L points. In the case where Team A wins by exactly L points, the original bets are returned. The quantity L is referred to as the point spread and is determined by bookmakers.

Although the -110 line is the most common line observed in sports gambling, it is in no way the only line that can be made. Many different odds may be given. For example, American odds of -120 stipulate that a winning wager of \$120 returns the original \$120 plus an additional \$100. When the American odds are positive, this suggests that an event is less likely. For example, American odds of +140 stipulate that a winning wager of $100 returns the original $100 plus an additional $140.

A nearly equivalent expression of (7.1) is based on European odds and appears as

$$\begin{array}{lll} \text{Team A} & -L & 1.91 \\ \text{Team B} & +L & 1.91 \end{array} \tag{7.2}$$

Here, a winning wager of x dollars returns $1.91x$ dollars. In the case of a $110 wager, the return is $110(1.91) = $210.10 which is nearly equivalent to the $110 + $100 = $210 situation described above.

7.1.2 Gambling on Baseball

When betting on baseball games there are no betting lines which offer varying game-to-game point spreads. Because the total amount of points (runs) tallied in a game is quite small compared to other sports (basketball, American football) there is only one standard handicapped line called the Run Line. The Run Line in baseball always handicaps the favored team by -1.5 runs. Whereas the point spread is more flexible in nature and usually settles upon a line with even odds, the odds given in the run line are much more variable from game-to-game.

Another betting line with even greater variability in odds is the Money Line. Like the point spread, the Money Line is used to equal out the attractiveness of the favorite and the underdog for the typical bettor. Money Line results are decided by a game's straight-up winner, without regard to any point spread. Odds makers set the Money Line so that more money must be risked on the favorite (the expected winner) and less money on the underdog in an effort to balance the willingness of bettors to back the respective sides of a contest.

A final popular betting line available in baseball is the Over-Under. The Over-Under line is an estimate of the total amount of runs that will be scored in a game. For example, if an Over-Under line is set at 10.5 runs for a particular match, a person who bets the over wins his bet if 11 or more total runs are scored, while a person who bets the under wins if 10 or less total runs are scored in the game. Because the totals given in the over-under line are flexible from game-to-game, the odds on either side of the line tend to be fairly even. When betting at even odds, a bettor needs to win 52.38% $\left(\frac{11}{21}\right)$ of his bets in order to break even. No such win/loss threshold can be formalized when odds are variable as winning bets placed on heavy underdogs can yield large profits and losing bets placed on heavy favorites can yield large losses.

A simple heuristic for wagering on baseball is to develop a procedure for estimating a Run Line, Money Line, and Over-Under line for a game between any two given teams. When one's personal line differs sufficiently from the posted line, this signals a condition to wager. A straightforward method for calculating such a line is via simulation. If a game is simulated 1,000 times between two teams, one may simply calculate the amount of times Team A or Team B won, the amount of times Team A or Team B won by at least 2 runs, and the amount times the two teams combined for more than t total runs. The size of the wager on any given bet may depend on the magnitude of the departure between the line set by odds makers and the estimated line determined from simulation.

7.2 Gambling Strategies

Gambling is inherently very risky with money bet at risk of complete and potentially instantaneous loss. It is well known in the gambling industry that a bettor risks financial ruin if they pursue a flat dollar betting strategy. Unfortunate runs of losing bets can eat up the entire stake quickly with flat betting so proportional betting is often used. Proportional betting, where only a proportion of wealth is bet on each bet, is the only strategy that avoids financial ruin. Any investigation of proportional betting strategies quickly leads one to the Kelly criterion (Kelly, 1956).

The Kelly criterion (or a fractional Kelly strategy) is a proportional betting strategy shown by numerous authors as the optimum money management strategy for betting. These authors include; Breiman, Hakansson, and Thorp (1975), Bell and Cover (1980), Ethier and Tavare (1983), Griffin (1984) and MaClean, Ziemba and Blazenko (1987). The Kelly criterion gives the optimum proportion of wealth to bet on each bet of a strategy in order to maximize terminal wealth. No other strategy can give a higher long term growth rate of wealth, which makes it a good candidate to use in the investigation of market efficiency.

7.2.1 The Kelly Criterion – A Derivation

One of the most concise, clear, and thorough derivations of the Kelly criterion is given by Thorp (1997) through a simple coin tossing example. Imagine that one is faced with an infinitely wealthy opponent who will wager even money bets made on repeated independent trials of a biased coin. Further, suppose that on each trial the win probability is $p > .5$ and the probability of losing is $q = 1 - p$. Let initial capital be denoted by X_0 and suppose the desired goal is maximizing the expected value $E(X_n)$ after n trials. How much should one bet, B_k, on the kth trial? Letting $T_k = 1$ if the kth trial is a win and $T_k = -1$ if it is a loss, then $X_k = X_{k-1} + T_k B_k$ for $k = 1,2,3\ldots$ and $X_n = X_0 + \sum_{k=1}^{n} T_k B_k$. It then follows that

$$E(X_n) = X_0 + \sum_{k=1}^{n} E(B_k T_K) = X_0 + \sum_{k=1}^{n} (p-q) E(B_K)$$

Since the game has a positive expectation and in order to maximize $E(X_n)$ one would want to maximize $E(B_k)$ at each trial. To maximize expected gain a player should bet all of his resources at each trial. Thus $B_1 = X_0$ and if he wins the first bet, $B_2 = 2X_0$, etc. However, the probability of ruin is given by $1 - p^n$ and with $p < 1$, $\lim_{n \to \infty} [1 - p^n] = 1$ so ruin is almost sure. Therefore this bold criterion of betting to maximize expected gain is usually undesirable.

Likewise, if one plays to minimize the probability of eventual ruin (i.e., $X_k = 0$ on the kth outcome) then the gambler's ruin formula in Feller (1966) shows that the player will minimize ruin by making a minimum bet on each trial, but this unfortunately also minimizes the expected gain. Thus this timid betting criterion is also unattractive.

This suggests an intermediate strategy which is somewhere between maximizing $E(X_n)$ (and assuring ruin) and minimizing the probability of ruin (and minimizing $E(X_n)$). An asymptotically optimal strategy was first proposed by Kelly (1956). In the coin-tossing game just described, since the probabilities and payoffs for each bet are the same, it seems plausible that an "optimal" strategy will involve always wagering the same fraction f of your bankroll.

If one bets according to $B_i = fX_{i-1}$, where $0 \leq f \leq 1$, this is sometimes called "fixed fraction" betting. Where S and F are the number of successes and failure, respectively, in n trials, then the capital after n trials is $X_n = X_0(1+f)^S(1-f)^F$, where $S + F = n$. With f in the interval $0 < f < 1$, $Pr(X_n = 0) = 0$. Thus ruin in the technical sense of the gambler's ruin problem cannot occur. Also note that since

$$e^{n\log\left[\frac{X_n}{X_0}\right]^{1/n}} = \frac{X_n}{X_0}$$

the quantity

$$G_n(f) = \log\left[\frac{X_n}{X_0}\right]^{1/n} = \frac{S}{n}\log(1+f) + \frac{F}{n}\log(1-f)$$

measures the exponential rate of increase per trial. Kelly chose to maximize the expected value of the growth rate coefficient, $g(f)$, where

$$g(f) = E\left\{\log\left[\frac{X_n}{X_0}\right]^{1/n}\right\} = E\left\{\frac{S}{n}\log(1+f) + \frac{F}{n}\log(1-f)\right\} = p\log(1+f) + q\log(1-f)$$

Note that $g(f) = \left(\frac{1}{n}\right)E(\log X_n) - \left(\frac{1}{n}\right)\log X_0$ so for n fixed, maximizing $g(f)$ is the same as maximizing $E(\log X_n)$. In this context one will usually talk about maximizing $g(f)$. Note that

$$g'(f) = \frac{p}{1+f} - \frac{q}{1-f} = \frac{p-q-f}{(1+f)(1-f)} = 0$$

when $f = f^* = p - q$.

It also follows that

$$g''(f) = \frac{-p}{(1+f)^2} - \frac{q}{(1-f)^2} < 0$$

so that $g'(f)$ is monotone strictly decreasing on [0,1). Also $g'(0) = p - q > 0$ and $\lim_{f \to 1^-} g'(f) = -\infty$. Therefore by the continuity of $g'(f)$, $g(f)$ has a unique maximum at $f = f^*$, where $g(f^*) = p\log p + q\log q + \log 2 > 0$. Moreover, $g(0) = 0$ and $\lim_{f \to q^-} g(f) = -\infty$ so there is a unique number $f_c > 0$, where $0 < f^* < f_c < 1$, such that $g(f_c) = 0$. The nature of the function $g(f)$ is now apparent.

The following theorem recounts the important advantage of maximizing $g(f)$. The details are omitted here but proofs of *(i), (ii), (iii),* and *(vi)* for the simple binomial case can be found in Thorp (1969); more general proofs of these and of *(iv)* and *(v)* are in Breiman (1961).

Theorem 1

(i) *If g(f) > 0, then* $\lim_{n\to\infty} X_n = \infty$ *almost surely, i.e., for each M,* $P(\lim_{n\to\infty} X_n > M) = 1$*;*

(ii) *If g(f) < 0, then* $\lim_{n\to\infty} X_n = 0$ *almost surely, i.e. for each* ε *> 0,* $P(\lim_{n\to\infty} X_n < \varepsilon) = 1$*;*

(iii) *If g(f) = 0, then* $\limsup_{n\to\infty} X_n = \infty$ *almost surely and* $\liminf_{n\to\infty} X_n = 0$ *almost surely*

(iv) *Given a strategy* Φ^* *which maximizes E(*logX_n*) and any other "essentially different" strategy* Φ *(not necessarily a fixed fractional betting strategy), then* $\lim_{n\to\infty} \frac{X_n(\Phi^*)}{X_n(\Phi)} = \infty$ *almost surely*

(v) *The expected time for the current capital* X_n *to reach any fixed pre-assigned goal C is, asymptotically, least with a strategy which maximizes E(*logX_n*).*

(vi) *Suppose the return on one unit bet on the ith trial is the binomial random variable* U_i*; further, suppose that the probability of success is* p_i*, where* $½ < p_i < 1$*. Then E(*logX_n*) is maximized by choosing on each trial the fraction* $f_i^* = p_i - q_i$ *which maximizes E(*log*(1+*f_iU_i*)).*

Part (*i*) shows that, except for a finite number of terms, the player's fortune X_n will exceed any fixed bound *M* when *f* is chosen in the interval $(0, f_c)$. But, if $f > f_c$, part (*ii*) shows that ruin is almost sure. Part (*iii*) demonstrates that if $f = f_c$, X_n will (almost surely) oscillate randomly between 0 and ∞. Parts (*iv*) and (*v*) show that the Kelly strategy of maximizing $E(\log X_n)$ is asymptotically optimal by two important criteria. An "essentially different" strategy is one such that the difference $E\ln X_n^* - E\ln X_n$ between the Kelly strategy and the other strategy grows faster than the standard deviation of $\ln X_n^* - \ln X_n$, ensuring $P(\ln X_n^* - E\ln X_n > 0) \to 1$. Part (*vi*) establishes the validity of utilizing the Kelly method of choosing f_i^* on each trial (even if the probabilities change from one trial to the next) in order to maximize $E(\log X_n)$.

7.2.2 The Kelly Criterion – An example

Consider the following example: Player A plays against an infinitely wealthy adversary. Player A wins even money on successive independent flips of a biased coin with a win probability of $p = .53$ (no ties). Player A has an initial capital of X_0 and capital is assumed to be infinitely divisible. Applying Theorem 1 (vi), $f^* = p - q = .53 - .47 = .06$. Thus 6% of current capital should be wagered on each play in order to cause X_n to grow at the fastest rate possible consistent with zero probability of ever going broke. If Player A continually bets a fraction smaller than 6%, X_n will also grow to infinity but the rate will be slower. If player A repeatedly bets a fraction larger than 6%, up to the value f_c, the same thing applies. Solving the equation $g(f) = .53\log(1 + f) + .47\log(1 - f) = 0$ numerically on a computer yields $f_c = .11973$. So, if the fraction wagered is more than about 12%, then even though Player A may temporarily experience the pleasure of a faster win rate, eventual downward fluctuations will inexorably drive the value of X_n toward zero.

The Kelly criterion can also easily be extended to uneven payoff games. Suppose Player A wins b units for every unit wager. Further, suppose that on each trial the win probability is $p > 0$ and $pb - q > 0$ so the game is advantageous to Player A. Methods similar to those already described can be used to maximize

$$g(f) = E(\log\left(\frac{X_n}{X_0}\right) = p\log(1+bf) + q\log(1-f)$$

Arguments using calculus yield

$$f^* = \frac{(bp-q)}{b} \tag{7.3}$$

the optimal fraction of current capital which should be wagered on each play in order to maximize the growth coefficient $g(f)$. In the context of sports gambling, b is the odds determined by the "house" for a particular event, p is the gambler's subjective probability for this same event, and $q = 1 - p$.

As an example, if a bet is seen as having a 40% chance of winning by a gambler ($p = 0.40$, $q = 0.60$), but the house is offering 2-to-1 odds on a winning bet ($b = 2$), then the gambler should bet 10% of the bankroll at each opportunity ($f^* = 0.10$) in order to maximize the long-run growth rate of the bankroll. If a gambler ever has zero or negative edge (i.e. $b \leq q / p$) then the gambler should bet nothing. For even-money bets (i.e. when $b = 1$) the formal can be simplified to: $f^* = p - q = 2p - 1$

7.2.3 Simultaneous Kelly Betting

In practice, it is likely that a sports bettor would like to bet on several games in a single day. If the bettor uses the Kelly fraction (7.3) on n such matches where $nf^* > 1$, then the total amount bet would exceed the current bankroll. The problem then is to determine fixed percentages that again satisfy some optimality.

As discussed in Insley, Mok, and Swartz (2004) the authors formally consider the situation where on day $j = 1, \ldots, m,$ a gambler wishes to place n_{ji} wagers on matches with European odds θ_i and where the probability of picking winners is p_i, $i = 1, \ldots, k.$ The question that arises is: what are the optimal betting fractions $f_{j1}^*, \ldots, f_{jk}^*$ on day j where n_{ji} wagers are placed with a fraction f_{ji} of the bankroll, $i = 1, \ldots, k$. Given an initial bankroll B_0, the bankroll at the completion of day j is

$$B_j = \prod_{x_{j1}=0}^{n_{j1}} \cdots \prod_{x_{jk}=0}^{n_{jk}} ((1 - \sum_{i=1}^{k} n_{ji} f_{ji}) B_{j-1} + \sum_{i=1}^{k} x_{ji} \theta_i f_{ji} B_{j-1})^{\Delta_{ji}}$$

where $\Delta_{ji} = I(X_{ji} = x_{ji})$, $i = 1, \ldots, k,$ in which the random variable X_{ji} denotes the number of winning wagers of type i on day j. The first term in the outer parentheses is the balance of the bankroll not bet in a given day and require $\sum_{i=1}^{k} n_{ji} f_{ji} < 1$ to prevent the possibility of ruin. Assuming that $X_{j1}, \ldots, X_{jk}$ are independent with $X_{ji} \sim Bin(n_{ji}, p_i)$, $i = 1, \ldots, k$, we are concerned with the maximization of the function $G = G(f_{j1}, \ldots, f_{jk}) = E(\log(B_j/B_{j-1}))$ where G is the exponential rate of growth and

$$G = E(\sum_{x_{j1}=0}^{n_{j1}} \dots \sum_{x_{jk}=0}^{n_{jk}} \Delta_{ji} \log(1 - \sum_{i=1}^{k} n_{ji} f_{ji} + \sum_{i=1}^{k} x_{ji} \theta_i f_{ji}))$$

$$= \sum_{x_{j1}=0}^{n_{j1}} \dots \sum_{x_{jk=0}}^{n_{jk}} (\prod_{i=1}^{k} \binom{n_{ji}}{x_{ji}} p_i^{x_{ji}} (1-p_i)^{n_{ji}-x_{ji}}) \log(1 + \sum_{i=1}^{k} f_{ji}(x_{ji}\theta_i - n_{ji}))$$

Insley, Mok, and Swartz establish the existence of a unique maximum $f^* = (f_{j1}^*, \dots, f_{jk}^*)$ and provide an algorithm that evaluates f^* for any set of independent simultaneous wagers.

7.3 Preliminary Results – Training Data

Before evaluating the accuracy of the simulations on an entire season's worth of data, training data was collected and analyzed so that any needed adjustments and improvement could be made. The training data consisted of 235 games from the first 33 days of the 2007 season. Money Line odds for each game were taken from the website www.pinnaclesports.com, an internet sports book. Simulations were not run on every game of each day for various reasons. If odds makers are uncomfortable on whatever grounds for setting a line they are under no obligation to do so and the game is termed offline. Also, because some games had starting pitchers with little or no past data to make predictions from, game simulations could not be run. The resulting 235 games represent matches from the first 33 days of the 2007 season which had posted odds and enough available player information to run simulations. Batting lineups for all AL teams consisted of the projected 9 man order listed on www.espn.com at the beginning of the season. Batting lineups for all NL teams consisted of the projected 8 man order listed on www.espn.com with the starting pitcher for that game always batting in the 9th spot. Batting lineups were not updated each day but instead consisted of the same 8 (NL) or 9 (AL) players each game. 2000 simulations were run for each game and the score and winner for each of these simulated games were recorded.

A qualitative analysis at the first set of results revealed game outcomes across all input methods (multilevel, polynomial, time weighted, Log5, Weighted Log5) that seemed rather reasonable but game scores which were inflated compared to the 2006 MLB average of 9.7 runs scored per game. Reasons for this inflation can possibly be attributed to overstated predictions from the *Log5* and *W-Log5* methods, game strategies not available within the simulation routine (defensive shifts, exact pitcher substitutions, etc.), or underestimating the effect of team defense. In an attempt to offset the exaggerated game scores without creating a bias toward any particular type of player (power hitters) or teams, a multiplicative constant value ($c < 1$) was used to deflate OBP^* for each batter/pitcher matchup. The average total runs scored from the simulated games were used as the barometer for adjustment and were set to match the average of 9.7 total runs scored per game from the 2006 season. The value of c that adjusted the scores accordingly was found at $c = .87$. While the total runs per game did change as a function of c, the percent of games won by each team remained stable as c varied.

After settling upon a suitable value for c, simulations for the 235 games were run again and a 50% Kelly betting fractions method was used as a means for evaluating simulation accuracy. Each portfolio was given a $1,000 initial bankroll. Next, adjustments to the home and away win probabilities were sought that would maximize the value of the bankroll after 33 days. The results across all input methods suggest that the simulation routine should incorporate a consistent advantage for the home team (or disadvantage for the away team) as bankroll values increased from both directions toward a 0.925 penalization value for away teams. Interestingly enough, baseball teams have historically won 7.5% more games at home than they have lost at home as mentioned in 3.4.2. Figure (7.1) shows results using the time weighted performance prediction and *Log5 OBP** input methods across different penalties against the away team. Although all portfolios are profitable after the 33 day training period, the portfolio which penalizes away team win percentages by 7.5% performs best (*away(.925)*).

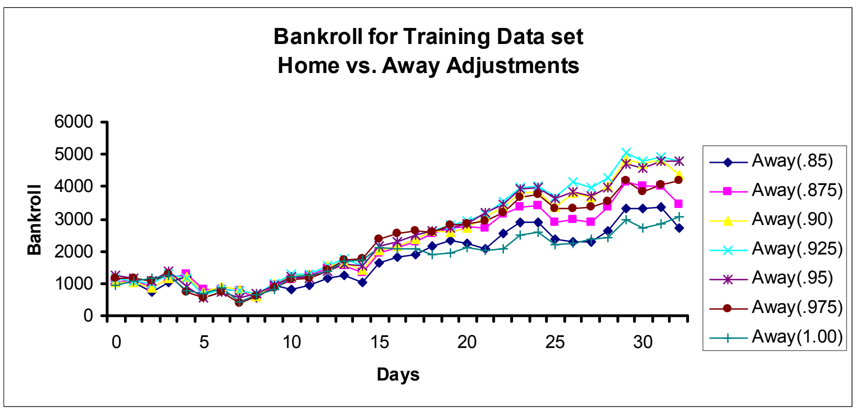

Figure 7.1: A search for appropriate home field advantage levels

7.4 Final Results – 2007 MLB Season

The improvements made upon prediction accuracy in the training data set are now put to the test against the rest of the collected data. Information for the complete data set was recorded from the same source as the training data set (www.pinnaclesports.com) and not all games throughout this period had sufficient information to make simulation predictions. The reasons for this are the same as those discussed earlier, namely, lines are not set on all games and some games contain players with a shortage of available player statistics. In total, the complete data set consists of Run Line, Money Line, and Over/Under odds for 999 games from the 2007 MLB regular season. 2,000 simulations were run for each of these 999 games with each simulation reporting back three separate summary measures: the percentage of games the favored team won by 2 runs or more and the percentage of games the underdog won or lost by less than 2 runs (subjective Run Line), the percentage of games that each team won with no handicap (subjective Money Line), and the percentage of games that combined for more total runs than the stated Over/Under line (subjective Over/Under Line). Each of these subjective odds calculated via simulation were compared to the posted line odds and a signal to wager occurred when a subjective line probability was greater than the posted line. It should be noted that some games are not bet on because the subjective percentages do not overcome the built in vigorish from the bookmaker. For example, an even money game with American odds [-110, -110] implies that both teams have a 52.4% chance of winning. The simulator calculates win percentages that always sum to 1 and so for a game where the simulator gives one team a 51% chance of winning and the other team a 49% chance of winning, no bet will be placed as the expected value from this bet is negative. An advantage from using the Pinnacle Sports website is that they are regarded as having the lowest vigorish and hence give the best odds on games. A typical even money game listed on Pinnacle is given American odds of [-104, -104], indicating that both teams are given a 50.98% chance of winning and a bettor need only win more than 50.98% of his even money bets in order to make a profit compared to the 52.4% hurdle discussed earlier. Even lower hurdle rates are possible via websites that offer customers the freedom of trading future contracts with other bettors.

7.4.1 Portfolio Specifications

Six different simulation models (*T-L5, T-WL5, Q-L5, Q-WL5, M-L5, M-WL5)* will offer predictions for each of the three types of baseball betting lines analyzed in this study. These simulation models are just an enumeration of the 6 possible combinations that can be made from the 3 performance prediction models (multilevel model, quadratic fit, and time-weighted predictions) and the 2 batter/pitcher matchup predictions for OBP^* (Log5, Weighted Log5). In addition to these 18 different simulation method/betting line combinations, 3 different betting strategies will also be imposed for a total of 54 portfolios with which to keep track. The 3 different betting strategies that will be utilized are termed: 50% Fractional Kelly, Constant Kelly, and Fixed Percentage wagering. The 50% Fractional Kelly wagering system is just a slight deviation from the original Kelly Criterion discussed earlier. The only difference is that for each day, only 50% of the total bankroll will be available for wagering. The Constant Kelly wagering system also utilizes the Kelly Criterion in its asset allocation but bets only off of a constant $5000 bankroll limit. Because of this constant bankroll amount with which to wager, the Constant Kelly system does allow the possibility for financial ruin. The Fixed Percentage wagering system does not allocate funds according to the Kelly Criterion but instead wagers a fixed percentage of the bankroll on each bet made. In this analysis, the Fixed Percentage system will allocate 5% of total bankroll for each bet placed. Also, for wager amounts determined by the Kelly Criterion, no single game is allowed to have more than 15% of all wagered funds for that day to be placed on its outcome. Each of these 54 portfolios was given a $5000 initial bankroll with which to wager. It is also important to remember that since each of these betting strategies are wagering differing amounts of money over the course of a season, one should look not only at the final bankroll values as a measure of success, but also at the percent of money returned relative to the total amounts wagered.

7.4.2 Over/Under Betting Line Performance

A quick look at Figure 7.2 and Table 7.1 reveal that across all input methods and betting strategies, the simulation predictions performed poorly against the Over/Under Betting Line. All of the Constant Kelly Betting portfolios finished with no money left to wager and each of the Fixed Percentage Betting portfolios outperformed the 50% Fractional Kelly portfolios in both total bankroll values and percent returned on investment. All but one of the input methods (*T-L5*) lost more bets then they won and because of the even odds the Over/Under Line tends to offer, an inability to overcome the 50.98% hurdle rate will inevitably result in money lost. Figure 7.2 shows a period between the 20-40 day mark when certain methods flirted with profitability but winning streaks are to be expected over such a large amount of days. In the long run though, the true inability of the simulator to give accurate predictions across all input methods and betting strategies for the Over/Under Betting Line made itself clear.

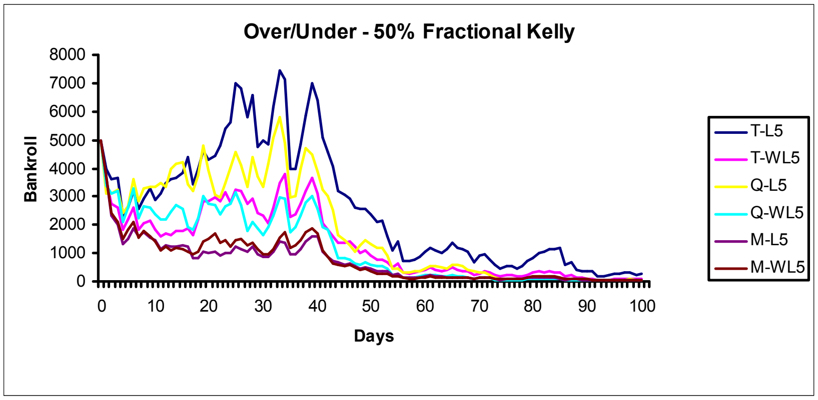

Figure 7.2: Over/Under Betting line results

Over/Under Line Performance of Models and Betting Strategies						
Performance Prediction Method	OBP* Prediction	Betting Strategy	Win vs. Loss	End Bankroll Value	Monies Wagered	% Return on Money Wagered
Time Weighted	Log5	50% Kelly	428 – 427	$261	$127,722	-3.71%
		Fixed Percentage	(50.1%)	$874	$150,367	-2.74%
		Constant Kelly		$0	$141,052	-3.54%
Time Weighted	Weighted Log5	50% Kelly	418 – 431	$78	$66,170	-7.44%
		Fixed Percentage	(49.2%)	$439	$87,147	-5.23%
		Constant Kelly		$0	$102,213	-4.89%
Quadratic Fit	Log5	50% Kelly	422 – 428	$65	$89,143	-5.54%
		Fixed Percentage	(49.6%)	$450	$119,587	-3.80%
		Constant Kelly		$0	$62,302	-8.03%
Quadratic Fit	Weighted Log5	50% Kelly	418 – 433	$36	$58,424	-8.49%
		Fixed Percentage	(49.1%)	$327	$81,906	-5.70%
		Constant Kelly		$0	$63,459	-7.88%
Multilevel Model	Log5	50% Kelly	403 – 430	$40	$34,254	-14.48%
		Fixed Percentage	(48.4%)	$238	$83,261	-5.72%
		Constant Kelly		$0	$9,438	-52.97%
Multilevel Model	Weighted Log5	50% Kelly	411 – 443	$39	$37,939	-13.08%
		Fixed Percentage	(48.1%)	$183	$66,296	-7.27%
		Constant Kelly		$0	$24,132	-20.70%

Table 7.1: Over/Under Line Performance of Models and Betting Strategies

7.4.3 Run Line Betting Performance

Figure 7.3 and Table 7.3 show improvements in the accuracy of the simulation predictions when measured against the Run Line. Of the 18 different input method/betting strategy combinations, only 5 had a final bankroll less than the original $5000. This is not overwhelming evidence of success but improvement nonetheless.

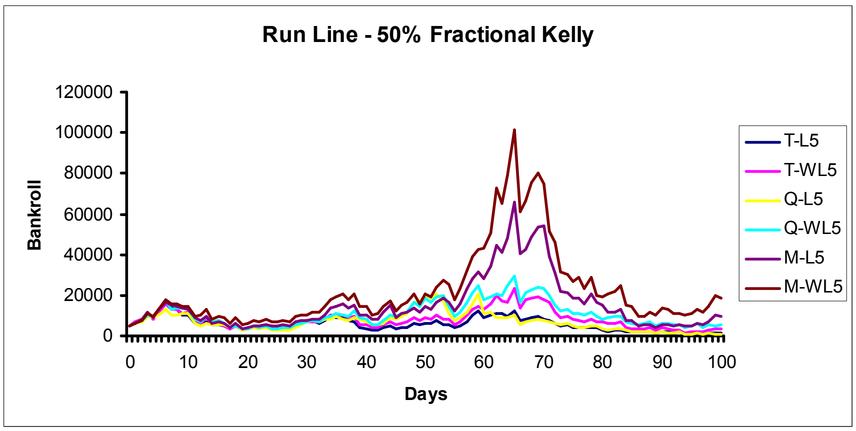

Figure 7.3: Run Line Betting Results

Figures 7.4(a-f) reveal no consistent pattern of superior performance by either of the two *OBP** prediction methods (*Log5* and *W-Log5*) or by any of the three performance prediction methods (multilevel, quadratic fit, time-weighted). Across all methods though, the Constant Kelly allocation approach demonstrated the most efficient use of money bet by having the highest return on investment in each case. Figures 7.4(a-f) also show the more tempered nature of the Constant Kelly approach as its performance across the 101 day period remains fairly stable and less volatile in comparison with the other betting strategies. Table 7.2 also shows that the fixed percentage wagering system earned both a higher percent return and finished with wealthier total bankrolls compared to the 50%

Fractional Kelly system. Across the 6 input methods, the average final bankroll using the 50% Fractional Kelly system was $6,623 and the average final bankroll using Fixed Percentage wagering system was $12,132. Similar to the behavior found in the Over/Under bankroll values, the Run Line bankrolls also exhibit an extended winning streak period between the 60-70 day range, only to decline to normal levels just as quickly as they rose.

Betting Line	Betting Strategy	Average End Bankroll	Average Money Wagered	Average Percentage Return
Money Line	50% Kelly	$46,720	$1,274,960	3.27%
	Fixed Percentage	$37,339	$1,137,630	2.84%
Run Line	50% Kelly	$6,623	$551,481	0.29%
	Fixed Percentage	$12,132	$505,395	1.41%

Table 7.2: Money and Run Line Season Performance

The failure of the Kelly approach so far in outperforming an unweighted betting scheme can most likely be attributed to the subjective nature of the posted odds as well as the simulation predictions. Section 7.2 showed the Kelly approach to be optimal when the exact amount of edge in an investment scenario is known but these properties are not as useful when both the objective (betting line estimates of game outcomes) and subjective (simulation predictions) success probabilities are still just estimates of an unknown process. Even if the predictions given by simulation are accurate enough to overcome the implied hurdle rates in the posted lines, optimal asset allocation via the Kelly approach may still be unattainable if the winning percentages implied by the posted odds are not accurate representations of reality.

One also may wonder how it is possible for the simulation prediction winning percentages over the course of a season to be lower than the hurdle rates yet still return profitable portfolios. As mentioned earlier in the chapter regarding the Run Line, all favored teams are handicapped by the same amount, -1.5 runs, and so the posted odds for each side tend to fall farther away from the even odds generally found in point spread or Over/Under betting. Because of this, a gambler may lose more bets than he wins and still be profitable as long as he more frequently bets on the underdog side of the line where less money is being wagered for possible (though presumably unlikely) greater returns. Although the win percentages from the simulation predictions in Run Line betting are not far below the hurdle rate, the overall profitable portfolios indicate that the simulator must be choosing underdog winners at a greater rate than it chooses predetermined favorites.

Run Line Performance of Models and Betting Strategies						
Performance Prediction Method	OBP* Prediction	Betting Strategy	Win vs. Loss	End Bankroll Value	Monies Wagered	% Return on Money Wagered
Time Weighted	Log5	50% Kelly	440 – 450	$1,458	$280,734	-1.26%
		Fixed Percentage	(49.4%)	$10,235	$382,673	1.37%
		Constant Kelly		$12,420	$238,884	3.11%
Time Weighted	Weighted Log5	50% Kelly	449 – 448	$3,659	$388,707	-0.35%
		Fixed Percentage	(50.1%)	$13,077	$464,072	1.74%
		Constant Kelly		$16,284	$241,052	4.68%
Quadratic Fit	Log5	50% Kelly	460 – 450	$963	$330,374	-1.22%
		Fixed Percentage	(50.5%)	$16,947	$664,073	1.80%
		Constant Kelly		$9,621	$239,992	1.92%
Quadratic Fit	Weighted Log5	50% Kelly	455 – 448	$5,334	$512,765	0.06%
		Fixed Percentage	(50.4%)	$23,432	$924,796	1.99%
		Constant Kelly		$16,609	$239,414	4.85%
Multilevel Model	Log5	50% Kelly	448 – 461	$9,748	$732,890	0.65%
		Fixed Percentage	(49.3%)	$4,467	$238,330	-0.22%
		Constant Kelly		$20,840	$240,263	6.59%
Multilevel Model	Weighted Log5	50% Kelly	450 – 458	$18,573	$1,063,415	1.28%
		Fixed Percentage	(49.6%)	$4,633	$358,425	-0.10%
		Constant Kelly		$23,290	$238,125	7.68%

Table 7.3: Run Line Performance of Models and Betting Strategies

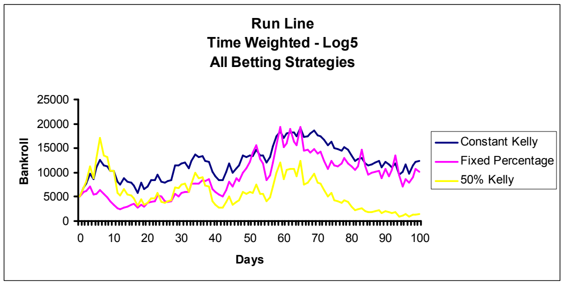

Figure 7.4(a)

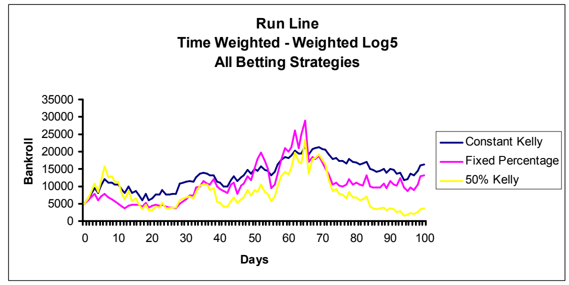

Figure 7.4(b)

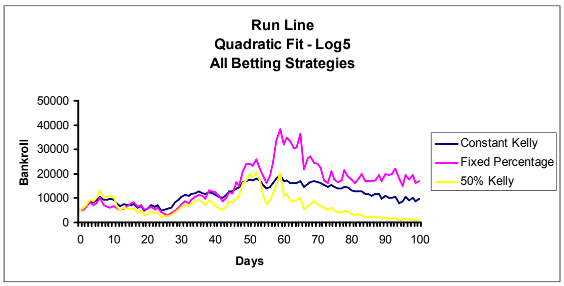

Figure 7.4(c)

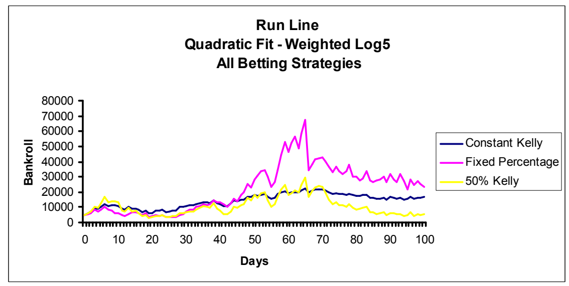

Figure 7.4(d)

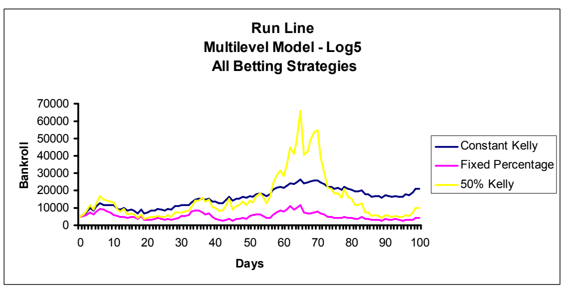

Figure 7.4(e)

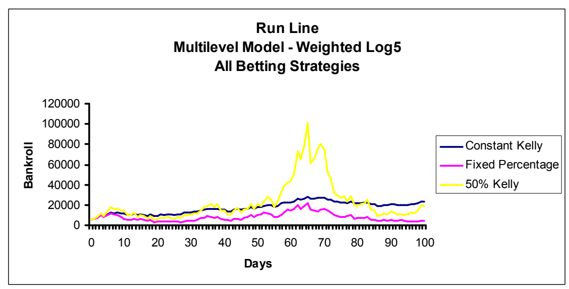

Figure 7.4(f)

7.4.4 Money Line Betting Performance

Figure 7.5 and Table 7.4 show strong improvements in the accuracy of the simulation predictions when measured against the Money Line. Of the 18 different input method/betting strategy combinations, all 18 had final bankroll values greater than the original $5000, some significantly greater. This is strong evidence of inefficiency in the MLB Money Line betting market and strong evidence of the simulation approach's ability to achieve consistent gains in this market. Figures 7.6(a-f) reveal a pattern of superior ability by the time-weighted performance prediction method in earning profits over the other two input methods but no clear advantage between the two OBP^* prediction methods is apparent (*Log5* and *W-Log5*). Perhaps not surprisingly, the Multilevel input approach outperformed the Quadratic Fit input approach, following the same order of accuracy found among the three methods in predicting *OBP* discussed in Chapter 5. Similar to what was observed with the Run Line results, the Constant Kelly allocation approach again demonstrated the most efficient use of money bet by having the highest return on investment in each case and maintained a noticeably stable behavior across the 101 day range. When profits are accruing, this type of stability is to be expected as the amount being bet is becoming disproportionately smaller in comparison to the total bankroll. The percentage scaling approach utilized in both the 50% Fractional Kelly and Fixed Percentage wagering systems is the dominant factor in the observable volatile behavior of both bankrolls. As opposed to what was observed with Run Line betting, Table 7.2 shows that the 50% Fractional Kelly wagering system earned both a higher percent return and finished with wealthier total bankrolls compared to the Fixed Percentage wagering system. Across the 6 input methods, the average final bankroll using the 50% Fractional Kelly system was $46,720 and the average final bankroll using Fixed Percentage wagering system was $37,339. Similar to the behavior found in both the Over/Under and Run Line running bankroll values, Money Line bankrolls also exhibit periods of extended winning and losing streaks.

Of the 6 input methods, 4 of them also finished with final win percentages above the 50.98% hurdle rate, essentially guaranteeing a profit for bets placed at even odds. As mentioned earlier though, the Money Line offers odds that tend to vary significantly far from even. And with the magnitude of many of the final bankrolls observed against the Money Line, it is most likely the case that a significantly greater amount of bets were placed and won on underdog teams than bets placed and won on favored teams.

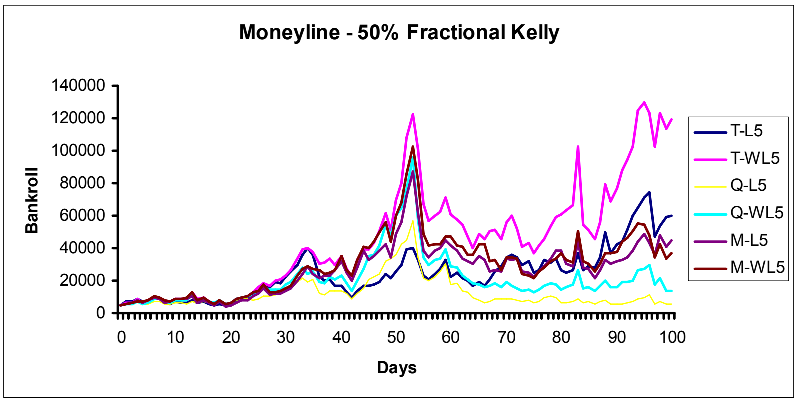

Figure 7.5: Money Line Betting Results

Moneyline Performance of Models and Betting Strategies						
Performance Prediction Method	OBP* Prediction	Betting Strategy	Win vs. Loss	End Bankroll Value	Monies Wagered	% Return on Money Wagered
Time Weighted	Log5	50% Kelly	476 – 410	$59,751	$1,150,603	4.76%
		Fixed Percentage	(53.7%)	$81,945	$1,611,594	4.77%
		Constant Kelly		$26,971	$239,300	9.18%
Time Weighted	Weighted Log5	50% Kelly	473 – 438	$119,478	$2,183,966	5.24%
		Fixed Percentage	(51.7%)	$58,215	$1,633,487	3.26%
		Constant Kelly		$30,111	$241,765	10.39%
Quadratic Fit	Log5	50% Kelly	475 – 431	$5,251	$602,306	0.04%
		Fixed Percentage	(52.4%)	$17,943	$926,870	1.40%
		Constant Kelly		$15,362	$242,302	4.28%
Quadratic Fit	Weighted Log5	50% Kelly	467 – 438	$13,817	$1,002,527	0.88%
		Fixed Percentage	(51.8%)	$33,150	$1,225,740	2.30%
		Constant Kelly		$20,650	$241,561	6.48%
Multilevel Model	Log5	50% Kelly	460 – 448	$45,115	$1,276,410	3.14%
		Fixed Percentage	(50.7%)	$14,836	$501,201	1.96%
		Constant Kelly		$25,586	$241,974	8.51%
Multilevel Model	Weighted Log5	50% Kelly	464 – 470	$36,907	$1,433,966	2.23%
		Fixed Percentage	(49.7%)	$17,943	$926,870	1.40%
		Constant Kelly		$24,331	$241,322	8.01%

Table 7.4: Moneyline Performance of Models and Betting Strategies

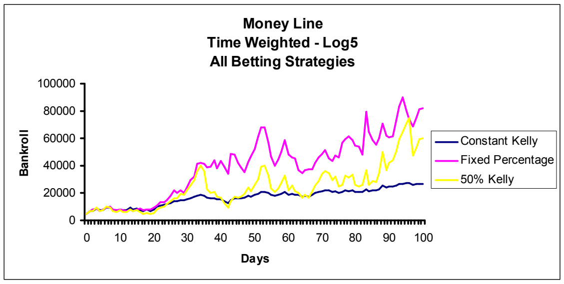

Figure 7.6(a)

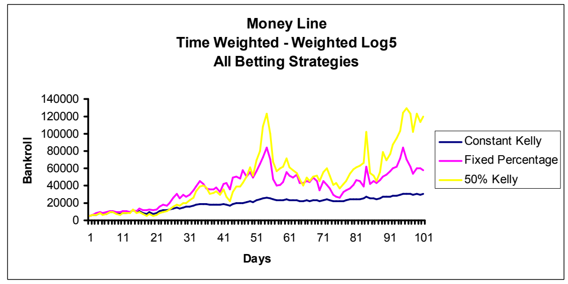

Figure 7.6(b)

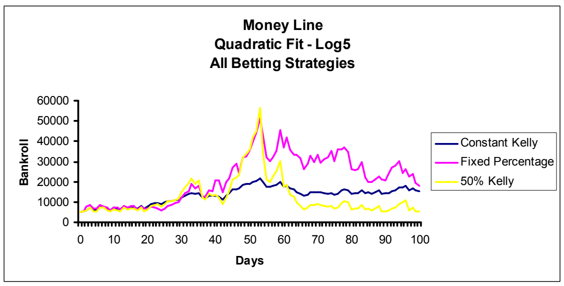

Figure 7.6(c)

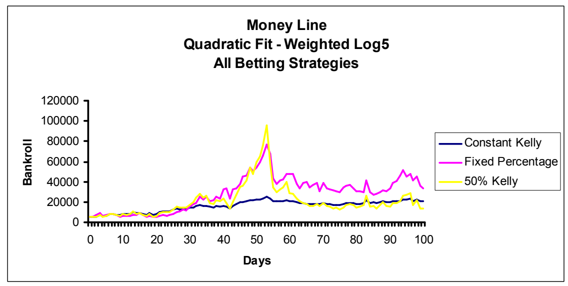

Figure 7.6(d)

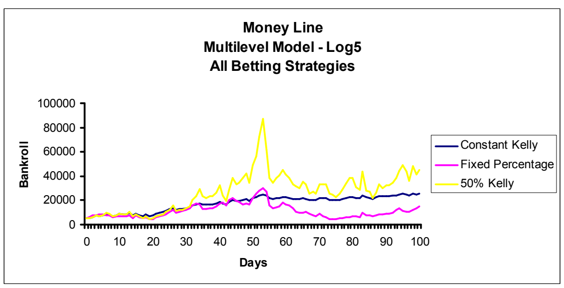

Figure 7.6(e)

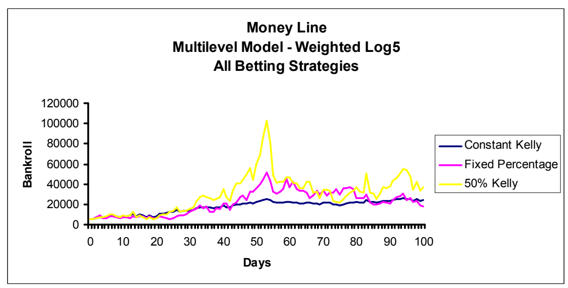

Figure 7.6(f)

7.4.5 A closer look at simulation versus posted odds predictions

At this point in the analysis, it would be reasonable to take a closer look at the behavior of the simulator relative to the posted odds by analyzing those games in which the probability of predicted winners differed the greatest. Table (7.5) shows the top 10 most disparate predictions between the Money Line posted odds and simulator probabilities using the Time-Weighted Log5 (*T-WL5*) approach. One of the first noticeable trends is that in each of these 10 games, the simulator is favoring the home team with a much greater magnitude than the posted odds. In fact, of the 911 games wagered on using this method, the simulator predicted the home team to win with a greater frequency than the probabilities implied by the posted odds 649 times. Although this may indicate that the advantage given by the simulator to the home team is too great, there is no apparent difference in the quality of predictions as the simulator was 335-314 (51.6%) when picking the home team to win, and 138-124 (52.6%) when picking the away team to win. In this particular subset of heavily wagered games, the simulator was 5-5.

Top 10 Games With The Greatest Discrepancy in Predicted Win Percentages									
Date	**Teams**	**Underdog/ Favorite**	**Posted Odds Prediction**	**Simulation Prediction**	**Prediction Difference**	**Final Score**	**Amount wagered**	**Amount won/lost**	**Starting Pitchers**
Game 1	Dbacks	U	.4434	.2161	.2273	1			Gonzalez
Sept. 11	Giants	F	.5566	.7839		2	$13,166	$10,125	Correia
Game 2	Indians	F	.6250	.4266	.1984	1			Carmona
June 22	Nationals	U	.3750	.5734		4	$1,245	$2,068	Bowie
Game 3	Angels	F	.5760	.3863	.1897	3			Colon
Sept. 14	White Sox	U	.4240	.6137		5	$18,679	$24,842	Contreras
Game 4	Yankees	F	.5777	.3899	.1878	7			Wang
Aug. 25	Tigers	U	.4223	.6100		2	$6,488	-$6,488	Bonderman
Game 5	Indians	F	.5726	.3874	.1852	3			Carmona
July 20	Rangers	U	.4274	.6126		2	$3,413	-$3,414	McCarthy
Game 6	Dbacks	U	.4280	.2435	.1845	8			Gonzalez
June 16	Orioles	F	.5720	.7565		4	$1,482	-$1,482	Cabrera
Game 7	Phillies	F	.5546	.3781	.1765	6			Durbin
Sept. 1	Marlins	U	.4454	.6219		12	$9,450	$11,435	Kim
Game 8	Dbacks	F	.5923	.4166	.1757	9			Webb
Sept. 12	Giants	U	.4077	.5834		4	$11,565	-$11,565	Sanchez
Game 9	Indians	F	.5923	.4248	.1675	8			Byrd
Sept. 11	White Sox	U	.4077	.5752		3	$10,390	-$10,390	Danks
Game 10	Astros	F	.6013	.4362	.1651	2			Oswalt
June 23	Rangers	U	.3987	.5638		7	$964	$1,436	Wright

Table 7.5: Top 10 Games With The Greatest Discrepancy in Predicted Win Percentages

A retrospective look at each of these games also reveals some common themes as to why the simulator predictions may have deviated so much from the posted odds. First, many of these games 1, 3, 7, 8, and 9 were played particularly late in the season and involve one team with playoff hopes and another with no mathematical chance of making the postseason. Although managers and coaches alike may claim that they always try to win each game with the same level of intensity, this is just not the case. Often, when a team is in no position to make the post season, they will start playing some of their minor league prospects in the hopes of evaluating their major league potential. Similarly, teams in a position to make the playoffs have a much greater motivation to win than teams that are not in the same situation. The posted odds account for these end of season quirks and the simulator does not. This can be seen in the table as the teams whose seasons are effectively over, are given a much greater chance of winning by the simulator than by the posted odds. Another reason for some of the significant departures between the two estimates can be attributed to the frequency with which both player lineups and player statistics were updated for the simulation method. Probable everyday lineups were updated for the simulator only twice throughout the entire season. Once at the beginning of recorded betting (May 31, 2007) and a second time after the trade deadline (July 31, 2007). Player statistics were also only updated twice throughout the entire season. Once at the beginning of recorded betting (May 31, 2007) and a second time during All-Star Weekend (July 4, 2007). In Table (7.5) the effects of using such aged information can be seen in games 1, 4, and 8. In games 1 and 8 the abilities of Barry Bonds were included into the lineup for the Giants when in actuality he did not play in those games. Since

Bonds has such monstrous offensive numbers, the simulator gave to great an edge to the Giants for games in which the Homerun King actually did not participate. Game 4 in the table shows the effects of not updating player statistics on a more regular basis. In this game, the starting pitcher for the Tigers (Jeremy Bonderman) was in the midst of a late season decline. Bonderman began the year with great numbers and was 10-1 at the All-Star break. Nagging injuries began to catch up with Bonderman and he finished the season 1-8 before being put on the disabled list. Since this game was played after the All-Star break, Bonderman's numbers fed into the simulator would have been stellar and accounts for why the simulator favored the Tigers in this game in such an extreme fashion over the posted odds. Although these games are the most extreme cases of disparity between simulator and posted line odds, it can only be assumed that aged player statistics and static player lineups had the same effects on other games as well. The good news, however, is that increasing the frequency with which statistics and lineups are updated should only improve simulator accuracy.

7.5 A Two-Way Analysis of Portfolio Returns

Tables 7.6 and 7.7 provide two way layouts of the average final bankroll values across betting strategies for both the Run Line and Money Line. In an effort to gauge the significance of profitability differences across the different input methods, a procedure is needed that can accurately measure these differences when only one observation (final season bankroll) is available for each case (the Over/Under betting line data is not under consideration as all portfolios lost money and have homogenous final bankroll values).

	Average Money Line Performance OBP* Calculation – Factor B		
Prediction Method Factor A	j = 1 Log5	j = 2 Weighted Log5	Average
i = 1 Time Weighted	\$56,222	\$69,268	\$62,745
i = 2 Quadratic Fit	\$12,852	\$22,539	\$17,695
i = 3 MLM	\$28,512	\$26,393	\$27,452
Average	\$32,529	\$39,400	\$35,964

Table 7.6: Two-Way Layout of Money Line Performance

	Average Run Line Performance OBP* Calculation – Factor B		
Prediction Method Factor A	j = 1 Log5	j = 2 Weighted Log5	Average
i = 1 Time Weighted	\$8,038	\$11,006	\$9,522
i = 2 Quadratic Fit	\$9,177	\$15,125	\$12,151
i = 3 MLM	\$11,685	\$15,498	\$13,591
Average	\$9,633	\$13,876	\$11,755

Table 7.7: Two-Way Layout of Run Line Performance

When there is only one case for each treatment, the option of working with a two-factor ANOVA model

$$Y_{ijk} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{7.4}$$

is no longer available because no estimate of the error variance σ^2 will be available. Recall that

$$SSE = \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij\bullet})^2 = \sum_i \sum_j \sum_k \varepsilon_{ijk}^2$$

is a sum of squares made up of components measuring the variability within each treatment, $\sum (Y_{ijk} - \overline{Y}_{ij\bullet})^2$. With only one case per treatment, there is no variability within a treatment, and *SSE* will then always be zero. A way out of this difficulty is to change the model.

$$E\{MSAB\} = \sigma^2 + n\frac{\sum\sum(\alpha\beta)_{ij}^2}{(a-1)(b-1)} \tag{7.5}$$

(7.5) indicates that if the two factors do not interact so that $(\alpha\beta)_{ij} \equiv 0$, the interaction mean square MSAB has expectation σ^2. Thus, if it is possible to assume that the two factors do not interact, we may use MSAB as the estimator of the error variance σ^2 and proceed with the analysis of factor effects as usual. If it is unreasonable to assume that the two factors do not interact, transformations may be tried to remove the interaction effects.

The two-factor ANOVA model with fixed factor levels in (7.4) when all interactions are zero so that $(\alpha\beta)_{ij} \equiv 0$, becomes for $n = 1$, the case considered here:

$$Y_{ij} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + \varepsilon_{ij} \tag{7.6}$$

Note that the third subscript has been dropped from the Y and ε terms because there is now only one case per treatment. The factor effects sums of squares *SSA* and *SSB* are calculated as before from $SSA = nb\sum_i(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$ and $SSB = na\sum_i(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2$

respectively, with $n = 1$. The interaction sum of squares

$$SSAB = n\sum_i\sum_j(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2$$

with $n = 1$ is now expressed as follows:

$$SSAB = \sum_i\sum_j(Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet})^2 \tag{7.7}$$

The number of degrees of freedom associated with SSAB in (7.7) is *(a – 1)(b – 1)*.

No new problems arise in the test for factor A and factor B main effects, or in estimating these effects. Since the expected value of MSAB is σ^2 in (7.5), the F-test statistic for testing factor A and factor B main effects can now utilize MSAB in the denominator, instead of the usual MSE.

Factor A main effects: $F^* = \frac{MSA}{MSAB}$

Factor B main effects: $F^* = \frac{MSB}{MSAB}$

7.5.1 Tukey Test for Additivity

The no interaction model when $n = 1$ just described enables us to obtain an estimate of the error variance in this case where only one seasons worth of data is available. Before continuing though, it would be prudent to determine whether or not the two factors being tested (prediction method and OBP^* calculation method) can reasonably be considered non-additive. The Tukey Test for Additivity provides a framework for measuring the validity of such an hypothesis by first assuming that:

$$(\alpha\beta)_{ij} = D\alpha_i\beta_j \tag{7.8}$$

where D is some constant. The motivation for this restriction is that if $(\alpha\beta)_{ij}$ is any second degree polynomial function of α_i and β_j, then it must be of the form (7.8) because of the restrictions:

$$\sum \alpha_i = 0 \qquad \sum \beta_j = 0 \qquad \sum_i (\alpha\beta)_{ij} = 0 \qquad \sum_j (\alpha\beta)_{ij} = 0$$

Using (7.8) in a regular two-factor ANOVA model with interactions for the case $n = 1$, we obtain:

$$Y_{ij} = \mu_{\bullet\bullet} + \alpha_i + \beta_j + D\alpha_i\beta_j + \varepsilon_{ij} \quad (7.9)$$

where each term has the usual meaning. The interaction sum of squares $\sum\sum D^2\alpha_i^2\beta_j^2$ now needs to be obtained. Assuming the other parameters are known, the least squares and maximum likelihood estimator of D turns out to be:

$$\hat{D} = \frac{\sum_i\sum_j \alpha_i\beta_j Y_{ij}}{\sum_i \alpha_i^2 \sum_j \beta_j^2}$$

The usual estimator of α_i is $\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$ and that of B_j is $\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}$. Replacing the parameters in $\hat{D}$ by these estimators, we obtain:

$$\hat{D} = \frac{\sum_i\sum_j (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})Y_{ij}}{\sum_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \sum_j (\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})^2}$$

The sample counterpart of the interaction sum of squares $\sum\sum D^2\alpha_i^2\beta_j^2$ will be denoted by SSAB* to remind us that this interaction sum of squares is for the special form of interaction in model (7.9). Substituting the sample estimates into $\sum\sum D^2\alpha_i^2\beta_j^2$, we obtain the interaction sum of squares:

$$SSAB^* = \sum_i\sum_j \hat{D}^2(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})^2 = \frac{\left[\sum_i\sum_j (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})Y_{ij}\right]^2}{\sum_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \sum_j (\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})^2}$$

The analysis of variance decomposition for the special interaction model (7.9) is therefore:

$$SSTO = SSA + SSB + SSAB^* + SSRem^*$$

Where $SSRem^*$ is the remainder sum of squares.

It can be shown (Johnson, 1972) that if $D = 0$ – that is, if no interactions of the type $D\alpha_i\beta_j$ exist – $SSAB^*$ and $SSRem^*$ are independently distributed as chi-square random variables with 1 and $ab - a - b$ degrees of freedom, respectively. Hence, if $D = 0$, the test statistic:

$$F^* = \frac{SSAB^*}{1} \div \frac{SS\,\mathrm{Re}\,m^*}{ab - a - b}$$

is distributed $F(1, ab - a - b)$.

7.5.2 Results

Tables 7.8 and 7.9 give results for both the Tukey Test for Additivity and two-way layout factor design for the Run Line and Money Line portfolio data. Both cases show no significant signs of interaction between the two factor effects and can hence use MSAB as an appropriate substitute for σ^2.

Tests for possible Interaction and Factor Effects for Run Line data presented in Table 7.5			
Hypothesis	Decision Rule	F^*	P-value
H_0: D = 0 (no interactions present)	F(.95, 1, 1) = 161	1.26	0.463
H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ (no prediction method effects)	F(.95, 2, 2) = 19	7.22	.122
H_0: $\beta_1 = \beta_2 = 0$ (no OBP^* calculation effects)	F(.95, 1, 2) = 18.5	7.63	.110

Table 7.8: Interaction and Factor Effects for Run Line Data

Tests for possible Interaction and Factor Effects for Money Line data presented in Table 7.4			
Hypothesis	Decision Rule	F^*	P-value
H_0: D = 0 (no interactions present)	F(.95, 1, 1) = 161	3.22	0.323
H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ (no prediction method effects)	F(.95, 2, 2) = 19	35.42	.027
H_0: $\beta_1 = \beta_2 = 0$ (no OBP^* calculation effects)	F(.95, 1, 2) = 18.5	0.74	.480

Table 7.9: Interaction and Factor Effects for Money Line Data

Both sets of data also exhibit no significant differences between the two OBP^* calculation methods but significant differences are found among the three different performance prediction methods within the Money Line portfolios. Referring back to Table 7.4 shows that this should not come as a surprise as the time-weighted, quadratic fit, and multilevel approaches earned \$62,745, \$17,695, and \$27,452 respectively.

Stronger conclusions regarding the true significance of these differences can only be made as future seasons worth of prediction data are collected.

Chapter 8

Applications for Simulation Baseball

With a player based approach to baseball simulation in place, a range of available applications are now made possible. A player's worth is no longer confined to an assortment of defensive and offensive statistics whose relative importance are uncertain, but player value can now be more succinctly surmised in the only real statistic of interest to fans and managers of the game, how many wins a player is responsible for creating. And unlike previous measures, this wins created variable is not restricted to the position in a lineup a particular batter bats, or the ability of one's teammates, or the home ballpark a player plays in. The true value of a player's worth can more accurately be evaluated because all of these extraneous circumstances can be accounted for by simulation. Simulating a player's worth over the course of an entire season among other teammates or within a different home ball park or in a different spot of a lineup is entirely possibly and is the strength of this individual approach to baseball simulation. Possible applications from this approach include: an objective measure of player worth for salary arbitration proceedings, trade and free agency acquisition evaluations, and optimal batting orders.

8.1 The Economics of Major League Baseball

Similar to any other business corporation, Major League Baseball franchises seek to maximize their profits. Ticket sales and local broadcasting rights are generally the two highest revenue generating devices for teams (Burger & Walters, 2003). The demand, however, for these items is strongly correlated with team performance. And as one might expect, the cost of team performance comes most directly in the form of player salaries. Being able to efficiently allocate player salaries is therefore a key component for running a successful organization in MLB.

8.1.1 The Salary Cap in Major League Baseball

Of the four major United States sports, Major League Baseball is the only one that does not have a salary cap for total team payroll. They do have what is called a salary tax, where all total salary above a certain level is taxed and shared with the smaller market teams. There is no rule as to a maximum that an MLB team can spend. By contrast, the National Hockey League, National Football League, and National Basketball Association all have hard caps – where teams are restricted to a maximum that can be spent on salaries. Each league has various minor differences, but essentially teams are restricted to a maximum team salary.

Due to the salary caps and rules in different sports, the science of allocating salary to players has become critical to success in professional sports. Baseball is unique in the incredible disparity in total team salary among the different teams (due to the lack of a salary cap). In 2007, the total payroll for the New York Yankees was $195 million. The Boston Red Sox had the second highest payroll, $143 million, and the New York Mets were third with $116 million. The Red Sox won the 2007 World Series, the Yankees were eliminated in the first round of the playoffs, and the Mets barely missed qualifying for the postseason after a historic late season collapse. By comparison, the Cleveland Indians and Arizona Diamondbacks both made it to their conference championships with only the 23rd and 26th highest payrolls respectively, and the Colorado Rockies made it to the World Series with only the 25th highest payroll. The Yankees won only 4 games more in a 162 game season than the Diamondbacks in 2007 but had a payroll nearly 4 times as large. Figure (8.1) shows the number of wins for each MLB team in 2007 as a function of total team payroll.

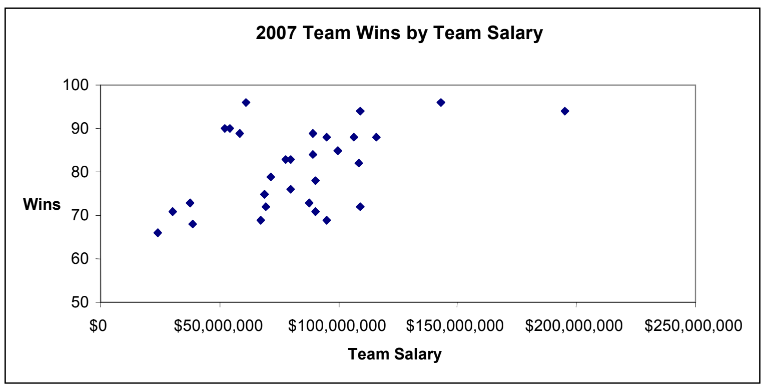

Figure 8.1: Relationship between wins and payroll for 2007 MLB teams

Table 8.1 provides an exact list of the 30 Major League Baseball teams from the 2007 season with the number of games they won, and ordered by total team salary. Values only include salaries of players on their respective 2007 rosters.

Team	Payroll	2007 Wins	Team	Payroll	2007 Wins
Yankees	$195,229,045	94	Athletics	$79,938,369	76
Red Sox	$143,123,714	96	Blue Jays	$79,925,600	83
Mets	$116,115,819	88	Brewers	$77,986,500	83
White Sox	$109,290,167	72	Twins	$71,439,500	79
Angels	$109,251,333	94	Reds	$69,654,980	72
Dodgers	$108,704,524	82	Rangers	$68,818,675	75
Mariners	$106,516,833	88	Royals	$67,366,500	69
Cubs	$99,936,999	85	Indians	$61,289,667	96
Tigers	$95,180,369	88	Padres	$58,235,567	89
Orioles	$95,107,808	69	Rockies	$54,424,000	90
Giants	$90,469,056	71	Diamondbacks	$52,067,546	90
Cardinals	$90,286,823	78	Pirates	$38,604,500	68
Braves	$89,492,685	84	Nationals	$37,347,500	73
Phillies	$89,368,213	89	Marlins	$30,507,000	71
Astros	$87,759,500	73	Devil Rays	$24,124,200	66

Table 8.1: List of 2007 MLB Team wins and payroll

Even without a salary cap in MLB, most teams are still faced with decisions on salary allocation and a critical component of success in this high dollar market is to hire players who will produce more than expected from what they are paid. In 2003, Michael Lewis wrote the influential book *Moneyball* which discusses how Oakland Athletics General Manager Billy Beane has had success evaluating potential players empirically, rather than on their physical tools. The Athletics have a smaller total payroll than most teams because of the limited income potential of their smaller market. *Moneyball* describes how Beane uses statistical techniques to identify empirical traits of players who are undervalued in the MLB market. As part of his analysis, Beane finds that OBP is a valuable asset for producing runs and is an undervalued statistic in the MLB community. He then assembled a team with high OBP averages, but lower salaries, thus producing a more productive team for the total payroll to which he was constrained. This ability to not only identify and classify players as undervalued or overvalued but to also be able to quantify how much any player is worth to a particular team is again a feature of the simulation approach.

8.1.2 Previous Research Estimating Player Worth

Previous work has been done in the area of evaluating an individual's worth or value added to a team, but not within a simulation framework. James (2002) and Berry (2000) develop models for measuring the overall value of individual players by predicting the amount of runs a particular player is responsible for creating over the course of a season. This runs created (*RC*) variable is then used to predict how many wins these estimated amount of runs created are equivalent to for a team. Similar measures are used to predict the amount of runs prevented, and hence wins created, by pitchers. Defense is in large part ignored. In Berry (2006) the details of modeling the offensive aspects of baseball are presented. The approach is to use the cumulative team totals for runs scored and model the individual contribution of hits, walks, steals, and outs to runs scored. Using the team season totals for every team from 1995 through 2004, the linear regression model for runs scored found by Berry is:

$$Runs = 100.0 + .59(1B) + .71(2B) + .91(3B) + 1.48(HR) + .30(BB) + .27(SB) - .14(OUTS) - .20(CS)$$

The model has an $R^2 = 0.927$, and a standard error of 24.3 (the mean runs scored in a season is 778 with a standard deviation of 89). The above model is used to provide a runs created value for each player (ignoring the intercept for an individual). For example, in the 2007 season, AL MVP Alex Rodriguez had 583 at-bats, 98 singles, 31 doubles, 0 triples, 54 homeruns, 95 walks, 24 stolen bases, 4 times caught stealing, and 415 outs made resulting in a predicted value of 135.83 runs created (*RC*). Berry reports that through 1995-2004, the average number of runs created per plate appearance is .118. This means that an average player – given Rodriguez's 699 *PA* - would create 82.48 runs. The differential of the runs created from Rodriguez's 2007 MVP season compared to an average player is 53.4. Similar methodology is used for pitchers to measure their difference in runs allowed from average.

Next, both Berry and James call on a commonly used "property" of baseball – the Pythagorean Theorem of Baseball – to convert their estimates for runs created into estimates of wins created. The Pythagorean Theorem of Baseball proposes that the ratio of wins to losses of a baseball team will be approximately the ratio of runs scored squared to runs allowed squared.

$$W\hat{i}n\% = \frac{RunsScored^2}{RunsScored^2 + RunsAllowed^2}$$

Throughout the last 10 years in MLB, a hypothetical average team scores and allows 778 runs – clearly resulting in an expected 0.500 record of 81-81 for a 162 game season. If a team could increase its runs scored by 9.665 over a season while maintaining the same runs allowed, it's expected winning proportion would be

$$\frac{(778+9.665)^2}{(778+9.665)^2+778^2}=.50617$$

The expected number of wins in 162 games with this wining proportion is 82.0. This now places in context the difference one player can be expected to make to a baseball team. If a player is 10 runs better over the course of a season than an average player, he will contribute approximately one more win per season than an average player. Estimating the economic value of a win to a club is the next step in determining player worth.

8.1.3 The Economic Value of a Win

Previous research on pay and performance in baseball (Scully, 1989) notes that team revenues are directly related to the club's win percentage and to the size of the market from which it draws fans. This suggests that players have greater value in larger markets than they do in smaller ones. But in Scully's model of players' marginal revenue products, market size and team performance independently affect team revenue, implying variation in market size influences just the intercept of each team's revenue function and not its slope. By Scully's estimates, a player who produces, say, 10 extra wins for his team would have the same marginal revenue product whether he played in Kansas City (small market) or New York (large market).

In Burger & Walters (2003) the authors build upon the original Scully model and outline a general model of fan behavior and team revenues in which they allow team performance to affect the intensity with which fans follow their team and the likelihood that the team will benefit from bandwagon effects as it wins more games. As a result, market size and team performance will interact in a unique fashion in determining the marginal value of an extra win. This allows for the possibility that a larger market team might enjoy a revenue function with a greater intercept (in a plot of revenues against wins) and a steeper slope. As a result, larger market teams would attach a greater value to a player of given marginal productivity (in wins) than would smaller market teams—and, ceteris paribus, rationally would bid more for available talent.

Their argument begins by assuming there are two types of fans who generate revenue for sports franchises (whether by purchasing tickets or merchandise or by tuning in to broadcasts of games). In each market, there are "purists," who follow the home team regardless of its performance, and "bandwagoners," who follow the home franchise only when it is having a successful season. Mathematically, this can best be described by:

$$P = \alpha M \tag{8.1}$$

$$B = \phi M \tag{8.2}$$

$$0 \le \alpha + \phi \le 1 \tag{8.3}$$

$$F = P + B \tag{8.4}$$

where parameters α and ϕ represent the fraction of a team's market population M who are purists P or bandwagoners B, respectively, and the team's total fan base F is the sum of these different types of fans.

By definition, the population of purists (P) is not affected by team performance (although the intensity with which they follow the team is so affected, as discussed below). The population of bandwagon fans (B), however, is directly affected by team performance, as some fans "jump on the bandwagon" only when the team is successful—that is, is "in contention"—and pay no attention otherwise.

Specifically:

$$\phi = \Phi(W) \quad (8.5)$$

$$\Phi'(W) > 0 \quad (8.6)$$

$$\Phi(W) = \begin{cases} 0 \ldots W < W^{*} \\ \theta(W) \ldots W \geq W^{*} \end{cases} \quad (8.7)$$

where *W* denotes team wins and *W** denotes the threshold number of wins a team must achieve (or be capable of achieving) for fans to consider it in contention, thus inducing the bandwagon effect. Now (8.2) can be rewritten to reflect this definition of ϕ:

$$B = \Phi(W)M \quad (8.8)$$

As a result, each team's fan base (*F*) is a function of market size and team performance:

$$F = P + B = \alpha M + \Phi(W)M = \Gamma(M, W) \quad (8.9)$$

Team revenue will be sensitive not only to the total number of fans but also the intensity with which they follow the team. Once fans are in the market (be they purists or bandwagoners), we assume they will attend or tune in to more games as team performance improves, so that:

$$I = \Psi(W) \quad (8.10)$$

and

$$\Psi'(W) > 0 \quad (8.11)$$

where *I* denotes the level of fan intensity, defined as the amount of revenue generated per fan. Team revenue (*TR*) is then determined by multiplying the number of fans by the intensity function:

$$TR = F \cdot I = \Gamma(M,W) \cdot \Psi(W) \tag{8.12}$$

Note the model's unique implications regarding the marginal value of a win. The derivative of team revenue with respect to wins is the sum of two components:

$$\frac{dTR}{dW} = \Psi(W)\frac{d\Gamma}{dW} + \Gamma(M,W)\frac{d\Psi}{dW} \tag{8.13}$$

Because the first component, $\Psi(W) \cdot (d\Gamma / dW)$ drops out for non-contending teams, (8.13) shows that, ceteris paribus, the marginal value of a win will be greater for a contender than a non-contender. That is, each team's revenue function will be kinked; its slope will increase significantly at the win threshold needed to achieve contender status. (8.13) also shows that market size and team performance will have an interactive effect on marginal revenue. Even if a team is not a contender, extra wins will have value because they increase the intensity of purists, and extra wins will have more value the greater is the population of purists. For example, increasing the intensity with which a million fans follow a team will deliver more revenue than increasing the intensity of, say, half a million. In addition, if a team crosses the threshold into contender status, extra wins will increase the fraction of the population that becomes bandwagoners, φ, and the intensity with which they (and, again, the purists) follow the team. Thus, market size affects the marginal value of a win through the intensity and bandwagon channels. As a result, the revenue function for a team in a larger market will likely have a greater slope than that for a team in a smaller market, both below and above the win threshold needed to achieve contender status.

Burger & Walters conclude that the marginal value of a win to a particular team—a crucial variable for translating an anticipated level of performance into appropriate pay for players—will depend on an interactive relationship between the size of the team's market and the team's expected performance. Table 8.2 is taken from Burger & Walters and shows their estimates for the marginal value of win for each team in either contender or non-contender status. These estimates are based primarily on the local population size of the market within which each team plays. More precise estimates for marginal win values would require an in depth look at the actual financials of each team.

Franchise	1999 Population (millions)	MV of Win as a Contender (millions of 1999 dollars)	MV of Win as a Noncontender (millions of 1999 dollars)
Anaheim	7.96	2.85	0.460
Arizona	3.03	1.09	0.175
Atlanta	3.86	1.38	0.223
Baltimore	7.35	2.63	0.425
Boston	5.66	2.03	0.327
Chicago Cubs	4.45	1.59	0.257
Chicago White Sox	4.45	1.59	0.257
Cincinnati	1.96	0.70	0.113
Cleveland	2.91	1.04	0.168
Colorado	2.42	0.87	0.140
Detroit	5.53	1.98	0.320
Florida	3.73	1.34	0.216
Houston	4.49	1.61	0.260
Kansas City	1.76	0.63	0.102
Los Angeles	7.96	2.85	0.460
Milwaukee	1.65	0.59	0.095
Minnesota	2.86	1.03	0.165
Montreal	3.40	1.22	0.197
New York Mets	10.11	3.62	0.584
New York Yankees	10.11	3.62	0.584
Oakland	3.46	1.24	0.200
Philadelphia	6.00	2.15	0.347
Pittsburgh	2.34	0.84	0.135
San Diego	2.83	1.01	0.164
San Francisco	3.46	1.24	0.200
Seattle	3.47	1.24	0.201
St. Louis	2.57	0.92	0.149
Tampa Bay	2.29	0.82	0.132
Texas	4.92	1.76	0.284
Toronto	4.50	1.61	0.260

Table 8.2: Local Revenues in MLB (source: Blue Ribbon Panel report)

Perhaps an even more realistic model than the one proposed by Burger & Walters would be one that did not assume a linear relationship between team wins and revenue earned with a single bump or kink in the revenue function once a team is deemed to have reached contender status. Work by Silver (2006) has shown that baseball teams can expect a substantial nonlinear bump in revenue as a result of making the postseason. He estimates that after all categories of revenue are accounted for and revenue-sharing payments are deducted, an additional regular-season win is worth about $650,000 and a playoff appearance $25 million for an average market team. Table (8.3) provides estimates by Silver of the additional revenues an average market MLB team can expect to make from each win and from making the playoffs.

Revenue Category	Additional Regular-Season win	Additional Playoff Appearance
Regular-season gate receipts	$705,000	$14, 869,000
Concessions	$214,000	$4,514,000
Luxury suites and club seats	$157,000	$3,314,000
Postseason gate receipts	$0	$5,797,000
Merchandise	$56,000	$1,182,000
Local broadcast rights	$0	$14,093,000
Revenue-sharing payments (34%)	($385,000)	($14,881,000)
Total	$747,000	$28,887,000

Table 8.3: Two-Tiered Model of revenue for an average MLB Franchise

This type of model would work well if all one wanted to do was look back and guess how much money a certain team had made over the course of a season. Using the information from Table (8.3) we would expect a team that won 93 games and made the playoffs, for example, to have made about $30 million more than a team that won 85 games and missed the playoffs.

For planning purposes, however, the above model is less useful, since a team cannot know in advance whether it will make the playoffs. Using historical data, however, we can estimate with some reasonable degree of certainty the probability of a playoff appearance given a particular number of wins. Between 1996 (the first full season with the Wild-Card system in place) and 2007, the team with the most number of wins that didn't make the playoffs was the 1999 Cincinnati Reds, which won 96 games but lost the division race to the Houston Astros and the Wild-Card race to the New York Mets, each of which won 97 games. Conversely, the team with the fewest number of wins that did reach the playoffs was the 1997 Houston Astros, which won the NL central with 84 wins. So we might say that a team that wins fewer than 80 games has essentially no chance of reaching the playoffs, while a team that wins more than 96 games is just about guaranteed to reach the playoffs. Figure (8.2) is an estimate of the probability of reaching the playoffs based on the number of regular-season wins.

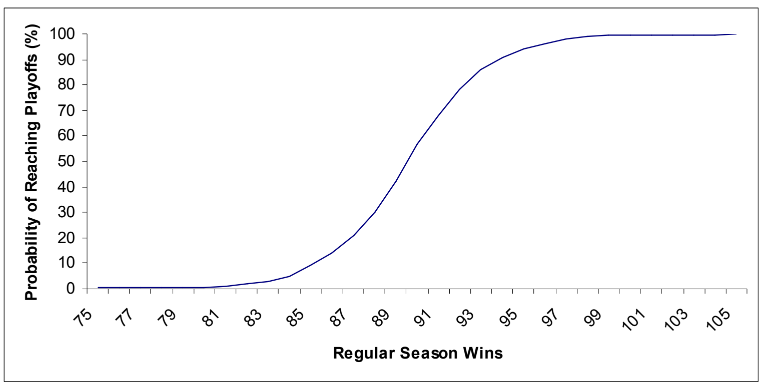

Figure 8.2: Probability estimates for reaching the playoffs given number of regular season wins

The most sensitive part of Figure 8.2 is between 86 and 93 wins. Winning 90 games rather than 89, for example, improves a team's chances of making the playoffs by about 13 percent. We can look at a couple of other presentations that help to illustrate what a profound impact this critical region can have on optimal team behavior. Figure (8.3) estimates the amount of local revenue that a team can expect to have, starting with a baseline of $22 million at 60 wins and attributing additional monies both directly based on increasing regular season wins and indirectly based on improved probability of reaching the playoffs. As can be seen, revenues take a sharp turn upward at about 83 wins - when fans deem their team to be within playoff contention - but slow down again at about 94 wins, when additional wins become superfluous.

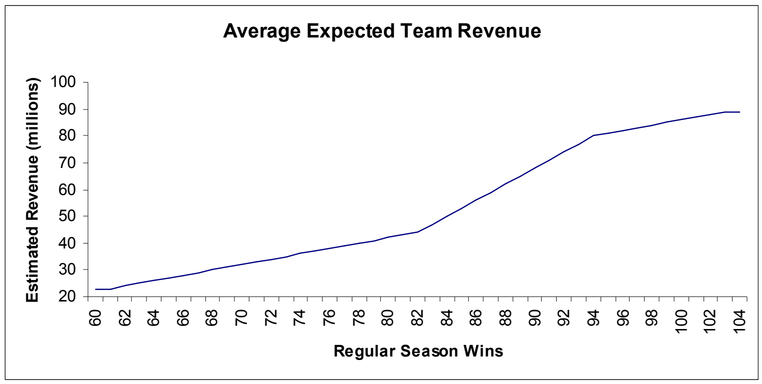

Figure 8.3: Total Revenue earns the greatest returns from wins within the 83-93 win range

We can further extrapolate the marginal economic value of winning one additional game as shown in Figure (8.4).

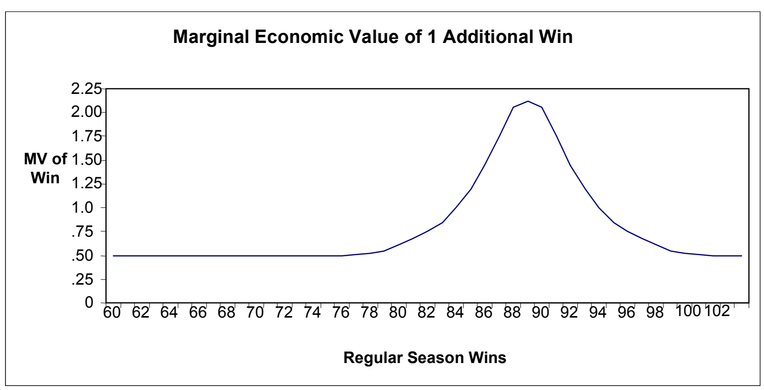

Figure 8.4: The Marginal Value of a win is most profitable within the 83 to 93 win range

This approach of estimating unique marginal values for wins within the playoff contention range (83-96 wins) seems much more plausible than a demand curve with a single kink representing contention status. Front office personnel would be wise to utilize this type of information when assessing possible trades or free agent signings. A player who is expected to contribute 6 more wins to a team with him on their roster is much more valuable to a team which won 83 wins in the previous year, than a team which won 96 games the previous year. As was mentioned earlier in this chapter, this player would also be worth more to a large market team, where marginal revenues are greater per win, than a smaller market team.

8.2 Salary Arbitration in Major League Baseball

Currently, Major League Baseball has a dispute resolution method known as Final Offer Arbitration ("FOA") as an exclusive method for resolving contract based salary disputes between players and owners. This relatively new form of dispute resolution, final offer, or last-best offer arbitration, limits an arbitrator to choosing the final offer made by one of the parties. It is designed to motivate each party to negotiation in good faith and genuinely attempt to compromise in order to create a final offer that an arbitrator will select as most reasonable. When Final Offer Arbitration is used to resolve a bargaining impasse, each party submits a proposed monetary award to the arbitrator before the hearing, at the conclusion of which, the arbitrator must choose one award without modification or compromise. This approach imposes drastic restrictions on the arbitrator's discretion and provides each party an incentive to offer a reasonable or defensible proposal, in the hope that it will be selected by the decision maker.

FOA facilitates good faith bargaining by motivating the parties to calculate and present their most reasonable positions prior to the arbitration. Final offers in salary arbitration tend to be reasonable, because they may not have been proposed in prior negotiations and the offers are exchanged simultaneously, making it impossible for the parties to base counteroffers upon the other's offer. Therefore, it forces each party to independently determine a reasonable and defensible monetary valuation of a player that is likely to be found more reasonable than the value reached by the other side.

The distributive incentive to settle in FOA arises out of the midpoint between the final offers, which provides the arbitrator a starting point for valuation and a range of numbers in which the parties can negotiate. This enables each party to evaluate the offers presented and the likelihood that the arbitrator would determine the value of the player to be greater or less than the midpoint. The offer that is closer to the value determined by the arbitrator is likely to win the arbitration, which allows the parties to settle accordingly. Thus, when organizations submit offers significantly lower than the players' proposals, they often scramble to settle on the player's side of the midpoint before hearings conclude out of fear of losing entirely.

Players and clubs may end up in salary arbitration through two primary routes. First, any player's salary may be determined by final offer arbitration if both he and the club consent to the procedure. Second, some players are eligible to have their salaries determined in arbitration without their club's consent. This second route is by far the most common way in which players and clubs end up in arbitration. Originally, any player with at least two years of major league experience was eligible to file for arbitration without club consent. But as part of the settlement of a strike in 1985, the players and clubs agreed that, effective in 1987, the minimum eligibility requirement for arbitration without club consent would be increased to three years' experience. Since then, the eligibility requirement has again been changed, so that today 17% of the players with at least two but less than three years of experience are also eligible. Currently, players and clubs must notify each other of their intention to seek arbitration between January 5 and January 15 of the year of the season for which the player's salary is to be determined, and final salary offers must be submitted within three days after such notice is received, or as soon as practicable. Arbitration hearings are scheduled during the period February 1-20. In the interim between submission of the offers and the hearings, players and clubs are permitted to continue to negotiate, which they almost always do.

Only if the player and club do not reach a settlement by a specified date does an arbitrator conduct a hearing. At the hearing, the player and club are each permitted one hour to present evidence and one-half hour to rebut the other side's case. Following the hearing, the arbitrator has 24 hours to choose one of the offers, which will then be the player's salary for the following season. The arbitrator cannot impose a compromise and is prohibited from issuing an opinion explaining the reasons for the final decision. Table 8.4 summarizes data on baseball arbitration cases for the time period 1984-91. The table shows the means of the players' and clubs' offers for each year in current dollars. Although these figures demonstrate that salaries have increased substantially, the means of final offers are not accurate measures of salary levels by year because the numbers of cases are small and players' abilities vary widely.

Year	Number of Cases	Mean of Club Offers	Mean of Player Offers	Number of Player Wins	Percent Player Wins
1984	10	$358,000	$491,500	4	40
1985	13	$425,769	$588,769	6	46
1986	35	$412,143	$573,400	15	43
1987	26	$602,692	$762,500	10	38
1988	18	$616,497	$767,667	7	39
1989	12	$626,250	$879,083	7	58
1990	24	$725,417	$1,060,000	14	58
1991	17	$1,167,794	$1,725,294	6	35

Table 8.4: Baseball Arbitration Cases, 1984 – 1991 (Sources: Scully (1989:163-164))

The rules of FOA specify that the arbitrator is obligated to consider six criteria in making a decision: the quality of the player's contribution to his club during the past season; the length and consistency of the player's career contribution; the player's past compensation; comparative baseball salaries (the arbitrator is given a tabulation of the preceding season's salaries for all players on major league club rosters); physical or mental defects affecting the player's performance; and the club's recent performance.

A visible problem with FOA in MLB is the limited information an arbitrator is given for making a decision as well as the subjective nature of the arbitrator’s final decision. For each salary dispute case, there is a host of possible arbitrators to choose from. It is highly unlikely that arbitrators presented with the same information and salary offers will always reach the same conclusion. Similar concerns are made in an article regarding salary evaluation for professional baseball players by Lackritz (1990)

> The time has come for the owners and the players to sit down and construct some reasonable models for salary evaluation. Arguments will also revolve around the measurement of intangibles such as hustle, desire, clutch performance, and guts. What should we do about the players who give up an at bat to advance a runner from second to third base? Players would obviously be unwilling to admit to the possibility that many are overpaid. Players will not be worth the same in both leagues; in fact, adjustments of the models (Lackritz 1986b) will vary players' worth from team to team. Additional research and testing must be done to fairly weight in-park factors such as whether the park is domed, weather conditions, hitter's versus pitcher's parks, day games, artificial turf, and other factors. Nevertheless, the players could easily generate enough information to develop an appropriate scale, according to conditions. Otherwise, players will want to go to good-hitting parks like Boston, Minnesota, and Seattle, but pitchers will want to pitch in Houston, San Francisco, and Chicago (White Sox), where the parks are thought to help the earned run average (ERA). I hope a model such as this can be useful in arbitration cases, free agency, trades, and long-term contracts.

A straightforward solution to these shortcomings is player based simulation. Substituting a player within different lineups, in different leagues and conditions are all possible. Some metric can then be devised that takes into account the amounts of wins a player would be predicted to create in different environments and a more objective measure for player value will result. Whatever offer (team or player) is closer to the salary generated via simulation will become that player’s salary for the upcoming season. In the absence of free agency where players can sell their skills to highest bidder, it seems reasonable that a more objective standard for salary determination should be utilized within Major League Baseball.

8.3 Is Alex Rodriguez Worth $27.5 Million a Year?

The most lucrative player contract in MLB history was made in the off season of 2008 when Yankees third baseman and 2007 MVP Alex Rodriguez signed a 10 year, $275 million dollar contract to remain with the team. Although Rodriguez is a proven talent with years of similar productivity still expected from him, many have raised the question if any player is worth this salary amount. As mentioned earlier, previous attempts to answer this question used static methods that focused solely on the individual and could not factor in other contributing factors to a player's overall worth such as teammate abilities, park effects, batting order, etc. With simulations taking place at the individual level though, players can be substituted in and out of fantasy lineups and projections can be made of how a player and team would fare through the course of an entire 162 game season. For example, using the projected starting lineups and pitching rotations of all MLB teams given by www.espn.com for the 2008 season, the player based simulator predicts the following final standings for the 2008 MLB season. These results are based on 2,000 simulated seasons of 2430 games each.

American League 2008 Predictions								
AL East			**AL Central**			**AL West**		
Team	**Wins**	**Losses**	**Team**	**Wins**	**Losses**	**Team**	**Wins**	**Losses**
Yankees*	95		Tigers	89*		Angels	85*	
Red Sox*	92		Indians	86		Rangers	79	
Blue Jays	83		White Sox	79		Athletics	79	
Rays	76		Twins	79		Mariners	74	
Orioles	73		Royals	71				

National League 2008 Predictions								
NL East			**NL Central**			**NL West**		
Team	**Wins**	**Losses**	**Team**	**Wins**	**Losses**	**Team**	**Wins**	**Losses**
Mets	91*		Cubs	92*		Padres	88*	
Braves	85		Astros	84		Dodgers	86*	
Phillies	82		Brewers	84		Rockies	78	
Nationals	75		Cardinals	78		Dbacks	78	
Marlins	68		Reds	75		Giants	75	
			Pirates	71				

Table 8.5: 2008 MLB Season Predictions

With Alex Rodriguez in their everyday starting lineup, the 2008 Yankees are projected to win 95 games and win their division. An important question to ask however is how the team would perform without Rodriguez and if his absence could be financially profitable for the club.

Before Rodriguez finalized his deal with the Yankees, there were rumors that he might sign with a handful of other teams. Because few teams in MLB can actually afford a player of Rodriguez's ability, one of the more likely contenders for Rodriguez's talents was the large market Los Angeles Dodgers. Although it is hard to say how each team's lineups would have looked if Rodriguez did sign with the Dodgers, let us assume that the two teams just traded their respective third basemen along with their contracts. For the Dodgers, this would mean giving up Nomar Garciaparra, a previous all-star caliber short stop who now is one of the more average starting third basemen in MLB. Table () shows the predicted season results given by the simulator had such a trade been made.

American League 2008 Predictions – Garciaparra to Yankees								
AL East			**AL Central**			**AL West**		
Team	**Wins**	**Losses**	**Team**	**Wins**	**Losses**	**Team**	**Wins**	**Losses**
Boston*	93		Tigers	89*		Angels	85*	
Yankees*	91		Indians	86		Rangers	80	
Blue Jays	84		White Sox	79		Athletics	80	
Rays	77		Twins	79		Mariners	74	
Orioles	75		Royals	72				

National League 2008 Predictions – Rodriguez to Dodgers								
NL East			**NL Central**			**NL West**		
Team	**Wins**	**Losses**	**Team**	**Wins**	**Losses**	**Team**	**Wins**	**Losses**
Mets	91*		Cubs	91*		Dodgers	90*	
Braves	85		Brewers	85		Padres	86*	
Phillies	82		Astros	83		Dbacks	77	
Nationals	75		Cardinals	78		Rockies	76	
Marlins	69		Reds	75		Giants	72	
			Pirates	71				

Table 8.6: 2008 MLB Predictions Under Hypothetical Rodriguez/Garciaparra Trade

Interestingly enough, the results predict a 4 game change in both teams' fortunes. For the Yankees who now have an average starting third baseman, their expected number of wins drops to 91 and they also surrender the title of AL East Division Champions. For the Dodgers, their expected win total increases to 90 games and also puts them in post season play as a division winner and not a wild card team. Outside of their divisions, however, the impact on other teams from this hypothesized trade is expectedly negligible.

Although the ability to simulate different lineups through entire seasons is useful to any front office decision maker, the true economic affect of such a trade is hard to imagine for any team. As discussed, the more wins a team can earn, the better, but the returns become more and more marginal after a certain threshold. Also, the more wins a team can collect the better their chances of making the postseason, but once the postseason begins, no guarantees can be made as to how far a team will advance. There are also a plethora of peripheral issues such as a player's marketability that are hard to quantify and could significantly affect team revenue outside of team performance. To seriously tackle such a risk-reward tradeoff scenario, a deep understanding and analysis of a team's financial statements would have to occur.

8.4 Conclusion

Because of it's discrete, start-and-stop nature with finite set of possible outcomes, the game of baseball lends itself well to study via Markov chain simulations. Although this framework for baseball simulation has been widely discussed in academic literature for decades, actual and realistic implementation of this model has been sparse due to former time prohibitive computational capabilities, as well as the lack of availability to modern sabermetric baseball statistics.

In this study, teams have been broken down to their true component parts - individual players - and various estimates were used to predict any given player's current level of ability while also adjusting for an assortment of situational affects. The role and informational value of batter-pitcher matchup data was given particular attention through use of a hierarchical beta-binomial model and little to no interaction affect was found to exist between the two sides.

The accuracy of this player based approach was measured by profitability of wagers versus the daily betting lines offered on individual games throughout the 2007 Major League Baseball Season. Although the predicted run totals for each game were not able to outperform the market hypotheses, money line wagering did prove to be quite profitable across all input and prediction methods. Besides its capability to be utilized as a tool in sports betting, other natural applications for this player based approach to baseball simulation were shown to include the cost-effectiveness of potential free-agent signings for major league teams, optimal batting lineup orderings, strategic in-game decision making, and a benchmark for teams, players, agents, and Major League Baseball in salary arbitration hearings. And although this approach to predicting game outcomes and evaluating player worth does appear to be quite robust and promising, a better estimate of its value and accuracy can only be assessed as future seasons begin to be played out in the present.

REFERENCES

[1] Alamar, B., Ma, J., Desjardins, G., Lucas, R. (2006). Who Controls the Plate? Isolating the Pitcher/Batter Subgame. *Journal of Quantitative Analysis in Sports.* **2**(3), Article 4.

[2] Albert, Jim. (2007). "Hitting in the Pinch." *Statistical Thinking in Sports.* Ed. Jim Albert, Ruud Koning. CRC Press. 111 – 133.

[3] Albert, James (2006). Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-Time. *Journal of Quantitative Analysis in Sports.* **2**(1), Article 2.

[4] Albert, Jim (2002). Smoothing Career Trajectories of Hitters. *By The Numbers.* **12**(3), 9-20.

[5] Albert, J. (2001). Using Play-by-Play Baseball Data to Develop a Better Measure of Batting Performance. Techincal report, Department of Mathematics and Statistics, Bowling Green State University.

[6] Albert, Jim. (1994). Exploring Baseball Hitting Data: What About Those Breakdown Statistics? *Journal of the American Statistical Association.* **89**(427), 1066-1074.

[7] Albright, Christian. (1993). A Statistical Analysis of Hitting Streaks in Baseball. *Journal of the American Statistical Association.* **88**(424), 1175-1183.

[8] Allen, Erik. Using Bayesian Markov Chain Monte Carlo to Study the Informational Value of Early Season Baseball Statistics. (Forthcoming).

[9] Barry, D., Hartigan, J. (1993). Choice Models for Predicting Divisional Winners in Major League Baseball. *Journal of the American Statistical Association.* **88**(423), 766-774.

[10] Bell, Robert M., Thomas M. Cover. (1980). Competitive optimality of logarithmic Investment. *Mathematics of Operations Research*. **5**(2), 161-166.

[11] Bellman, R. (1976). "Dynamic Programming and Markovian Decision Processes, with Application to Baseball." *Management Science in Sports*. Ed. Robert E. Machol, Shaul P. Ladany, Donald G. Morrison. North-Holland Publishing Company. 77-85.

[12] Bennett, J., Flueck, J. (1983). An Evaluation of Major League Baseball Offensive Performance Models. *The American Statistician.* **37**(1), 76-82.

[13] Berry, S. (2006). Budgets and Baseball. *Chance.* **19**(1), 57-60.

[14] Berry, S. (2000). Modeling Offensive Ability in Baseball. *Chance*. **13**(4), 52-57.

[15] Birnbaum, Phil. (2005). Clutch Hitting and the Cramer Test. *By The Numbers.* **15**(1), 7-13.

[16] Bissinger, B. (2005). *Three Nights in August: Strategy, Heartbreak and Joy Inside the Mind of a Manager.* Mariner Books.

[17] Breiman, Hakansson, Thorp. (1975). *Stochastic optimization models in finance.* Academic Press: New York

[18] Breiman, L. (1961). Optimal gambling systems for favorable games. *Fourth Berkeley Symposium of Probability and Statistics.* 65–78.

[19] Bukiet, B., Harold, E. (1997). A Markov Chain Approach to Baseball. *Operations Research.* **45**(1), 14-23.

[20] Cook, E. (1976). "An Analysis of Baseball as a Game of Chance by the Monte Carlo Method." *Management Science in Sports*. Ed. Robert E. Machol, Shaul P. Ladany, Donald G. Morrison. North-Holland Publishing Company. 50-54.

[21] Cover, T., Keilers, C. (1977). An Offensive Earned-Run Average for Baseball. *Operations Research.* **25**(5), 729-740.

[22] Efron, B., Morris, C. (1977). Stein's paradox in statistics. *Scientific American.* **236**(5), 199-127.

[23] Ethier, S., Tavare, S. (1983). The proportional bettor's return on investment. *Journal of Applied Probability*. **20**, 563-573.

[24] Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, *3rd Edition*. Wiley: New York

[25] Fox, D. (2005). Tony LaRussa and the search for significance. `www.hardballtimes.com/main/article/tony-larussa-and-the-search-for-significance.`

[26] Gelman, A., Carlin, J., Stern, H., Rubin, D. (2003). *Bayesian Data Analysis* (2nd ed.). Boca Raton: CRC Press/Chapman & Hall.

[27] Griffin, P. (1984). Different measures of win rate for optimal proportional betting. *Management Science*. **30**(12), 1540-1502.

[28] Hanrahan, Tom (2001). Does Good Hitting Beat Good Pitching? *By The Numbers.* **11**(3), 13-17.

[29] Howard, R. (1976). "Baseball a la Russe." *Management Science in Sports*. Ed. Robert E. Machol, Shaul P. Ladany, Donald G. Morrison. North-Holland Publishing Company. 89-93.

[30] Insley, R., Mok, L., Swartz, T. (2004). Issues Related to Sports Gambling. *Aust. N.Z. J. Stat.* **46**(2), 219-232.

[31] James, B. (2006). *The Bill James Handbook 2007*. Skokie, IL.: ACTA Sports.

[32] James, B. (2002). *Win Shares.* Chicago, Illinois: STATS Publishing.

[33] James, B. (1988). *The Bill James Historical Baseball Abstract*. New York: Villard Books.

[34] James, B. (1987). *The Bill James Baseball Abstract*. New York: Villard Books.

[35] James, B. (1981). *The Bill James Baseball Abstract.* New York: Ballantine Books.

[36] Johnson, D., Graybill, F. (1972). Estimation of σ^2 in a Two-Way Classification Model with Interaction. *Journal of the American Statistical Association*. **67**(338), 388-394.

[37] Kaplan, David (2006). A Variance Decomposition of Individual Offensive Baseball Performance. *Journal of Quantitative Analysis in Sports.* **2**(3), Article 2.

[38] Kelly, J. (1956). A new interpretation of information rate. *Bell System Tech. Journal*, **35,** 917– 926 .

[39] Lehman, E.L., Casella, G. (2003). *Theory of Point Estimation* (2^{nd} ed.) New York: Springer

[40] Levitt, Dan (1999). Hits and Baserunner Advancement. *By The Numbers*. **9**(3), 20-22.

[41] Lindsay, George (1963). An Investigation of Strategies in Baseball. *Operations Research.* **11**(4), 477-501.

[42] Lindsey, George (1961). The Progress of the Score During a Baseball Game. *Journal of the American Statistical Association.* **56**(295), 703-728.

[43] MaClean, L., Ziemba, T., Blazenko, B. (1992). Growth versus security in dynamic investment analysis. *Management Science*. **38**(11), 1562-1585.

[44] McCracken, V. (2001). Pitching and Defense: How much control do hurlers have? http://www.baseballprospectus.com/article.php?articleid=878

[45] Morrison, D., Kalwani, M. (1993). The Best NFL Field Goal Kickers: Are They Lucky or Good? *Chance*. **6**(3), 30-37.

[46] Rosner, B., Mosteller, F., Youtz, C. (1996). Modeling Pitcher Performance and the Distribution of Runs per Inning in Major League Baseball. *The American Statistician.* **50**(4), 352-360.

[47] Schall, T., Smith, G. (2000). Do Baseball Players Regress Toward the Mean? *The American Statistician*. **54**(4), 231-235.

[48] Sela, R., Simonoff, J. (2007). "Does Momentum Exist in a Baseball Game?" *Statistical Thinking in Sports*. Ed. Jim Albert, Ruud Koning. CRC Press. 135 – 151.

[49] Skoog, G. (1987). Measuring runs created: the value added approach. In James, B. *The Bill James Baseball Abstract 1987*. New York: Ballantine Books.

[50] Stefani, Ray. (2007). "Measurement and Interpretation of Home Advantage." *Statistical Thinking in Sports*. Ed. Jim Albert, Ruud Koning. CRC Press. 203 – 215.

[51] Stern, H., Sugano, A. (2007). "Inferences about Batter-Pitcher Matchups in Baseball from Small Samples." *Statistical Thinking in Sports.* Ed. Jim Albert, Ruud Koning. CRC Press. 153-165.

[52] Stern, H. (1994). A Brownian Motion Model for the Progress of Sports Scores. *Journal of the American Statistical Association.* **89**(427), 1128-1134.

[53] Tango, Tom, and Mitchel G. Lichtman (2006). *The Book: Playing the Percentages in Baseball.* TMA Press.

[54] Thorn, J., Palmer, P. (1985). *The Hidden Game of Baseball.* New York: Doubleday.

[55] Thorp, Edward (1997). The Kelly Criterion in Blackjack, Sports Betting, and the Stock Market. *10th International Conference on Gambling and Risk Taking.*

[56] Thorp, E. (1969). Optimal gambling systems for favorable games. *Review of the International Statistical Institute.* **37**(3).

[57] Trueman, R. (1976). "Analysis of Baseball as a Markov Process." *Management Science in Sports*. Ed. Robert E. Machol, Shaul P. Ladany, Donald G. Morrison. North-Holland Publishing Company. 69-76.