Chapman & Hall/CRC

Handbooks of Modern Statistical Methods

Handbook of Statistical Methods and Analyses in Sports

Edited by

Jim Albert
Mark E. Glickman
Tim B. Swartz
Ruud H. Koning

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Handbook of Statistical Methods and Analyses in Sports

Chapman & Hall/CRC Handbooks of Modern Statistical Methods

Series Editor

Garrett Fitzmaurice
Department of Biostatistics
Harvard School of Public Health
Boston, MA, U.S.A.

Aims and Scope

The objective of the series is to provide high-quality volumes covering the state-of-the-art in the theory and applications of statistical methodology. The books in the series are thoroughly edited and present comprehensive, coherent, and unified summaries of specific methodological topics from statistics. The chapters are written by the leading researchers in the field, and present a good balance of theory and application through a synthesis of the key methodological developments and examples and case studies using real data.

The scope of the series is wide, covering topics of statistical methodology that are well developed and find application in a range of scientific disciplines. The volumes are primarily of interest to researchers and graduate students from statistics and biostatistics, but also appeal to scientists from fields where the methodology is applied to real problems, including medical research, epidemiology and public health, engineering, biological science, environmental science, and the social sciences.

Published Titles

Handbook of Mixed Membership Models and Their Applications
Edited by Edoardo M. Airoldi, David M. Blei,
Elena A. Erosheva, and Stephen E. Fienberg

Handbook of Statistical Methods and Analyses in Sports
Edited by Jim Albert, Mark E. Glickman,
Tim B. Swartz, and Ruud H. Koning

Handbook of Markov Chain Monte Carlo
Edited by Steve Brooks, Andrew Gelman,
Galin L. Jones, and Xiao-Li Meng

Handbook of Big Data
Edited by Peter Bühlmann, Petros Drineas,
Michael Kane, and Mark van der Laan

Published Titles Continued

Handbook of Discrete-Valued Time Series
Edited by Richard A. Davis, Scott H. Holan, Robert Lund, and Nalini Ravishanker

Handbook of Design and Analysis of Experiments
Edited by Angela Dean, Max Morris, John Stufken, and Derek Bingham

Longitudinal Data Analysis
Edited by Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs

Handbook of Spatial Statistics
Edited by Alan E. Gelfand, Peter J. Diggle, Montserrat Fuentes, and Peter Guttorp

Handbook of Cluster Analysis
Edited by Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci

Handbook of Survival Analysis
Edited by John P. Klein, Hans C. van Houwelingen, Joseph G. Ibrahim, and Thomas H. Scheike

Handbook of Spatial Epidemiology
Edited by Andrew B. Lawson, Sudipto Banerjee, Robert P. Haining, and María Dolores Ugarte

Handbook of Missing Data Methodology
Edited by Geert Molenberghs, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke

Handbook of Neuroimaging Data Analysis
Edited by Hernando Ombao, Martin Lindquist, Wesley Thompson, and John Aston

Chapman & Hall/CRC

Handbooks of Modern Statistical Methods

Handbook of Statistical Methods and Analyses in Sports

Edited by

Jim Albert
Bowling Green State University, Ohio, USA

Mark E. Glickman
Harvard University, Cambridge, Massachusetts, USA

Tim B. Swartz
Simon Fraser University, Burnaby, British Columbia, Canada

Ruud H. Koning
University of Groningen, The Netherlands

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20161109

International Standard Book Number-13: 978-1-4987-3736-4 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

A strong relationship has always existed between sports and the statistics that are used to measure player and team performance. Many interesting questions about sports have led to serious research in sports statistics. Comprehensive surveys of statistics in sports research have been provided by books such as *Management Science in Sports* (1976), *Optimal Strategies in Sports* (1977), and *Statistics in Sport* (1998). The American Statistical Association created a section on Statistics in Sports in 1992, the International Statistical Institute created a Sports Statistics Committee in 1993, and the journal *Chance* has devoted a regular column to sports statistics.

In the approximate 20 years since the publication of *Statistics in Sport*, there has been a remarkable change in both the accumulation of sports data and the opportunities to address sports questions using statistical methods. Researchers and sports professionals have seen a recent explosion in the proliferation of data collected on a variety of aspects of sports. Motion-tracking technology has permitted the accumulation of detailed information on player-level dynamics, and large data archives for sports have become much more accessible worldwide. For example, a baseball researcher has opportunities to explore season data by Lahman database (`www.seanlahman.com`), game and play data using Retrosheet (`www.retrosheet.org`), and pitch-by-pitch data through the PitchFX system (`www.sportvision.com/baseball/pitchfx`). This accumulation of data has led to the development of new statistical methodologies. As an example, the traditional measurement of fielding in baseball is the fielding percentage, the fraction of plays by a particular fielder that is successful. With the development of new tracking systems, one can now measure the movement of a fielder toward a ball that is hit in his direction and obtain a much better measure of fielding performance that accounts for the range of a player.

The opportunities for statistical research in sports have correspondingly grown immensely, as reflected in new dedicated journals to the subject area. Two examples are the *Journal of Quantitative Analysis in Sports* founded in 2006 and the *Journal of Sports Analytics* founded in 2014. In addition, meetings focusing on statistics in sports such as MathSport International and the New England Symposium on Statistics in Sports are regularly scheduled events. Also, opportunities exist to present research on statistics in sports at industry professional meetings such as the MIT Sloan Sports Analytics Conference and the SABR Analytics Conference.

Volumes such as *Statistical Thinking in Sports* (2007, ed. Albert and Koning) and *Anthology of Statistics in Sports* (2005, ed. Albert, Bennett, and Cochran) consist of general arrays of statistical articles but are not designed to provide the reader with a complete survey of the state-of-the-art methods in statistics in sports. The general aim of this handbook is to provide a basic reference for statistical researchers and students with an interest in sports applications to learn about the fundamental background, problems, and ongoing challenges in statistical methods in sports. The chapters in this book provide both overviews of statistical methods in sports and in-depth treatment of critical problems and challenges confronting statistical research in sports. This handbook intends to provide the reader the necessary background to conduct serious statistical analyses for sports applications and to appreciate scholarly work in this expanding area.

This handbook should be of interest for three types of readers. First, the handbook can serve as the basis for a graduate course or seminar in statistical methods in sports. Using the

handbook in this fashion can take advantage of connections between the methods typically used in a sports context with the methods that graduate students are learning in theoretical coursework. Second, the handbook can serve as a reference for statistical practitioners in professional sports who may not be aware of the breadth of statistical issues and problems in their area, or who may simply want a refresher in the problem areas they are likely to encounter. Finally, the handbook can provide statistical researchers who are interested in delving more into sports applications the requisite background to produce sound scholarly work that is set in a proper context. The handbook is organized by major sport (baseball, American football, basketball, hockey, and soccer) followed by a section on other sports.

The four chapters on baseball provide a general description of measures of player performance and situational and streakiness effects. The chapter by Ben Baumer and Pamela Badian-Pessot describes the wide range of measures proposed for evaluating batters and base runners. Carson Sievert and Brian Mills discuss, in their chapter, traditional measures of pitching performance and explore the opportunities for further insight about pitching using pitch-level data from the PITCHf/x system. Measuring fielding performance has been one of the more challenging problems in baseball analysis. The chapter by Mitchel Lichtman describes a plethora of fielding measures that have been proposed and the use of modern technology systems such as Statcast and Fieldf/x that can construct improved defensive metrics. In sports, fans are fascinated with streaky and "clutch" performances of players and teams, and Jim Albert, in his chapter, describes the use of statistical models to detect and estimate the size of situational and streaky effects.

The topics of the American football chapters address the issues of evaluating college talent, measuring player and team abilities, and decision-making within a game. The section on football starts with a chapter by Mark Glickman and Hal Stern describing methods for estimating NFL team abilities based on game outcomes. This is followed by a chapter describing the methods of evaluating NFL quarterbacks and placekickers by Drew Pasteur and John David. The next chapter by Julian Wolson, Vittorio Addona, and Rob Schmicker is devoted to forecasting the success of NFL players based on college performance. The final chapter in this section by Keith Goldner presents a discussion of quantitative methods for making optimal strategic decisions within a football game.

The four chapters on basketball analytics cover a range of topics about player abilities and game outcomes. The chapter by James Lackritz describes the ongoing controversy surrounding streak shooting and methods for detecting the hot hand in NBA basketball. This is followed by a chapter by Jeremias Engelmann on the history and current usage of plus/minus methods for evaluating player contributions from possession-based data. Brian Skinner and Matthew Goldman present the basics of optimal strategy in basketball in their chapter, with a focus on optimizing when players should take shots, at what point in the shot clock a team should take a shot, and under what circumstances teams should try high-risk tactics. Finally, the basketball section is concluded by a chapter by Luke Bornn, Daniel Cervone, Alexander Franks, and Andrew Miller that explores the current state of the art of analyzing player tracking data to measure both offensive and defensive player abilities.

The next group of chapters concerns hockey analytics. A chapter by Andrew Thomas develops the structure for an NHL match simulator using Poisson/Exponential models. The approach is comprehensive in that it takes into account various game situations (e.g., penalties, score, etc.) that affect the play of the game. In the next chapter, Bobby Gramacy, Matt Taddy, and Sen Tian use regularized logistic regression to assess player performance. Their approach may be seen as a generalization and an improvement of traditional plus/minus methods. A chapter by Michael Schuckers surveys the statistics used

to evaluate goaltending. Some of the statistics take into account the detailed aspects of goaltending, including the type of shots and the location of shots. In the final hockey chapter, Peter Tingling looks at drafting with a specific focus on the nuances of the NHL draft.

Soccer is an example of a low-scoring sport. Compared to, say, baseball, relatively few performance measures are available, and for that reason, focus has been on models for outcomes and the effect of interventions on these outcomes. The chapter by Phil Scarf and Jose Rangel Sellitti discusses recent developments in the literature on score modeling, focusing on the dependence between the number of goals scored by the home team and the number scored by the away team. At a slightly more abstract level, information about scores results in information about team quality. Ruud Koning, in his chapter, discusses a range of models that have been proposed to measure team quality. At a more disaggregate level, the chapter by Ian McHale and Samuel Relton describes the analysis of individual player ratings using data that have become available only recently. The soccer section concludes with chapters on intervention issues in the quantitative analysis of soccer. The chapter by Martin van Tuijl discusses whether dismissing a coach midseason results in better performance. Another issue explored by the chapter by Rob Simmons is whether referees are biased or impartial judges of the game, a topic that is relevant to other sports as well.

The final section of the handbook contains three contributions related to other sports. The first chapter by Mark Broadie and Bill Hurley investigates the two major research areas in golf. They look at the use of detailed ShotLink data in golf analytics and the age-old problem of handicapping in golf. In the second chapter, Tim Swartz surveys statistical research in cricket, the second most popular sport in the world. One take-away from this chapter is that the game has been underexplored and that opportunities exist in cricket analytics. The final chapter by Elmer Sterken discusses the issue whether inequality in performance in Olympic sports between top athletes increases or decreases over time. As discussed in this chapter, a general observation is that the performance of athletes tends to improve, but once a sport has reached maturity, improvement can level off.

We thank the chapter writers for their important contributions and our many colleagues who have inspired us to create a handbook that we hope sets a standard for researchers and practitioners in statistics in sports.

Jim Albert
Mark E. Glickman
Tim B. Swartz
Ruud H. Koning

Contributors

Vittorio Addona
Department of Mathematics, Statistics, and Computer Science
Macalester College
Saint Paul, Minnesota

Jim Albert
Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, Ohio

Pamela Badian-Pessot
School of Operations Research and Information Engineering
Cornell University
Ithaca, New York

Benjamin S. Baumer
Program in Statistical & Data Sciences
Smith College
Northampton, Massachusetts

Lucas M. Besters
Department of Economics
Tilburg University
Tilburg, the Netherlands

Luke Bornn
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, British Columbia, Canada

Mark Broadie
Graduate School of Business
Columbia University
New York, New York

Babatunde Buraimo
Department of Economics, Finance, and Accounting
University of Liverpool Management School
Liverpool, United Kingdom

Daniel Cervone
Center for Data Science
New York University
New York, New York

John A. David
Department of Applied Mathematics
Virginia Military Institute
Lexington, Virginia

Jeremias Engelmann
Consultant
ESPN.com
Heidelberg, Germany

Alexander Franks
Department of Statistics
University of Washington
Seattle, Washington

Mark E. Glickman
Department of Statistics
Harvard University
Cambridge, Massachusetts

Matthew Goldman
Microsoft Research
Redmond, Washington

Keith Goldner
numberFire
New York, New York

Robert B. Gramacy
Department of Statistics
Virginia Tech
Blacksburg, Virginia

William J. Hurley
Department of Mathematics and Computer Science
Royal Military College of Canada
Kingston, Ontario, Canada

Ruud H. Koning
Department of Economics, Econometrics and Finance
University of Groningen
Groningen, the Netherlands

James Lackritz
MIS Department
College of Business Administration
San Diego State University
San Diego, California

Mitchel Lichtman
Baseball Analyst
Canandaigua, New York

Ian G. McHale
Centre for Sports Business
Salford Business School
University of Salford
Manchester, United Kingdom

Andrew Miller
Department of Computer Science
Harvard University
Cambridge, Massachusetts

Brian M. Mills
Department of Tourism, Recreation, and Sport Management
University of Florida
Gainesville, Florida

R. Drew Pasteur
Department of Mathematics and Computer Science
The College of Wooster
Wooster, Ohio

José Sellitti Rangel Jr.
Salford Business School
University of Salford
Manchester, United Kingdom

Samuel D. Relton
School of Mathematics
University of Manchester
Manchester, United Kingdom

Phil Scarf
Salford Business School
University of Salford
Manchester, United Kingdom

Michael E. Schuckers
Department of Mathematics, Computer Science, and Statistics
St. Lawrence University
Canton, New York

Robert Schmicker
Department of Biostatistics
University of Washington
Seattle, Washington

Dirk Semmelroth
Department of Management
University of Paderborn
Paderborn, Germany

Carson Sievert
Department of Statistics
Iowa State University
Ames, Iowa

Rob Simmons
Department of Economics
Lancaster University Management School
Lancaster, United Kingdom

Brian Skinner
Department of Physics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Elmer Sterken
Institute of Economics, Econometrics, and Finance
University of Groningen
Groningen, the Netherlands

Hal S. Stern
Department of Statistics
University of California, Irvine
Irvine, California

Tim B. Swartz
Department of Statistics and
Actuarial Science
Simon Fraser University
Burnaby, British Columbia, Canada

Matt Taddy
Microsoft Research New England
Cambridge, Massachusetts
and
Booth School of Business
University of Chicago
Chicago, Illinois

Andrew C. Thomas
Minnesota Wild
National Hockey League
Saint Paul, Minnesota

Sen Tian
Stern School of Business
New York University
New York, New York

Peter M. Tingling
Beedie School of Business
Simon Fraser University
Vancouver, British Columbia, Canada

Jan C. van Ours
Department of Applied Economics
Erasmus University Rotterdam
Rotterdam, the Netherlands
and
Department of Economics
University of Melbourne
Parkville, Australia

Martin A. van Tuijl
Department of Economics
Tilburg University
Tilburg, the Netherlands

Julian Wolfson
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, Minnesota

1

Evaluation of Batters and Base Runners

Benjamin S. Baumer and Pamela Badian-Pessot

CONTENTS

Since Henry Chadwick began publishing boxscores around the turn of the twentieth century (Schwarz, 2005a), people have been interested in evaluating the performance of baseball players. In particular, the offensive contributions made by position players—in the form of both batting and baserunning—are the most obviously varied and carefully studied contributions. In this chapter, we will catalog the most enduring sabermetric models for evaluating batters and base runners. Our approach is model centric, in that we will attempt to categorize metrics based on the type of model on which they are built. It should not be surprising that over time these models have become more sophisticated, both in terms of the model complexity and the rigor with which any parameters are estimated.

The fundamental challenge in evaluating batters and base runners is that run scoring in baseball is the result of interdependent actions among teammates and opponents. While separating the individual contributions of these team actions is considered easier in baseball than in many other sports, it is far from trivial. For example, consider an inning in which the first batter leads off with a walk, steals second base, advances to third base on a groundout, and then scores on a sacrifice fly. What is unambiguous is that the team scored one run. What is debatable is how much of that run is attributable to each player. Does the second batter make a positive contribution by advancing the runner, even though he made an out? How much credit did the first player accrue though baserunning? The tools developed in this chapter will help us answer these questions.

One common technique that will prime the search for models is

- To recognize that neither the runs scored statistic (R), which gives full credit for the team run to the player who crossed the plate, nor the runs batted in statistic (RBI), which gives full credit for the team run to the player who drove in the run, are reasonable ways to apportion the run
- To find a model for R that works well for teams
- To apply that model to individual players

Later in this chapter, we will show how this method can be used to evaluate the accuracy of offensive metrics.

In Section 1.1, we motivate this line of inquiry, outline some basic tools necessary to understand these models, and define our notation. In Section 1.2, we explore models that can be computed using data that are aggregated at the seasonal level. These are the simplest but most numerous and storied models for offensive performance. Models that require more detailed play-by-play data are explored in Section 1.3. Predictive models, including Bayesian models, are discussed in Section 1.4. We conclude the chapter in Section 1.5 by pointing toward some open problems.

1.1 Basic Tools

1.1.1 Overview

Several comprehensive assessments of offensive players have advanced the field of sabermetrics. Many have drawn inspiration from the work of Bill James interspersed in his books (James, 1986; James and Henzler, 2002). Perhaps the first book-length treatise was Cook (1964). This was followed 20 years later by Thorn and Palmer (1984), an important book that was reprinted in 2015 (Thorn and Palmer, 2015). The article-length analysis of Bennett and Flueck (1983) was greatly expanded by Albert and Bennett (2003) into what remains probably the best place to start reading about sabermetrics. The more recent book by Tango et al. (2007) not only focuses more on in-game strategy, but also includes some measures of offensive assessment. Methods for analyzing baseball data using the statistical computing environment R—as we do in this chapter—are explicated by Marchi and Albert (2013).

1.1.2 Expected Run Matrix

One of the fundamental tools in sabermetric analysis is the *expected run matrix*, which gives the expected number of runs scored in the remainder of an inning, given that the inning is currently in one of the 24 (*base*, *out*) states. Throughout this chapter, we will use the notation **R** to refer to this 8×3 matrix, and the notation **r** to refer to the corresponding vector of length 24. Specifically, the notation $\mathbf{R}_{23,2}$ indicates the value of the expected run matrix when runners are on second and third with two outs.* The complete expected run matrix for 2013 is presented in Table 1.1.

1.1.3 Notation

In this chapter, we denote the value of player i's batting and baserunning contributions as y_i. Typically, but not always, player value is measured in the units of runs. Since, as we discussed earlier, there is no way to *directly* measure the run value of these contributions for individual players, we consider y_i to be unknown. The goal of this chapter is to describe models for y_i in a coherent fashion. The estimates for y_i will be denoted by $\hat{y}_i$. Note that from the example at the beginning of this chapter, each of the three offensive players has a y_i—we just don't know what they are.

An example may make this clearer. It is an undisputed fact that the St. Louis Cardinals scored $y_{STL} = 798$ runs in 1987. This number represents the total batting and baserunning contributions of the entire team. However, we don't know how many of those 798 runs are attributable to each player. We know that Ozzie Smith scored 104 of those runs and drove in another 75, but neither of those numbers represents y_{Smith}. Furthermore, neither is

* There are several commonly used notations for referring to the configuration of base runners indicated by *base*. The most intuitive is a notation like `23` (or ×23), which indicates that there are runners on second and third. We will use this notation in the text of this chapter. However, this notation is very inconvenient computationally. In our computations, we use the following notation: imagine each of the three bases as a binary digit that can either be unoccupied (0) or occupied (1). Then the binary string 110 indicates that runners are on second and third. This has the decimal equivalent of 6. Thus, the notations `x23`, 110, and 6 all refer to having runners on second and third. We trust the reader will be able to keep this clear in context.

TABLE 1.1

The Expected Run Matrix for 2013

BaseCode\Outs	0	1	2
000	0.456	0.240	0.091
001	0.812	0.491	0.211
010	1.096	0.617	0.301
011	1.382	0.838	0.402
100	1.261	0.925	0.344
101	1.828	1.108	0.480
110	2.080	1.390	0.558
111	2.179	1.568	0.714

Source: MLBAM, GameDay files, http://gd2.mlb.com/components/game/mlb/, accessed July 1, 2016.

Note: The rows correspond to the configuration of the bases, while the columns correspond to the number of outs. The entry (110, 2) implies that based on 2013 data, about 0.558 runs are expected to be scored in an inning in which there are two outs and runners on second and third.

TABLE 1.2

Some Basic Statistics for Eight Notable National League Players in 1987

Player	Team	*PA*	*R*	*H*	*2B*	*3B*	*HR*	*RBI*	*BB*	*HBP*	*SB*	*CS*
Ozzie Smith	SLN	706	104	182	40	4	0	75	89	1	43	9
Dale Murphy	ATL	693	115	167	27	1	44	105	115	7	16	6
Tony Gwynn	SDN	680	119	218	36	13	7	54	82	3	56	12
Andre Dawson	CHN	662	90	178	24	2	49	137	32	7	11	3
Darryl Strawberry	NYN	640	108	151	32	5	39	104	97	7	36	12
Tim Raines	MON	627	123	175	34	8	18	68	90	4	50	5
Eric Davis	CIN	562	120	139	23	4	37	100	84	1	50	6
Jack Clark	SLN	558	93	120	23	1	35	106	136	0	1	2

Note: Dawson won the MVP Award, with Smith and Clark placing next in the voting. Gwynn's season is considered the best by WAR, followed by Davis, Murphy, and Raines.

a particularly good *estimate* of y_{Smith}. We will see later in this chapter that our best estimate of y_{Smith} is closer to 98.

For seasonal metrics, we will follow a list of eight prominent National League players from the 1987 season to illuminate the metrics discussed herein. The list of players, along with their unambiguous counting statistics, is shown in Table 1.2. It is worth noting that Andre Dawson of the Chicago Cubs—largely by virtue of his NL-leading 49 home runs—won the NL Most Valuable Player (MVP) Award, becoming the first player from a last-place team to do so.*

* In hindsight, Tony Gwynn of the San Diego Padres produced the highest Wins Above Replacement (*rWAR*), according to Baseball-Reference.com. We do not discuss WAR in this chapter since it involves the evaluation of pitching and fielding.

1.2 Models Based on Seasonal Data

It should not be surprising that the most storied and widely used models for offensive performance in baseball are linear models. That is, models in which some number of discrete batting or baserunning events are associated with a specific value, and each player's value is quantified by summing these values over the number of these events attributed to that player.

Formally, let Ω be a set of commonly recorded offensive and baserunning events. Ω includes things like singles ($1B$), doubles ($2B$), triples ($3B$), and home runs (HR), as well as things like stolen bases (SBs) and caught stealings (CSs). To build a linear model, we might choose p of the events in Ω that seem important and tally the number of occurrences of these p events for player i over some fixed time period—say, a season. This becomes the vector $x_i \in \mathbb{N}^p$. Note that each element of x_i is a nonnegative integer, since it represents a count.

Next, we could assign a value to each of these p events. These values are most often in the units of *runs*, but they could be measured in any unit. Our coefficient vector becomes $\beta \in \mathbb{R}^p$, where here we allow negative values, since events like strikeouts (SO) and caught stealings are clearly detrimental to a team's offense.

Finally, we consider $y_i = x_i^T \beta$. The scalar value y_i is in the same units as the elements of β. If we have n players, then we can extend this framework to the matrix equation $y = \mathbf{X}\beta$, where $y \in \mathbb{R}^n$ and $\mathbf{X}$ is an $n \times p$ matrix with the ith row being x_i. That is, each row in $\mathbf{X}$ represents a player, and each column a variable. This expression is *linear*—in what follows we catalog the most significant models for offensive player value of this form. In each case, the primary differences in these models are

1. Which events to consider? That is, which elements of Ω make up the columns of $\mathbf{X}$ and the elements of β? We denote this choice by $\omega \subseteq \Omega$.
2. How to choose (or estimate) the elements of β? We will see that many early models simply asserted values for β or chose them based on natural elements of the game. More recently, sabermetricians have employed techniques for optimizing β according to some criteria.

The remaining differences are choices about units and about converting y to a *rate*. That is, as presented earlier y is a sum of weighted counts. As the playing time of players varies, it is often more useful to think about a player's value in terms of his value *per game*, or *per plate appearance*. In this case, we let $z_i \in \mathbb{N}^q$ be a vector of counts for q events $\psi \subseteq \Omega$ that record player i's playing time, and $\alpha \in \mathbb{R}^q$ a vector of corresponding weights. Then $z_i^T \alpha$ is a scalar that assesses player i's playing time, and y_i/z_i is a rate that measures player i's contributions on a per unit of playing time basis. We will reinforce this notation throughout this chapter.

1.2.1 Batting Models

We begin by considering models that incorporate batting statistics only.

1.2.1.1 The Triple-Slash Models

The so-called "triple-slash" statistics for batting are batting average (AVG), on-base percentage (OBP), and slugging percentage (SLG). These are the most commonly cited statistics in

baseball and are ubiquitous on television broadcasts, the back of baseball cards, and web sites. We have also included on-base plus slugging (*OPS*) in this section, as it is the simple sum of *OBP* and *SLG*.

1.2.1.1.1 Batting Average

The most widely known offensive statistic is batting average, which dates back to the 1800s and Henry Chadwick (Schwarz, 2005a). Defined simply as hits over at-bats, the appeal of *AVG* is that it reflects the common sense perception that better offensive players will get hits in more of their opportunities.

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

In our framework, we have the trivial decomposition where $\psi = \{AB\}$, $\omega = \{H\}$, and $\beta = 1$. Since, by definition, $H \leq AB$, the units here are a proportion.

While batting average remains popular, researchers have reached consensus that it is a relatively poor measure of offensive batting prowess, for both descriptive and predictive uses (Baumer, 2008; Albert, 2016) (see Table 1.11).

1.2.1.1.2 On-Base Percentage

While *AVG* only counts when a player reaches base by getting a hit, there are other ways of reaching base that also contribute to run scoring—most commonly through walks and hit by pitches. On-base percentage improves upon *AVG* in two ways: by including walks and hit by pitches in the numerator and by using plate appearances—excluding sacrifice hits—as a measure of opportunities in the denominator. *OBP* is essentially the rate a player reaches base, or—perhaps more importantly—the proportion of the time he doesn't make an out.

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF} = \frac{H + BB + HBP}{PA - SH},$$

where

SF is sacrifice fly

SH is sacrifice hit—more commonly known as sacrifice bunts

Here, we have $\psi = \{PA, SH\}$, $\omega = \{H, BB, HBP\}$, $\alpha = \begin{pmatrix} 1 & -1 \end{pmatrix}^T$, and $\beta = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^T$.

1.2.1.1.3 Slugging Percentage

Slugging percentage is not a measure of the rate a player reaches base—it instead asks how far a player reaches, that is, how many bases he produces per at-bat.

$$\begin{aligned} SLG &= \frac{TB}{AB} = \frac{1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR}{AB} \\ &= \frac{H + 2B + 2 \cdot 3B + 3 \cdot HR}{AB} \\ &= AVG + \frac{2B + 2 \cdot 3B + 3 \cdot HR}{AB} \end{aligned}$$

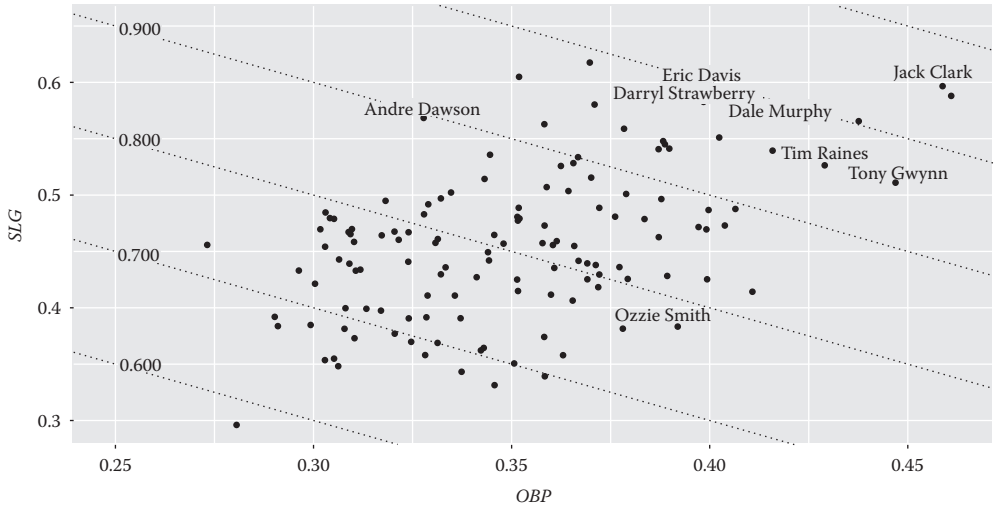

FIGURE 1.1
Scatterplot of *SLG* vs. *OBP* for all qualified (at least 502 plate appearances) NL batters in 1987. Dotted lines indicate *OPS* isometries.

In this manner, *SLG* differentiates between players who hit mostly singles and those who hit for power, unlike *AVG* and *OBP*. *SLG* weights hits by the number of total bases (*TB*s): one for singles, two for doubles, three for triples, and four for home runs. The term slugging percentage is itself a misnomer, in that it is the rate at which a player produces bases, not a percentage.

In our notation, $\omega = \{1B, 2B, 3B, HR\}$ and $\beta = \begin{pmatrix}1 & 2 & 3 & 4\end{pmatrix}^T$. The latter term in the final equation is known as *isolated power* (*ISO*) and accordingly measures a player's extra base power net of his batting average.

1.2.1.1.4 On-Base Plus Slugging

OBP improves upon *AVG* by including ways to reach base other than hits and *SLG* improves upon *AVG* by distinguishing between types of hits. On-base plus slugging (*OPS*) incorporates both of these ideas by simply adding these values together. Players with the highest *OPS* both reach base often and hit for power. While these skills are correlated, they are not the same. Figure 1.1 illustrates the relationship between *OBP* and *SLG* among qualified* NL players in 1987.

$$OPS = OBP + SLG$$

The triple-slash metrics for our eight notable players are shown in Table 1.3. We note that Gwynn's exceptionally high batting average does not carry the day according to *OPS*. Moreover, Dawson's low *OBP* and Smith's low *SLG* reveal the former's inability to draw walks and the latter's lack of power.

* To qualify for the batting title, a player needs 3.1 plate appearances per team game. For a full 162 game season, this translates to at least 502 plate appearances.

TABLE 1.3

Triple-Slash Statistics for Seven Notable National League Players in 1987

Player	Team	*AVG*	*OBP*	*SLG*	*OPS*
Jack Clark	SLN	0.286	0.459	0.597	1.055
Dale Murphy	ATL	0.295	0.417	0.580	0.997
Eric Davis	CIN	0.293	0.399	0.593	0.991
Darryl Strawberry	NYN	0.284	0.398	0.583	0.981
Tony Gwynn	SDN	0.370	0.447	0.511	0.958
Tim Raines	MON	0.330	0.429	0.526	0.955
Andre Dawson	CHN	0.287	0.328	0.568	0.896
Ozzie Smith	SLN	0.303	0.392	0.383	0.775

Since *OBP* is measured as a proportion, whereas *SLG* is an average, their units are incomparable, leading many to question whether the simple sum of the two measures is best. That is, if we generalize *OPS* and consider all models of the form

$$\hat{y}_i = a \cdot OBP_i + b \cdot SLG_i,$$

for constants a and b, what are the optimal values of a and b? In *Moneyball*, it is reported that the proper ratio a/b is closer to 3, as opposed to the ratio of 1 posited by *OPS* (Lewis, 2003). Wang (2006) estimates the value to be approximately 1.8 using trial and error—we confirmed this using data for all 1162 team seasons from 1960 to 2005 and estimating the optimal ratio using linear regression. The R^2 for the fit is 0.929, and the ratio of the coefficients is 1.781. While there is variability associated with this estimate, the conclusion of Wang (2006) seems to be in the right ballpark. That is, a modified *OPS* that weights *OBP* more heavily produces estimates that are "closer" to team runs scored than *OPS*.

1.2.1.2 Averaging Changes to the Expected Run Matrix

SLG posits the weights $\beta = (1, 2, 3, 4)$ as values for singles, doubles, triples, and home runs, respectively, without considering how many runs, on average, a hit is actually worth. That is, is a double really twice as valuable as a single? How many runs is a double worth, on average? Using data from the 1959 to 1960 season, Lindsey (1963) estimated this figure to be 0.817. His idea was to tabulate the change in run expectancy for each situation in which a double could be hit and then take the weighted average according to how often those situations occur. Mathematically, the average value of a double β_{double} is given by

$$\beta_{double} = \sum_{(base,out)} \delta_{double}|(base, out) \cdot freq(base, out),$$

where

δ_{double} is the change in run expectancy for a double
$freq(base, out)$ is the likelihood of being in each $(base, out)$ state

More generally, the change in run expectancy $\delta_{f,i}$ from state i to state f is

$$\delta_{f,i} = \mathbf{R}_f + r_i - \mathbf{R}_i, \tag{1.1}$$

where

$\mathbf{R}_i$ and $\mathbf{R}_f$ are the expected runs scored in the remainder of the inning before and after the play, respectively

r_i is the number of runs scored as a result of the play

This value $\delta_{f,i}$ arises in other contexts later in this chapter.

For example, Lindsey found an expected run value of $\mathbf{R}_{12,2} = 0.403$ runs when a batter hits with two outs and men on first and second, but that he will bat in this situation only 3.3% of the time. If this batter hits a double, Lindsey supposed that half the time both runners will score and the other half one runner will score and the other will be left on third. The run value of a double in this situation is then

$$\begin{aligned}\hat{\delta}_{double}|(12,2) &= \frac{1}{2}(\mathbf{R}_{23,2} + 1) + \frac{1}{2}(\mathbf{R}_{2,2} + 2) - \mathbf{R}_{12,2} \\ &= \frac{1}{2}(0.687 + 1) + \frac{1}{2}(0.297 + 2) - 0.403 \\ &= 1.589.\end{aligned}$$

Thus, with two men on base and two outs, the value of hitting a double is quite high—nearly 1.6 runs. However, this situation is relatively rare, occurring only 3.3% of the time. Much more likely is that the batter will come up with no one on and no one out—24.2% of plate appearances occur in this configuration. The value of a double here is simply

$$\hat{\delta}_{double}|(0,0) = (\mathbf{R}_{2,0} + 0) - \mathbf{R}_{0,0} = 1.194 - 0.461 = 0.733.$$

This is less than half of the value of the previous example, illustrating how the value of different events can change based on the situation. The *average* value of hitting a double $\hat{\beta}$ is the sum of the values of a double in all initial possible states, weighted by the frequency of each state.

Ultimately, Lindsey found the average value of a single to be 0.41 runs, and doubles, triples, and home runs to be worth 0.82, 1.06, and 1.42 runs, respectively, on average. With an eye toward measuring the value of hits relative to the average value of a single, Lindsey proposed altering the weights used for *SLG* from $1, 2, 3, 4$ to $1, 2, 2.5, 3.5$. Although this proposal was not adopted, it marks the beginning of using the expected run matrix to methodically determine optimal weights for different offensive plays. This idea has informed much of the subsequent analysis in all aspects of the game, as well as other sports, and is a truly enduring contribution.

Using the expected run matrix for the 2013 season shown in Table 1.1, we can use Lindsey's methodology to estimate the run values of any event. We show the results in Table 1.4. There are a number of interesting observations one can make based on these data—most of which were identified by Lindsey. These observations have informed sabermetric orthdoxy for decades. Among the most obvious are the following:

TABLE 1.4

Estimated Run Values for the 22 Most Common Events, 2013

Event	*N*	*Freq*	$\bar{r}$	$\hat{\beta}$
Home run	4,661	0.025	1.54	1.37
Triple	772	0.004	0.62	1.02
Double	8,185	0.044	0.40	0.75
Field error	1,516	0.008	0.20	0.49
Single	28,448	0.154	0.22	0.44
Hit by pitch	1,536	0.008	0.01	0.31
Walk	13,622	0.074	0.02	0.30
Intent walk	1,018	0.005	0.01	0.18
Sac fly	1,204	0.006	1.01	−0.00
Sac bunt	1,382	0.007	0.04	−0.10
Groundout	35,171	0.190	0.02	−0.20
Flyout	23,080	0.125	0.00	−0.23
Lineout	9,493	0.051	0.00	−0.23
Strikeout	36,573	0.197	0.00	−0.25
Popout	8,877	0.048	0.00	−0.26
Forceout	3,946	0.021	0.07	−0.31
Grounded into DP	3,731	0.020	0.03	−0.75

Source: MLBAM, GameDay files, http://gd2.mlb.com/components/game/mlb/, accessed July 1, 2016.

Note: The rightmost column shows the average change in run expectancy, which is a measure of the value of the play ($\hat{\beta}$), in runs. The column second from right ($\bar{r}$) shows the average number of runs that score on plays of each type.

- Walks are worth less than singles.
- A walk and a hit by pitch have the same run value, since they have the same result. However, an intentional walk is less valuable even though it too has the same result, because by design it is only employed in situations where adding the runner on first does not contribute as much to the run expectancy.
- A fielding error is worth about the same as a single, since it usually means that the batter reached first and the other runners moved up one or two bases.
- A sacrifice fly necessarily results in one run being scored, but because it also results in an out being recorded, it has a negligible effect in terms of the change in run expectancy.
- A sacrifice bunt is a marginally negative event for the offense.
- All other types of outs (e.g., groundouts, flyouts, popouts, lineouts, and strikeouts) are of approximately equal negative value.

1.2.1.2.1 Batting Runs (BRs)

Batting runs (Thorn and Palmer, 1984) use the same basic framework as Lindsey (1963) and consequently find nearly the same average run values for singles, doubles, triples, and home runs. However, Palmer expands on Lindsey's work by also including walks, hit by pitches, and the negative run value of an out. By Palmer's estimation, the value of an out changes yearly to account for changes in the league's run-scoring environment.

TABLE 1.5

Aggregated Linear Batting Models for Selected 1987 NL Players

Player	*PA*	*OPS*	*BR*	*wOBA*
Jack Clark	558	1.055	58.5	0.447
Eric Davis	562	0.991	43.9	0.421
Dale Murphy	693	0.997	57.6	0.417
Darryl Strawberry	640	0.981	48.6	0.415
Tony Gwynn	680	0.958	52.8	0.408
Tim Raines	627	0.955	46.9	0.405
Andre Dawson	662	0.896	26.9	0.382
Ozzie Smith	706	0.775	14.3	0.347

Note: We note that Clark leads by all three metrics.

BR measures the number of runs contributed beyond what an average player would have contributed while consuming the same number of outs. Accordingly, a league average player will have a BR of 0. The formula is

$$BR = 0.47 \cdot 1B + 0.85 \cdot 2B + 1.02 \cdot 3B + 1.40 \cdot HR + 0.33 \cdot (BB + HBP) - ABF \cdot (AB - H),$$

where *ABF* is calculated so that *BR* for the league is zero. A typical value of *ABF* is approximately 0.3. We show *BR* for selected players in Table 1.5.

1.2.1.2.2 Weighted On-Base Average (wOBA)

wOBA is another metric that uses changes to the expected run matrix to determine the value of different events (Tango et al., 2007). However, *wOBA* is designed so that the league average *wOBA* is equal to the league average *OBP*, putting the former numbers on the familiar scale of the latter. It is necessary to inflate the run values by about 15% to make this happen. All weights are recalculated annually to adjust for changes to the run-scoring environment. The following equation defines *wOBA* for the 2013 season:

$$wOBA = \frac{0.690 \cdot uBB + 0.722 \cdot HBP + 0.888 \cdot 1B + 1.271 \cdot 2B + 1.616 \cdot 3B + 2.101 \cdot HR}{AB + BB - IBB + SF + HBP}$$

wOBA is presented as an improvement over *OPS*, as it incorporates both elements of reaching base and the value of the event on which the player reaches. Additionally, because *wOBA* uses the same scale as *OBP*, many baseball fans already have a reference for what values are poor, good, or excellent. Weighted Runs Above Average (*wRAA*)—which is used for the hitting component of Fangraphs' WAR (*fWAR*)—uses *wOBA* to calculate the number of runs a player contributes compared to an average player in the same number of plate appearances. In Figure 1.2, we illustrate how *wOBA* compares to *OBP* for all NL players in 1987. Note that the distribution of *wOBA* is similar to that of *OBP*.

In Table 1.5, we compare *OPS*, Batting Runs, and *wOBA* for our selected 1987 NL players. While there is widespread agreement, there are also discrepancies. Note the nearly 14-run discrepancy between Eric Davis and Dale Murphy in Batting Runs, despite their being nearly identical in *OPS* and *wOBA*. This is the result of Batting Runs being a counting stat, while the others are rate stats.

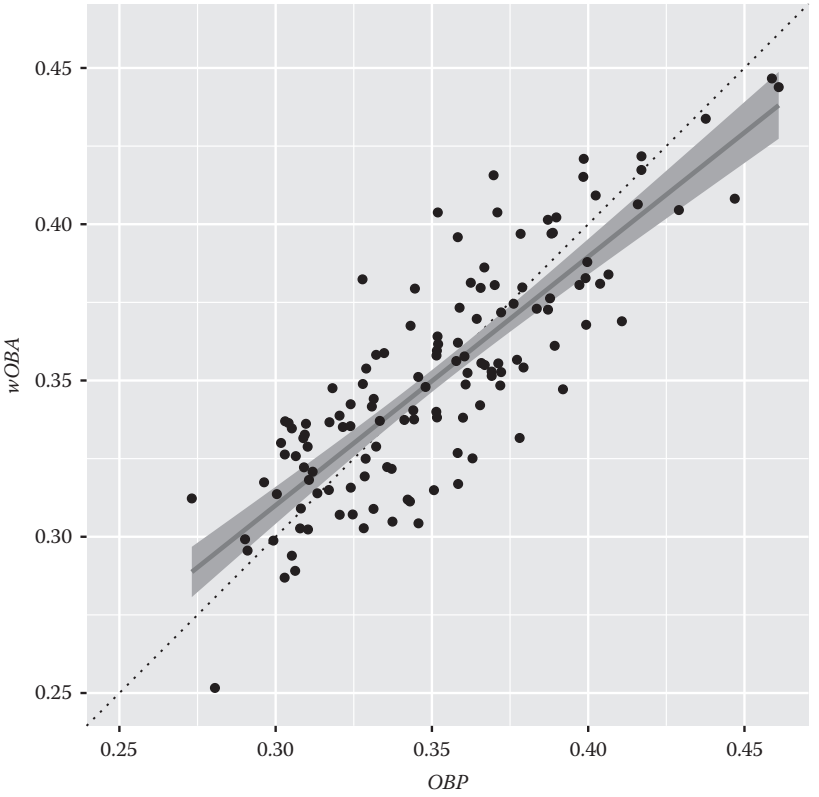

FIGURE 1.2
Comparison of *wOBA* to *OBP*. A regression line with error ranges has been added. Since *OBP* and *wOBA* are on the same scale, the dotted diagonal line allows a direct comparison of the values.

1.2.2 Baserunning Models

In this section, we discuss measures for baserunning alone. While thus far all the models covered use only batting statistics, a player's total offensive performance is clearly a combination of his hitting and baserunning skills. Both metrics covered in this section are offshoots of metrics from the previous section.

1.2.2.1 Basestealing Runs

Basestealing Runs are the basestealing counterpart of *batting runs*. Palmer used the same methods as in *BR* to find the average run value of stolen bases and caught stealing.

$$Basestealing\ Runs = 0.22 \cdot SB - 0.38 \cdot CS$$

1.2.2.2 Weighted Stolen Base Runs

wSB is the baserunning counterpart of *wOBA*. The values of stolen bases and caught stealings are found in the same way as *wOBA* and then compared to the league average. *wSB* measures runs created (RC) above (or below) the league average in the same number of

opportunities. This statistic is used as part of the baserunning component of Fangraphs' *fWAR* (Fangraphs Staff, 2015b).

$$wSB = \beta_{SB} \cdot SB + \beta_{CS} \cdot CS - lgwSB \cdot (1B + BB + HBP - IBB),$$

where
$\beta_{SB} = 0.2$
$\beta_{CS} = -(2 \cdot RunsPerOut + 0.075)$, and *RunsPerOut* is the total number of runs scored by all teams divided by the outs in a season

The constant

$$lgwSB = \frac{\beta_{SB} \cdot SB + \beta_{CS} \cdot CS}{1B + BB + HBP - IBB}$$

provides a scaling factor. Its denominator is an approximation of the number of times that the player was on first base with second base potentially open. Thus, *wSB* measures the weighted run value of an individual player's contributions on the basepaths, above what a league average runner would have contributed the same number of times on first.

Clearly, there are shortcomings in *wSB* in that it only measures stolen base contributions and ignores contributions from "taking the extra base" (i.e., advancing from first to third on a single). These unmeasured contributions are known to be significant (Click, 2005; Fox, 2005). Second, the approximation for the number of times on first base is not a great measure of opportunities, since it ignores things like reaching first on a fielder's choice.

Table 1.6 displays baserunning metrics for our selected 1987 NL players.

1.2.3 Combined Models

We now turn our attention to measures that combine both batting and baserunning.

TABLE 1.6

Basestealing Runs for Selected 1987 NL Players

Player	*PA*	*SB*	*CS*	*SB%*	*Base Runs*	*wSB*
Tim Raines	627	50	5	0.91	9.10	7.57
Eric Davis	562	50	6	0.89	8.72	7.19
Tony Gwynn	680	56	12	0.82	7.76	5.72
Ozzie Smith	706	43	9	0.83	6.04	4.39
Darryl Strawberry	640	36	12	0.75	3.36	1.82
Andre Dawson	662	11	3	0.79	1.28	0.70
Dale Murphy	693	16	6	0.73	1.24	0.33
Jack Clark	558	1	2	0.33	−0.54	−0.96

Note: We note that the plodding Clark accrued a negative value by being caught stealing twice against only one successful stolen base.

1.2.3.1 Total Average

Total Average (*TA*) is most similar to the intuitive statistics covered at the beginning of this chapter. Indeed, Boswell (1982) introduced it as a revision to slugging percentage. As opposed to *SLG*—which is the rate of total bases per at-bat—*TA* is the total number of bases attributed to a player, either through hitting, walking, or baserunning, divided by total outs. The change from at-bats to outs in the denominator is important because unlike at-bats, outs are generally constant at 27 per game. This suggests that bases per out should be more closely related to runs per game than bases per at-bat.

$$\begin{aligned} TA &= \frac{TB + BB + HBP + SB}{AB - H + CS + GIDP} \\ &= \frac{1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR + BB + HBP + SB}{AB - H + CS + GIDP} \end{aligned}$$

1.2.3.2 Offensive Performance Average

Offensive Performance Average uses the weights computed by Lindsey (1963) for hits while additionally including walks, hit by pitches, and stolen bases (Pankin, 1978). Recall that the weights are relative to the value of single, that is, a stolen base is about half as valuable as single.

$$OPA = \frac{1B + 2 \cdot 2B + 2.5 \cdot 3B + 3.5 \cdot HR + 0.8 \cdot (BB + HBP) + 0.5 \cdot SB}{AB + BB + HBP}$$

1.2.3.3 Estimated Runs Produced

A response to Bill James' runs created (see Section 1.2.5), Johnson's *Estimated Runs Produced* (*ERP*) combines an intuitive model with weights found through trial and error to predict runs scored (Johnson and James, 1985). *ERP* aims to track the positive contributions of a player using TB, BB, HBP, H, SB and the negative contributions, that is, the number of outs made ($AB - H, CS, GIDP$) by a player. The weights were determined with the value of each play in mind but were estimated through trial and error.

$$ERP = (2 \cdot (TB + BB + HBP) + H + SB - (0.605 \cdot (AB + CS + GIDP - H))) \cdot 0.16$$

After a bit of algebra,

$$\begin{aligned} ERP = {} & 0.48 \cdot 1B + 0.8 \cdot 2B + 1.12 \cdot 3B + 1.44 \cdot HR + 0.32 \cdot BB + 0.32 \cdot HBP \\ & + 0.16 \cdot SB - 0.0968 \cdot (AB - H) - 0.0968 \cdot CS - 0.0968 \cdot GIDP. \end{aligned}$$

1.2.3.4 eXtrapolated Runs

Finally, in eXtrapolated Runs (*XR*), Furtado (1999) used regression analysis to determine the best-fit coefficients for a series of batting and baserunning events.* The unit of *XR* is runs.

* Furtado also created two simplified versions of *XR*, eXtrapolated Runs Reduced (*XRR*) and eXtrapolated Runs Basic (*XRB*), which simply omit terms that might be difficult to find in seasonal data.

$$XR = 0.50 \cdot 1B + 0.72 \cdot 2B + 1.04 \cdot 3B + 1.44 \cdot HR + 0.34 \cdot (HBP + TBB - IBB))$$
$$+ 0.25 \cdot IBB + 0.18 \cdot SB - 0.32 \cdot CS - 0.090 \cdot (AB - H - K) - 0.098 \cdot K$$
$$- 0.37 \cdot GIDP + 0.37 \cdot SF + 0.04 \cdot SH$$

XR represents one of the first attempts to *fit* a model to the data and estimate the coefficient vector β using formal statistical methods. It is not hard to approximate the weights specified in *XR* using regression. The coefficients we obtain are

Event	$\hat{\beta}$ for 1955–1997	$\hat{\beta}$ for 1969–2008
1B	0.5230	0.5327
2B	0.6617	0.6805
3B	1.0180	0.9993
HR	1.4756	1.4419
SB	0.1225	0.1358
CS	−0.2119	−0.1867
BIPO	−0.0967	−0.0998
SO	−0.0827	−0.0934
BB	0.3380	0.3425
IBB	−0.1491	−0.1854
GIDP	−0.3928	−0.3163
SF	0.6301	0.6438
SH	0.0630	−0.0141

It is important to understand that these coefficients provide a best fit only during the time intervals over which they were estimated, which in the case of *XR* was from 1955 to 1997. If instead, we fit the coefficients over the time period from 1969 (when the pitching mound was lowered and expansion occurred) to 2008 (when improved drug testing was implemented), we can see that most of the coefficients change only slightly. One notable exception is the coefficient for sacrifice bunts (*SH*), which flips its sign from positive to negative.

Since these coefficients can be interpreted as run values associated with each of these events, one can see evidence for many sabermetric insights—similar to those discussed earlier in the work of Lindsey—therein. To name a few, the small, and perhaps even negative coefficient for *SH* suggests that there is little value to sacrifice bunts, and that perhaps they are detrimental to offense. The similar coefficients for strikeouts (*SO*) and ball in play outs (*BIPO*s) suggest that there may be little difference in their value. The substantial negative coefficient on intentional walks (*IBB*) confirms that intentional walks are not nearly as helpful to the offense as a walk or hit by pitch. The ratio of the values on *SB* and *CS* provides an approximation for the "break-even" stolen base percentage. In *XR*, one needs $0.32/0.18 = 1.78$ stolen bases for every caught stealing just to positively impact run scoring, implying an effective stolen base percentage of 64%. Thus, runners who are stealing bases at less than a 64% success rate are actually costing their team runs.

One notable difference between the coefficients in *XR* and the output from the regression model discussed earlier is the weight on sacrifice flies (*SF*s). Here, collinearity is at work, since every sacrifice fly results in exactly one run being scored. Thus, Furtado has (somehow) adjusted the value downward. *XR* is often converted into a rate statistic by

TABLE 1.7

Summary of Combined Offensive Models for Selected 1987 NL Players

Player	*PA*	*TA*	*OPA*	*ERP*	*XR*	*XR27*
Jack Clark	558	1.265	0.615	113.3	114.3	10.09
Eric Davis	562	1.199	0.632	113.8	114.4	8.90
Tim Raines	627	1.146	0.587	119.6	118.7	8.68
Tony Gwynn	680	1.116	0.574	129.0	126.4	8.62
Dale Murphy	693	1.120	0.598	133.0	132.0	8.56
Darryl Strawberry	640	1.134	0.612	124.0	123.5	8.40
Andre Dawson	662	0.874	0.552	111.1	109.6	6.42
Ozzie Smith	706	0.833	0.466	96.2	97.9	6.06

Note: While Clark and Davis rank highly in terms of the rate metrics *TA*, *OPA*, and *XR*27, they did not play as often, leading to lower totals for the count metrics *ERP* and *XR*. Clark's *XR*27 of 10.09 suggests that if Clark were to successively bat until he recorded 27 outs, his team would score 10.09 runs, on average.

dividing the total number of eXtrapolated Runs produced by the number of outs made and multiplying the result by 27 (the number of outs in a nine-inning game). This statistic, known as *XR*27, measures the expected number of runs the player would produce in a normal game if the lineup contained nine identical batters. We will return to this idea in the following section.

In Table 1.7, we summarize these combined offensive metrics for our 1987 NL players.

1.2.4 Comparison of Linear Models

Because of the linearity of the models discussed earlier, we can sensibly compare their coefficients. In Table 1.8, we summarize the value of each event considered by each model relative to a single. That is, for each event j, we report β_j/β_{single}. The similarities in these relative weights overwhelm the differences. It is easy to see, however, that home runs are overvalued by *SLG* relative to *wOBA* or *XR*, just as they are dramatically undervalued by *AVG* and *OBP*.

It should not be surprising then, that these metrics are highly correlated. In Table 1.9, we show the pairwise correlation matrix between these metrics for nearly 100,000 qualified player-seasons spanning 1871–2014.

The question of accuracy in matching team runs is mostly moot, since the regression-based results shown earlier provide the best fit among all possible linear models with least-squares criteria. Nevertheless, we discuss the accuracy of these metrics (and others) in Section 12.6.

1.2.5 Multiplicative Models

The linear models in the previous section were all predicated on the idea that each event has a fixed, average value, and that any accrual of those events would have a proportional impact upon scoring. Usually, this assumption is quite reasonable. However, it has been observed that the average run value of certain events will change based on the run-scoring environment. Moreover, one event might be worth a different amount depending on the value of another event. For example, consider the value of a walk. In the high run-scoring

TABLE 1.8

Comparison of Offensive Coefficients from Linear Models, Relative to the Value of a Single, that is, β_j/β_{single}

	BAVG	*OBP*	*SLG*	*BR*	*WOBA*	*TAVG*	*OPA*	*ERP*	*XR*
1B	1	1	1	1.00	1.000	1	1.0	1.0000	1.000
2B	1	1	2	1.81	1.431	2	2.0	1.6667	1.440
3B	1	1	3	2.17	1.820	3	2.5	2.3333	2.080
HR	1	1	4	2.98	2.366	4	3.5	3.0000	2.880
uBB	0	1	0	0.70	0.777	1	0.8	0.6667	0.680
IBB	0	1	0	0.70	0.000	1	0.8	0.6667	0.500
HBP	0	1	0	0.70	0.813	1	0.8	0.6667	0.680
SO	0	0	0	−0.64	0.000	0	0.0	−0.2017	−0.196
BIPOO	0	0	0	−0.64	0.000	0	0.0	−0.2017	−0.180
GIDP	0	0	0	−0.64	0.000	0	0.0	−0.4033	−0.920
SF	0	0	0	0.00	0.000	0	0.0	0.0000	0.740
SH	0	0	0	0.00	0.000	0	0.0	0.0000	0.080
SB	0	0	0	0.00	0.000	1	0.5	0.0000	0.360
CS	0	0	0	0.00	0.000	0	0.0	−0.2017	−0.640

TABLE 1.9

Correlation Matrix for Linear Models, Based on All Qualified Batters from 1871 to 2014

	BAVG	*OBP*	*SLG*	*WOBA*	*BR*	*TAVG*	*OPA*	*ERP*	*XR*
BAVG	1.00								
OBP	0.71	1.00							
SLG	0.59	0.65	1.00						
WOBA	0.70	0.87	0.94	1.00					
BR	0.70	0.89	0.92	1.00	1.00				
TAVG	0.62	0.87	0.89	0.97	0.97	1.00			
OPA	0.62	0.78	0.96	0.97	0.97	0.97	1.00		
ERP	0.67	0.83	0.96	0.99	0.99	0.96	0.98	1.00	
XR	0.68	0.84	0.94	0.98	0.98	0.98	0.98	0.99	1.00

environments of the late 1990s and early 2000s, it may have been the case that walks were quite valuable relative to their value in the low run-scoring environment in the early 2010s.* In the former era, home runs were comparatively likely. Thus, simply getting on base meant that your chances of scoring were much higher than in the current era, where a runner on first is more likely never to come around and score due to the lack of extra base power.

In linear models, the two separate skills of getting on-base (i.e., *OBP*) and moving runners around the bases (i.e., *SLG*) are not allowed to interact. That is, the benefit to increasing

* Running the regressions alluded to the aforementioned results in a run value of 0.35 for walks during the period 1995–2005, but a value of 0.31 for the period from 2006 to 2014.

slugging is the same whether on-base percentage is high or low. In the previous section, we considered the formula for a generic *OPS*:

$$OPS = a \cdot OBP + b \cdot SLG$$

Here, additional slugging power will result in b being contributed, but this is essentially independent of *OBP*.* Conversely, for the models discussed in the following text, additional slugging power could contribute more or less to overall offensive output based on the value of *OBP*. These models share a multiplicative functional form.

1.2.5.1 Runs Created

Bill James recognized linearity as a constraint and sought a metric that captured the interplay between getting on-base and moving runners around the bases (James, 1986). The basic notion of *Runs Created* (*RC*) was that slugging percentage was more valuable when *OBP* was also high, since there would be more runners on base to move around the bases with extra base hits. Thus, he modeled offensive production using a multiplicative model:

$$RC = \frac{(H + BB) \cdot TB}{AB + BB} = \frac{H + BB}{AB + BB} \cdot \frac{TB}{AB} \cdot AB \approx OBP \cdot SLG \cdot AB$$

Here, the contribution of *OBP* is nonlinear, since it depends on the value of *SLG*. Not so obviously, the value of *RC* is on the scale of runs, as we show in Figure 1.3.

1.2.5.2 Batter's Run Average

An even simpler multiplicative model is *batter's run average* (*BRA*), which was posited by Cramer and Palmer (1974). In its simplest version:

$$BRA = OBP \cdot SLG\,.$$

Subsequent versions incorporated stolen bases by expanding the formula to

$$\begin{aligned} BRA &= \frac{H + BB + HBP + (1/2)(SB - CS)}{AB + BB + HBP} \cdot SLG \\ &= \frac{H + BB + HBP + (1/2)(SB - CS)}{AB + BB + HBP} \cdot \frac{TB}{AB}\,. \end{aligned}$$

While subsequent "improvements" to runs created have resulted in more complicated formulas that hew closer to the truth, the sabermetric value of runs created is the simplicity and insight of its original formulation. It is not surprising that linear weights estimators and other metrics that have been fit to data are more accurate than runs created. On the other hand, what is surprising is that runs created works as well as it does (see Figure 1.3). Specifically, the genius of creating a nonlinear model of the run creation process as an interaction between getting on-base and moving runners around is the lasting contribution of this line of inquiry.

* Of course, *SLG* and *OBP* are functionally dependent, since they both depend, they include hits, and thus the qualifier "essentially."

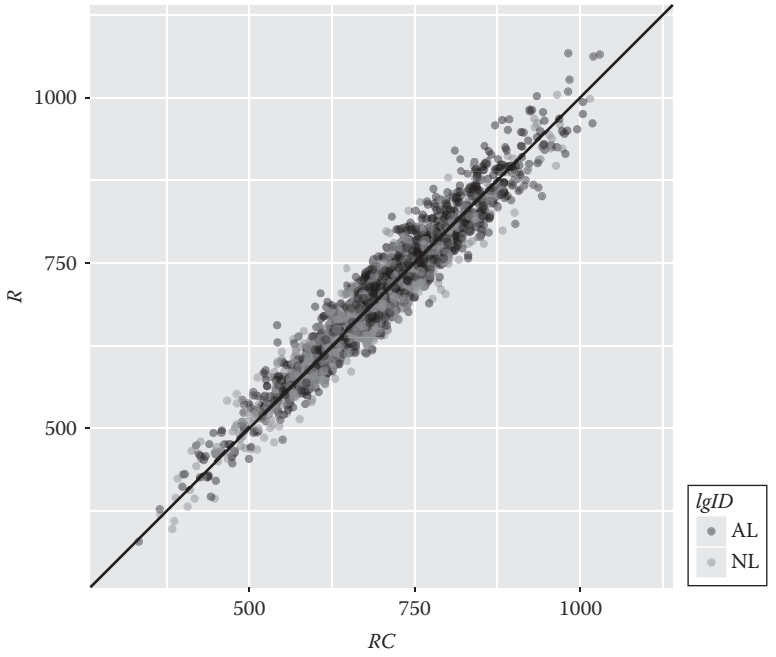

FIGURE 1.3
Scatterplot between runs created (*RC*) and runs scored (*R*) among all teams since 1915. The correlation is above 0.96.

1.2.5.3 Earnshaw Cook's Scoring Index

In his opus, Cook (1964) defined scoring index (DX) as

$$DX = SO \cdot \frac{H + BB + E + HBP - 2 \cdot SH - XBH}{PA} \cdot \frac{TB}{PA}.$$

The middle term in this equation is basically the "probability of reaching first base," and thus Cook's formula is very close in spirit to James' runs created, while preceding it by several years. However, unfortunately for Cook, Scoring Index was not adopted with the fanfare of Run Created.

In Table 1.10, we summarize the value of nonlinear models for our selected 1987 players.

1.2.6 Accuracy of Run Estimators

In Table 1.11 and Figure 1.4, we summarize the accuracy of the metrics discussed thus far on historical baseball data. As noted previously, it is not surprising that *XR* has the lowest root-mean-square error (*RMSE*), since it was constructed by mathematically optimizing for that criterion. It is more remarkable that simple, intuitive metrics such as *OPS* and *RC* perform so well.

The quest for a "best" model is quixotic. Nevertheless, this analysis confirms that eXtrapolated Runs (*XR*) is a good general purpose run estimator. The concept (i.e., linear weights) is simple, the units (runs) are intuitive and the performance is excellent. The downside is

TABLE 1.10

Comparison of Nonlinear Models for Select 1987 NL Players

Player	*PA*	*RC*	*BRA*	*DX*
Dale Murphy	693	135.8	0.242	20.16
Tony Gwynn	680	134.6	0.228	5.54
Darryl Strawberry	640	122.2	0.232	16.53
Tim Raines	627	119.2	0.226	7.71
Jack Clark	558	115.3	0.274	21.99
Andre Dawson	662	113.5	0.186	11.78
Eric Davis	562	112.3	0.236	19.07
Ozzie Smith	706	90.5	0.150	3.39

TABLE 1.11

Summary of Accuracy of Various Run Estimators, 1954–1999

Metric	*RMSE*	R^2
XR_G	0.1339	0.9434
ERP_G	0.1408	0.9374
BRA	0.1566	0.9226
RC_G	0.1587	0.9205
wOBA	0.1592	0.9200
OPS	0.1597	0.9195
TA	0.1597	0.9195
OPA	0.1727	0.9059
SLG	0.2177	0.8505
OBP	0.2531	0.7980
AVG	0.3171	0.6829
BR_G	0.3760	0.5539
DX_G	0.4272	0.4243

Note: The *RMSE* is computed as the root mean squared error of the simple linear regression model between team runs per game and each metric. (Compare with page 230 of Albert and Bennett (2003).) R^2 is the coefficient of determination (i.e., the percentage of the variability in runs explained by the model) for the same regression model.

that the formula is long, difficult to remember, and painstaking to type in. If one is using a computer, then *XR* is hard to beat. Conversely, for back-of-the-envelope calculations, *BRA* and *OPS* can be quicker (given that *OBP* and *SLG* are available). *RC* is of similar complexity but has the advantage of intuitively meaningful units (runs). It is harder to see the practical virtues of *ERP*, *wOBA*,* and *OPA*.

* Proponents of *wOBA* often tout the scale of *OBP* as intuitively meaningful, but how is that more intuitive than runs per game?

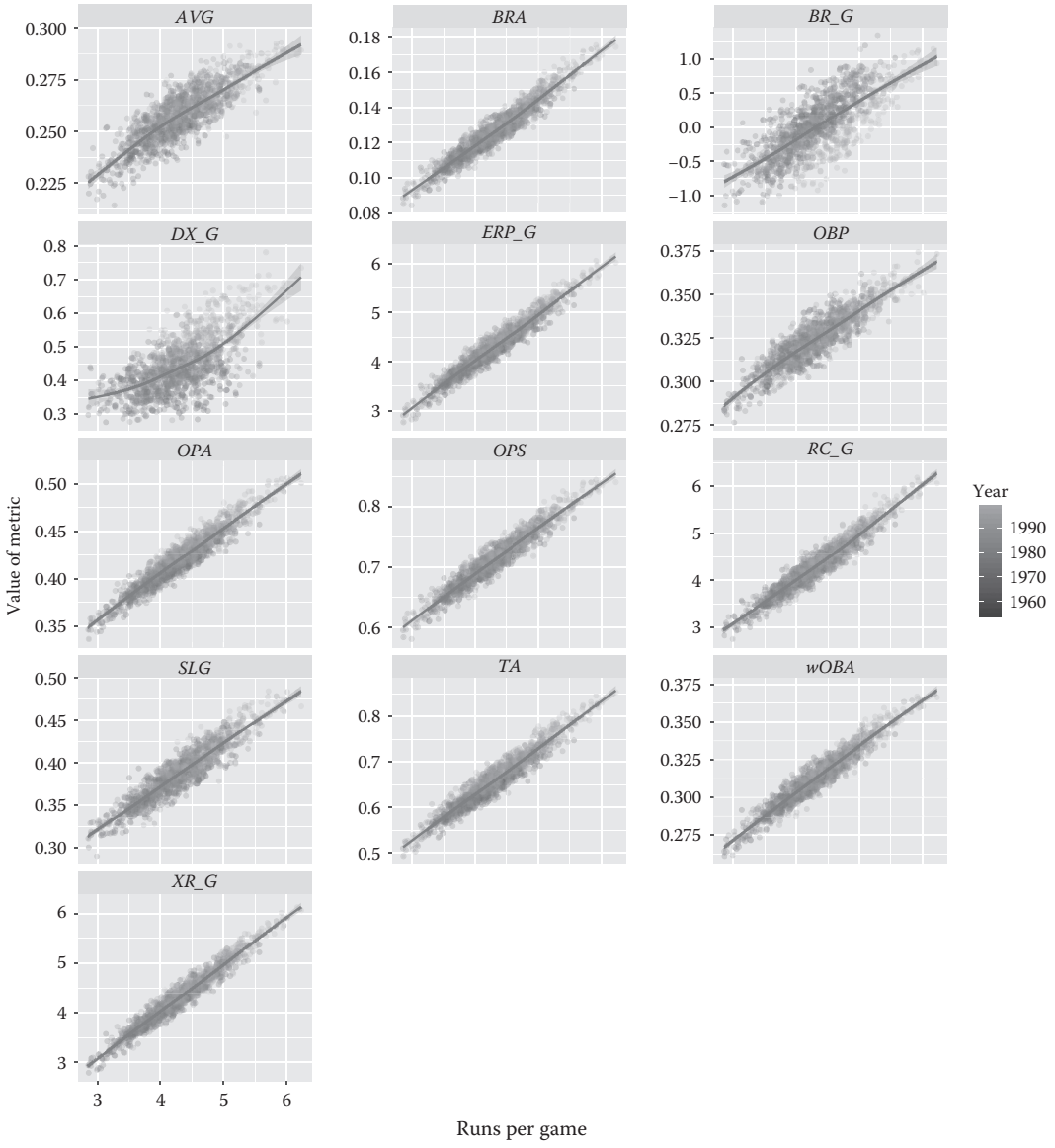

FIGURE 1.4
Comparison of accuracy of run estimators, 1954–1999. Note the tight fit for eXtraploted Runs (*XR*) and the relatively looser fit for Scoring Index (*DX*).

1.3 Models Based on Play-by-Play Data

The models in the previous section are computable using preaggregated data. That is, one only needs to know the counts of various events at the seasonal level to compute the metric. We now turn to models that use play-by-play data. These models are often situational or context-dependent. In this section, we will use examples from the 2013 major league season.

1.3.1 Markov Chain Models

A Markov chain is a *memoryless* stochastic process. That is, the behavior of the system in its present state does not depend on its previous state. Baseball is often assumed to have this property. For example, if a runner is on second and there is one out, it does not matter if the first batter hit a double and the next struck out, or if the first walked and moved up on ground out. Certainly, the run expectancy is the same in either case. Whether the psychological state of the pitcher is the same is debatable, but this simplifying assumption opens up a whole field of useful mathematical models (Trueman, 1977) (for a nice exposition, see D'Angelo (2010)).

Because baseball is naturally discrete and has 24 clearly defined (*base*, *out*) states, a Markov chain is a natural model to apply. These models can be useful in understanding the long-term probabilities associated with a fairly complex interdependent system.

Consider the 24 (*base*, *out*) states of an inning, and let's add a 25th state corresponding to having 3 outs. We denote the *state vector* of length 25 by $\mathbf{s}$. Since every inning always begins in the state $(0, 0)$, the initial state vector $\mathbf{s}_0$ has a 1 as its first entry, followed by 24 zeros.

Next, consider the probability of moving from any one of the 25 initial states to any other of the 25 states. Clearly, some of these probabilities will be zero, since you can't go from two outs and nobody on to one out and a man on first. But many of these probabilities will be nonzero, and moreover they will be normalized so that the sum of the probabilities from any initial state is one. For example, in Table 1.12 we show the transition probabilities from the $(0, 0)$ state in which all innings begin. Only five outcomes are possible: either there is a runner on exactly one of the three bases, or the bases are empty and there is one out (if the first batter made an out) or no outs (if the first batter hit a home run). By far, the most likely outcome is that the first batter made an out (about 68% of the time), and the state is now $(0, 1)$.

Extending this idea, let the transition matrix $\mathbf{T}$ be the 25×25 matrix containing the probabilities of the inning moving from one state to the next. Formally, $\mathbf{T}_{i,j} = \Pr(s_f = j | s_i = i)$ for $i, j = 1, \ldots, 25$, where s_i and s_f are the states of the inning before and after the play, respectively.

The Markov process allows us to compute the probability of being in any given state as the inning evolves. Thus, the vector $\mathbf{s}_1 = \mathbf{T}\mathbf{s}_0$ gives the probabilities shown in Table 1.12—these are the probabilities of being in any state after one batter (the probability of being in any other state is zero). Similarly, the vector $\mathbf{s}_2 = \mathbf{T}\mathbf{s}_1 = \mathbf{T}^2\mathbf{s}_0$ gives the probabilities after

TABLE 1.12

Transition Probabilities from the Inital State of an Inning, 2013

Begin State	Begin Outs	End State	End Outs	Freq
0	0	000	0	0.0296
0	0	000	1	0.6799
0	0	001	0	0.2354
0	0	010	0	0.0497
0	0	100	0	0.0054

Source: MLBAM, GameDay files, http://gd2.mlb.com/components/game/mlb/, accessed July 1, 2016.

Note: About 68% of the time, the following state is one out, no one on base.

two batters, and so on. After many batters have batted, the probability of being in the three out state approaches 1.

We learn a bit more with a little linear algebra. Let **Q** be the matrix obtained by deleting the three out row and column from **T**. Inverting this 24 × 24 matrix gives us

$$\mathbf{E} = (\mathbf{I} - \mathbf{Q})^{-1},$$

where **I** is the identity matrix. It can be shown that the entries of **E** are the expected number of times that any given inning will be in each state. By summing over the rows, we arrive at the expected number of batters remaining in the inning (Albert, 2003). These values are shown in Table 1.13. Thus, the average number of batters per inning is about 4.2, but with two outs, it is always about 1.45, pretty much regardless of the configuration of the base runners. This makes sense, since the number of batters remaining in the inning has a lot more to do with the probability of the batter reaching base than it does with the configuration of the bases.

Another neat trick is that if we know the expected number of runs scored on a single play **q**, we can use a little algebra to compute the expected run vector $\mathbf{r} = \mathbf{Eq}$. Table 1.14 displays the expected run matrix that results from the Markov chain model

TABLE 1.13

Expected Number of Batters Remaining in the Inning from Each of the States

BaseCode\Outs	0	1	2
000	4.244	2.857	1.458
001	3.992	2.685	1.452
010	4.295	2.890	1.505
011	4.005	2.678	1.437
100	4.336	2.951	1.518
101	4.106	2.733	1.483
110	4.400	3.072	1.543
111	3.967	2.704	1.421

TABLE 1.14

Expected Run Matrix Based on Markov Chain Model, 2013

BaseCode\Outs	0	1	2
000	0.464	0.245	0.093
001	0.819	0.491	0.213
010	1.088	0.634	0.301
011	1.394	0.870	0.411
100	1.340	0.929	0.349
101	1.700	1.131	0.488
110	1.986	1.384	0.558
111	2.143	1.535	0.725

Note: Compare these values to those presented in Table 1.1.

seeded by the transition matrix from the 2013 data. These values are very similar to the ones shown in Table 1.1, suggesting that the Markov chain model is a reasonable approximation to reality.

The transition matrix $\mathbf{T}$ described earlier is based on observations from the entire league. Suppose instead that we could estimate $\mathbf{T}_i$ for each player i. Then using the same equations outlined earlier, we could compute the vector $\mathbf{r}_i = \mathbf{E}_i\mathbf{q}$. The first entry in this vector is the expected number of runs scored in an inning with transition probabilities defined by $\mathbf{T}_i$, which are specific to player i. The interpretation is that if all nine batters were clones of player i, then $9\mathbf{r}_{i,1}$ is the expected number of runs scored per nine innings (e.g., per game) by a team of player i (Pankin, 1987, 2007). This is a useful measure of player i's offensive ability, comparable to *XR*27 as discussed earlier. A value of 8.48 runs per game, for example, would imply that the player was about twice as productive as the average batter.

However, the task of estimating $\mathbf{T}_i$ accurately is nontrivial. There are $25 \times 25 = 625$ entries in $\mathbf{T}_i$, many of which require observations to estimate. So even for players who play a full season and accrue 700 plate appearances, many entries in the transition matrix will have essentially no data. Thus, assumptions must be made. Most commonly, assumptions about baserunning are made to build transition matrices. Scoring index (Bukiet et al., 1997), offensive earned-run average (OERA) (Cover and Keilers, 1977), and modified offensive earned-run average (MOERA) (Katsunori, 2001) all use the same Markov model with slightly different baserunning assumptions.

1.3.2 *RE*24

Like the Markov models, *RE*24 is a context-dependent model that uses play-by-play data to evaluate batters (Fangraphs Staff; Appelman, 2008). Simply put, *RE*24 is the sum of the changes to the run expectancy matrix during a player's plate appearance. That is,

$$RE24 = \sum \delta_{f,i}$$

where $\delta_{f,i}$ is defined as in Equation 1.1. While the sum of the δ's over a long period of time should equal the number of runs scored, in each plate appearance a player may accrue either a negative or positive run value. Thus, the sum of these δ's over a period of time will provide a measure of the number of runs that were produced during each player's turns at-bat. *RE*24 forms the basis for the computation of *openWAR* (Baumer et al., 2015).

Skoog (1987) called this the "value-added" approach to runs created, while Albert (2001) considered this metric as well as the companion rate statistic *RE*24/*PA* by dividing *RE*24 by the number of plate appearances. In Table 1.15 we report the top players in terms of *RE*24 for the 2013 season.

*RE*24 is an appealing concept that has arisen in multiple places over time. However, its context-dependent nature is an important feature (or bug, depending on your point-of-view). Because *RE*24 measures the actual changes in run expectancy, players with widely disparate opportunities could score higher or lower than their performance probably warrants. For this reason, *RE*24 is probably not a good basis for a prediction metric, such as those discussed in Section 1.4.

1.3.3 Baserunning Models

The idea behind *RE*24—that changes to the expected run matrix provide a measure of offensive performance—can be extended to baserunning as well as batting.

TABLE 1.15

*RE*24 Leaders for 2013

Player	*PA*	*AB*	*H*	*HR*	*RE24*	*RE24_PA*
Cabrera, M	651	554	193	44	81.33	0.125
Davis, C	673	584	167	53	75.53	0.112
Trout	716	589	190	27	69.78	0.097
Goldschmidt	710	603	182	36	64.91	0.091
Freeman, F	629	551	176	23	57.69	0.092
Ortiz, D	600	519	160	30	54.81	0.091
Cano	681	605	189	27	52.81	0.078
Votto	726	581	177	24	48.87	0.067
Carpenter, M	719	628	201	11	48.49	0.067
Encarnacion	621	530	144	36	48.21	0.078

Source: MLBAM, GameDay files, http://gd2.mlb.com/components/game/mlb/, accessed July 1, 2016.

In both Ultimate Base Running (UBR) and Base Runner Runs (BRRs), the baserunning value on an event e can be inferred from the difference

$$\delta_{f,i} - \bar{\delta}_{f,i}(e),$$

where

$\delta_{f,i}$ is again defined as in Equation 1.1

$\bar{\delta}_{f,i}(e)$ is the average value of the change in run expectancy on the batting event e

The difference between the two systems is how to measure $\bar{\delta}_{f,i}(e)$. Click (2005) estimates the value of $\bar{\delta}_{f,i}(e)$ by averaging δ's across the league, taking into account the state of the inning, how the ball is hit and where. For example, a ground ball to shortstop will result in a different value than a fly ball to rightfield. This notion of using play-by-play data to determine the expected position of base runners—as opposed to making assumptions about the movement of base runners based on the type of hit—is also employed in the baserunning component of *openWAR* (Baumer et al., 2015). However, in that system, a formal regression model is employed, unlike the empirical estimates of Click (2005). Alternatively, in Base Runner Runs, Fox (2005) estimates $\bar{\delta}_{f,i}(e)$ using conservative assumptions about baserunning. For example, a base runner is assumed to move up only one base on a single, and runners accrue additional value by advancing further.

1.3.4 Nonoutcome-Based Models

According to Lewis (2003), the Oakland A's worked with a company called AVM Systems on a model for the expected run value of any batted ball based on its location, but regardless of its outcome. That is, instead of giving batters credit for a bloop single just because the fielders couldn't catch it, why not give the batter credit for only the part of the transaction that he was involved in (namely, the weakly-hit fly ball)? If a ball was crushed, but a fielder happened to make an exceptional play, shouldn't the batter be credited for the hard-hit ball?

This type of modeling leads to a fundamentally different way of evaluating batters. Traditionally, for any batted ball i, we focus on the outcome ω_i (e.g., single, double, fly out), which is simply a designation from our list of categories Ω. Instead, the system alluded to assigns an expected run value $\mathbb{E}[r_i|x, y]$ to each batted ball, conditional upon its (x, y)-location on the field, but ignoring ω_i. One could imagine including additional covariates for batted ball speed, trajectory, etc.

In this section, we explore models based on this concept, but unfortunately very few details about these kinds of systems are publicly known. Nevertheless, new data sources that contain more accurate descriptions of the paths of batted balls are renewing interest in these types of models (Lindbergh, 2015).

1.3.4.1 DIBS

In 2015, Ben Jedlovec of Baseball Info Solutions presented a trajectory-based hitting evaluation metric called Defense-Independent Batting Statistics (DIBS). While the details of this proprietary system are unknown, it appears to be similar to the AVM system alluded to discussed earlier. It is claimed that DIBS has demonstrably greater predictive value than conventional batting statistics. DIBS includes the following covariates (Eddy, 2015):

- Batted ball location (e.g., (x, y)-coordinates)
- Hang time or ground ball velocity
- Batter handedness, speed, and power
- Park factors

1.3.4.2 HITf/x and Statcast

As we argued earlier, the outcomes ω_i are noisy—and this provides motivation for using the (x, y)-batted ball locations instead. But those locations are *also* noisy, suffering both measurement error and the vagaries of weather, altitude, etc. One further abstraction is to record the physical characteristics of the ball *as it leaves the bat*.

Since the late 2000s, Sportvision has been licensing HITf/x data, which contains launch angle, exit velocity, and trajectory information for batted balls in major and minor league parks (Sportvision, 2009). Unfortunately, these data are quite expensive, but many major league teams buy them. It is not hard to imagine how such data could be used to create a model for the expected run value of a batted ball. Note that in this case even the location of the batted ball is not important—rather we are using physical measurements from the milliseconds after impact to measure the strength of the hit. This is one further step removed from traditional metrics that focus on the outcome ω_i.

Statcast is a new player tracking system unveiled in 2014 and published in 2015 that uses Doppler radar to track the ball and video technology to track the players (Albert, 2016; Keri, 2014). These data provide unparalleled granularity about the locations of batted balls and players over fractions of seconds as each play unfolds. Undoubtedly, this will lead to new public domain models for batting and fielding evaluation.

1.3.5 Simulation-Based Models

Given the high degree of interaction within an inning of baseball, it is hard to understand how run scoring will change when more than one variable is changed. For example, how

do changes in baserunning strategy affect run scoring for different profiles of offensive ability? These questions are not easily answered through parameter estimation but can be addressed using simulation.

The seeds of simulation-based study of baseball lie in the Strat-O-Matic tactile game that inspired so many sabermetricians (Guzzo, 2005). Since the late 1980s, Diamond Mind Baseball has produced a computer simulation game that its creator, Tom Tippett, has used for research purposes (Keene, 2016; Tippett). Here again, the program is proprietary, and Tippett has been an employee of the Boston Red Sox since the mid-2000s.

In two articles, Baumer (2009) and Baumer and Terlecky (2010) used a simulation-based approach to estimating base runners value. Baumer et al. (2012) used the same simulation engine to explore the relationship between baserunning and batting abilities. This approach enables, for example, the baserunning strategies of a particular base runners to be evaluated in the context of several different offensive profiles for the batter hitting behind him.

1.4 Predictive Models

While much of the early work in sabermetrics focused on developing methods for proper evaluation of the past performances of batters and base runners, in the past few decades interest has grown in predicting the future performance of these players. Indeed, while sportwriters and award voters may be primarily interested in the former problem, the question of how good a player *was* is almost irrelevant to the decisions that major league teams make. A good prediction of how good a player *will be* is far more germaine to the problem of constructing a winning team.

Thus, whereas so far in this chapter, we have focused on the problem of estimating y_t given $\mathbf{X}_t$, in this section, we consider the question of predicting y_{t+1} given $\mathbf{X}_t$.

As the ability to predict y_{t+1} accurately is of obvious importance for both general managers and fantasy players alike, it has become something of a sport unto itself. It would be impractical to review all of the publicly known projection systems here (for a larger list, see Fangraphs Staff (2015a), or Larson (2015a) for a collection of known projections). Instead, we focus on illuminating the guiding principles of a few of the better known projection systems and illustrating the key differences between them. The first characteristic distinction is between models that simply provide *point estimates* as opposed to those that provide *interval estimates*.

1.4.1 Models That Produce Point Estimates

1.4.1.1 Marcel

In comparing forecasting models, it is useful to have a benchmark. The most commonly used benchmark is the relatively simple Marcel projection system promulgated by Tango (2012). With great modesty, Tango describes Marcel as "most basic forecasting system you can have, that uses as little intelligence as possible." The system is not utterly trivial—it is a time-weighted 3-year average of each player's previous statistics but also takes into account each player's age and employs regression to the mean in the context of their playing time.

Let p_{ijt} be the rate at which player i achieves outcome j in season t. Since Marcel uses the previous three seasons of data, let $p'_{ijt} = \begin{pmatrix} p_{ij,t-1} & p_{ij,t-2} & p_{ij,t-3} \end{pmatrix}^T$ be the vector of those rates

over the past three seasons. The corresponding 3×1 vector q'_{it} gives the number of plate appearances for player i in the three seasons preceding t.

Marcel uses a fixed time-discounting vector we define as $t = \begin{pmatrix} 5 & 4 & 3 \end{pmatrix}^T$ and a fixed regression to the mean vector we define as $\bar{q} = \begin{pmatrix} 100 & 100 & 100 \end{pmatrix}^T$ (for a fuller description, see Baumer (2014)).

In this notation, Marcel computes the predicted rate for each player as

$$\hat{p}_{ijt} = w_{it} \cdot A_{it} \cdot p'_{ijt} + (1 - w_{it}) \cdot A_{it} \cdot \bar{p}'_{jt}, \tag{1.2}$$

where

$\bar{p}'_{jt}$ is the corresponding vector of league average rates for event j over the past three seasons

$w_{it} = \frac{t^* q'_{it}}{t^* q'_{it} + t^* \bar{q}}$

$A_{it} = \frac{t^* diag(q'_{it})}{t^* q'_{it}}$

$$diag(q'_{it}) = \begin{pmatrix} q_{i,t-1} & 0 & 0 \\ 0 & q_{i,t-2} & 0 \\ 0 & 0 & q_{i,t-3} \end{pmatrix}.$$

The quantity w_{it} depends only on player i's number of plate appearances over the past three seasons and is called *reliability*. Similarly, the quantity A_{it} depends on the same but acts as a scaling factor. Thus, by defining $\theta_{ijt} = A_{it} p'_{ijt}$ and $\bar{\theta}_{ijt} = A_{it} \bar{p}'_{jt}$, we have

$$\hat{p}_{ijt} = w_{it} \cdot \theta_{ijt} + (1 - w_{it}) \cdot \bar{\theta}_{ijt}.$$

This is immediately evocative of the well-known formula for the posterior mean for the frequency parameter in a beta-binomial Bayesian model. Here, θ_{ijt} is the time-discounted observed rate of event j for player i, and $\bar{\theta}_{ijt}$ is the corresponding league average rate, with the time-discounting depending on player i's playing time. Marcel's reliability w_{it} is apparent as a *shrinkage* factor.

The number of predicted plate appearances in season t is

$$\hat{q}_{it} = \frac{1}{2} q_{i,t-1} + \frac{1}{10} q_{i,t-2} + 200,$$

where this estimate is asserted without justification. The final computation is a piecewise linear age adjustment

$$f_i(t) = \begin{cases} -0.003 \cdot (29 - (t - birthYear_i)), & t - birthYear_i > 29, \\ 0.006 \cdot (29 - (t - birthYear_i)), & \text{otherwise}, \end{cases}$$

which has the effect of increasing the expected performance of those who have yet to reach their prime and decreasing the expected performance of those who are past their prime. While the intent is clear, no justification for this particular aging curve is given.

TABLE 1.16

Carlos Beltran's Recent History Preceding the 2004

Year	*PA*	*HR*	*HR/PA*
2001	680	24	0.0353
2002	722	29	0.0402
2003	602	26	0.0432

Consider the canonical example of predicting Carlos Beltran's home run count heading into the 2004 season. In Table 1.16, we present Beltran's relevant data. Beltran hit more home runs that the average player in each of these seasons. The home run rates for the average player are shown in Table 1.17.

Following Equation 1.2, for Beltran in 2004, we have

$$\begin{aligned}
&\hat{p}_{beltran,HR,2004} \\
&= \underbrace{\frac{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}680\\722\\602\end{pmatrix}}{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}680\\722\\602\end{pmatrix}+\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}100\\100\\100\end{pmatrix}}}_{w_{beltran,2004}} \cdot \underbrace{\frac{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}680 & 0 & 0\\0 & 722 & 0\\0 & 0 & 602\end{pmatrix}\begin{pmatrix}0.0353\\0.0402\\0.0432\end{pmatrix}}{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}680\\722\\602\end{pmatrix}}}_{A_{beltran,2004}\cdot p'_{beltran,HR,2004}} \\
&+ \underbrace{\frac{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}100\\100\\100\end{pmatrix}}{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}680\\722\\602\end{pmatrix}+\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}100\\100\\100\end{pmatrix}}}_{1-w_{beltran,2004}} \cdot \underbrace{\frac{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}680 & 0 & 0\\0 & 722 & 0\\0 & 0 & 602\end{pmatrix}\begin{pmatrix}0.0300\\0.0278\\0.0286\end{pmatrix}}{\begin{pmatrix}3 & 4 & 5\end{pmatrix}\begin{pmatrix}680\\722\\602\end{pmatrix}}}_{A_{beltran,2004}\cdot \bar{p}'_{HR,2004}} \\
&= \underbrace{\frac{7938}{9138}}_{w_{beltran,2004}} \cdot \underbrace{\frac{318}{7938}}_{\theta_{beltran,HR,2004}} + \underbrace{\frac{1200}{9138}}_{1-w_{beltran,2004}} \cdot \underbrace{\frac{228}{7938}}_{\bar{\theta}_{beltran,HR,2004}} \\
&= 0.8687 \cdot 0.0401 + 0.1313 \cdot 0.0287 \\
&= 0.0386
\end{aligned}$$

The number of plate appearances Beltran is expected to take in 2004 would be

$$\begin{aligned}
\hat{q}_{beltran,2004} &= \frac{1}{2}q_{beltran,2003} + \frac{1}{10}q_{beltran,2002} + 200 \\
&= \frac{1}{2}\cdot 602 + \frac{1}{10}\cdot 722 + 200 \\
&= 573.2
\end{aligned}$$

TABLE 1.17

MLB's Recent History Preceding the 2004 Season

Year	*PA*	*HR*	*HR/PA*
2001	177,941	5333	0.0300
2002	178,487	4958	0.0278
2003	177,376	5078	0.0286

Note: Pitchers are excluded by filtering players with at least 2 plate appearances per game played.

His predicted number of home runs is thus $573.2 \cdot 0.0386 = 22.1$. Since Beltran was 27 in 2004, this number gets adjusted up by $(29 - 27) \cdot 0.006 = 0.012$. Thus, our final estimate is that Beltran would have hit $1.012 \cdot 22.1 = 22.4$ home runs in 2004. (In reality, Beltran had a breakout year and hit 38 home runs.)

Despite its relative simplicity and idiosyncratic modeling choices, Marcel has proven to be a durable and effective projection system. Its primary contribution is to provide a benchmark against which all other projection systems can be compared.

1.4.1.2 ZiPS

ZiPS is a popular projection system created by Szymborski (2015). ZiPS is similar to Marcel in that is based on a time-discounted average of recent performance (Cockcroft, 2015). However, ZiPS makes several significant improvements over Marcel:

- It uses 4 years of data, rather than 3, with a $t = \begin{pmatrix} 8 & 5 & 4 & 3 \end{pmatrix}^T$ time-discount vector.
- Shrinkage toward the mean is different for different rate statistics. In particular, ZiPS uses McCracken (2001)'s Defensive-Independent Pitching Statistics (DIPS) theory as its starting point in determining shrinkage levels. This means that, for example, a batter's batting average on balls in play will be shrunk much further toward the mean than his strikeout rate.
- Age adjustments are based on a historical comparison with "similar" players. Thus, heavy-set sluggers may age differently than speedy leadoff hitters.
- Rookies and international players are projected based on major league equivalencies (MLEs).
- Park adjustments are included.

Of course, once season t is played, it is easy enough to measure the accuracy of any prediction system. The three obvious metrics are the mean absolute error ($MAE = \frac{1}{n}\sum_{i=1}^{n} |y_{it} - \hat{y}_{it}|$), root mean squared error ($RMSE = \frac{1}{n}\sum_{i=1}^{n}(y_{it} - \hat{y}_{it})^2$), and the Pearson correlation ($r(y_{it}, \hat{y}_{it})$). It is not surprising that by the later criteria, ZiPS tends to best Marcel (Meyer, 2014). What is perhaps more surprising is that Marcel occasionally bests ZiPS according to the *RMSE* criteria (Larson, 2015b). One explanation that has been proffered is that Marcel implicitly estimates an accurate run-scoring environment.

Another approach—common in machine learning—is to create an *ensemble* prediction based on weighted averages of existing projection systems (Highley et al., 2010).

1.4.2 Models That Produce Interval Estimates

Naturally, contientious forecasting of the future performance of baseball players involves uncertainty, and while the point estimates produced by the aforementioned models are certainly useful, general managers are better able to assess risk using forecasts that provide not just a point estimate for each player's future performance, but a probability distribution over that potential future performance.

1.4.2.1 Pecota

One of the first publicly available projection systems that published forecast distributions was Pecota (Schwarz, 2005b). Pecota is a proprietary system originally developed by Nate Silver, and now operated by Baseball Prospectus. In addition to the breakthrough that was publishing forecast distributions, Pecota is fundamentally different than most projection systems, in that it is based on *nearest neighbor* analysis (Silver, 2003). While the details of Pecota are proprietary and a full treatment of nearest neighbor algorithms is beyond what we can accomplish here, since this approach differs significantly from many of the other predictive models, we will describe the general framework upon which Pecota rests.

Note that in our notation, the matrix $\mathbf{X}$ has p columns and n rows, where each row corresponds to a player and each column to an attribute of that player. Each row is a vector in $\mathbb{R}^p$, and thus each player can be thought of as a point in a p-dimensional space. We can define a distance metric $d(s,t)$ for any two points $s,t \in \mathbb{R}^p$ (e.g., the Euclidean distance metric). Then given a desired number of comparison players k, for any candidate player x^*, we can find the k players who are "closest" to x^* using the distance metric d. The average performance of those k players provides an estimate of the performance of player x^*. Such k-NN algorithms are well documented in the machine learning literature (James et al., 2013).

In Pecota, we know that among the p attributes considered are

- Production metrics that measure past performance using typical sabermetric statistics
- Usage metrics that measure playing time
- Phenotypic attributes that are specific to the physical characteristics of the player
- Role metrics that focus on fielding position or starter vs. relief usage

We don't know much about the distance metric—a *similarity score* in the parlance of Bill James—used in Pecota (i.e., is it Euclidean?), but we do know that Pecota incorporates age and adjustments for ballpark and league.

1.4.2.2 Steamer

Steamer Projections are another commonly cited set of predictions. Like ZiPS, Steamer is a Marcel-rooted forecasting system with a host of additional complexities built-in. In particular, Steamer uses a DIPS-inspired framework with variable time-discounting.

More recently, Steamer has produced a Shiny app that allows a user to explore the forecast distribution while experimenting with parameters (Cross, 2014).

1.4.2.3 Bayesian Models

While all of the forecasting systems described earlier incorporate the notion of regression to the mean, and at least two (Marcel and Steamer) are evocative of a Bayesian framework, none of the models discussed earlier are explicitly Bayesian. However, the academic literature is rich with Bayesian models for all kinds of things, with batting and baserunning being no exceptions.

In a widely cited article, Efron and Morris (1975) motivated their use of Stein's estimator using an example from baseball. In particular, their goal was to use each player's first 45 at-bats of the 1970 season to predict their batting average over the remainder of the season. Thus, baseball has provided a natural context for Bayesian analysis for nearly as long as Bayesian analysis has been done.

More recently, Neal et al. (2010) use linear regression to predict second-half performance given first-half performance, improving upon the Bayesian models of Brown (2008). Jensen et al. (2009) formulate a rigorous hierarchical Bayesian model for home run rate that surpasses Marcel—but not Pecota—by the *RMSE* metric, while at the same time producing full posterior distributions.

It is common in baseball parlance to identify certain players as "home run hitters," while others are not identified as such. Statistical analysis confirms this intuition, as a batter's home run rate is relatively highly correlated across consecutive seasons. Yet while it is clear that home runs are an important characteristic for hitters, the question of which components of a batter's profile represent the strongest *signal* remains. From DIPS theory, we know that much of the *noise* in pitcher (and in turn hitter) performance is captured by the volatile statistic of batting average on balls in play. McShane et al. (2011) uses a hierarchical Bayesian variable selection model to tease out which other hitting statistics seem most persistent for hitters over time. They find that while many of the 50 commonly cited metrics were correlated, most of the signal is captured by just 5: strikeout rate, walk rate, ground ball rate, isolated power, and Bill James's speed score. This result—in conjunction with DIPS—provides a meaningful starting place for forecasting position players, in that it confirms that most of what we know about batters are their plate discipline, how fast they run, how often they hit for power, and how often they hit the ball on the ground. From the point of view of a forecaster, much of the rest may just be noise.

Rather than choosing specific rates to follow, in general, it is natural and convenient to consider the observations in $\mathbf{X}$ as realizations of a multinomial distribution with parameters n (i.e., the number of plate appearances) and ρ, a vector of length p giving the true probabilities for each of the p outcomes of a plate appearance (with $\sum_{i=1}^{p} \rho_i = 1$). Viewing the data as having this multinomial likelihood, a natural Bayesian approach is to employ the conjugate prior distribution: the Dirichlet, which has the hyperparameter vector κ. This framework is simply the multivariate extension of the beta-binomial framework suggested by Marcel.

Recognizing that this multinomial Dirichlet framework, however natural, has certain limitations (namely, it doesn't understand DIPS), Null (2009) constructed the *nested* Dirichlet distribution for extending this framework. This remains the current state of the art, although recent work by Albert (2016) provides a simpler model.

1.5 Conclusion

It is not a stretch to argue that the problem of evaluating batters (and base runners) has provided the primary motivation for sabermetrics, and in turn for sports analytics in general. This field has seen the development of fundamental tools (e.g., Lindsey's expected run matrix), brilliant insights (e.g., runs created), and sophisticated statistical models (e.g., Null's nested Dirichlet) that have collectively revolutionized the way we assess baseball players.

The problem of assessing the value of the contributions made by major league batters and base runners is largely solved. If the values of the expected run matrix can be estimated with reasonable accuracy, then *RE*24 answers the question of how the team's fortunes changed when a certain batter was at the plate, relative to a league average batter. The only remaining question is to separate the contributions of base runners from those of batter, for which we have discussed several reasonable attempts (Click, 2005; Baumer et al., 2015). While it is true that analysts continue to debate which particular metric is best, the variation in the estimates of competing metrics is generally small enough that it is unlikely that we will see major changes in the way that we evaluate batters and base runners in the future. This stands in stark contrast to, for example, the dramatic changes in the evaluation of fielders and catchers that have occurred in recent years.

On the other hand, the problem of predicting the future offensive performance of players is far from settled. Most of the widely cited projection systems (e.g., Pecota, ZiPS) are proprietary, and thus we don't really understand exactly how they work. Furthermore, while these projection systems do offer better accuracy than the relatively simple Marcel, the order of magnitude of the difference is not so large as to suggest that the problem has been "solved."

Moreover, these projection systems work well for major league players, less well for minor league players, and not at all for amateur players. Thus, there is still tremendous work to be done in predicting the future offensive performance of amateur players.

The most likely source of a breakthrough here is not a brilliant model, but rather a new data source—or the combination of a new model and a new data source. In what follows we outline some open problems.

1.5.1 Open Problems

Forecasting: The forecasting problem (i.e., given $\mathbf{X}_t$, predict y_{t+1}) is particularly well defined, especially if we restrict $\mathbf{X}_t$ to include only information in the `Lahman` (or Retrosheet) database. This makes it a nice problem to tackle. Unfortunately, there is little consensus surrounding a "best" forecasting model. What is needed is a simple and/or elegant model that can consistently beat Marcel by a significant margin. It seems clear that we have models that can consistently beat Marcel, but they are significantly more complex (or proprietary—either way, we haven't learned much). It also seems likely that we have relatively simple models that can consistently beat Marcel but not by a significant margin (Albert, 2016). We suspect that bringing machine-learning tools to bear on this problem (e.g., random forests) might be a worthwhile approach.

Amateur forecasting: Forecasting the major league performance of amateur players is a fundamentally different problem, since performance data from colleges and high schools are considerably less reliable than minor league or major league performance data.

Here, we would like to see a model that incorporates traditional scouting information with phenotypic attributes and performance data. If a 6′2″ pitcher throws 90 mph in high school, and a scout describes him as having a "quick" arm, does he have a higher expected career WAR than a 5′11″ high school pitcher throwing 90 mph who is a "slinger?" Major league teams certainly have such models, but they are not known to the public.

Biometric models: Pursuant to the above, are there biometric indicators that are useful in predicting future performance, particularly that of amateurs? How does exit velocity—which is neither directly biometric nor directly a measure of performance—presage future performance?

Player allocation: There are only 9 positions on the field and 25 spots on the roster. If one assumes a free market, and that each player has a value (i.e., his expected WAR) and a cost (i.e., his salary), then what is the optimum allocation of players to positions given a fixed budget? Note that WAR is position dependent. Intuitively, it seems clear that signing the nine best hitters and forcing them to play the field is not the optimal solution, even though it would maximize offense. This is a variation on the multiple knapsack problem, which is known to be NP-hard, and thus, there is not likely to be a polynomial-time solution. Nevertheless, is the problem tractable for this size? Is there a better solution than trying all the possible combinations?

References

Albert, C. The metrics system. *Sports Illustrated*, pp. 45–48, August 22, 2016. URL http://www.si.com/mlb/2016/08/26/statcast-era-data-technology-statistics, accessed July 1, 2016.

Albert, J. Using play-by-play baseball data to develop a better measure of batting performance. Technical report, Bowling Green State University, Bowling Green, OH, September 2001. http://bayes.bgsu.edu/papers/rating_paper2, accessed July 1, 2016.

Albert, J. *Teaching Statistics Using Baseball*. MAA, Washington, DC, 2003.

Albert, J. Improved component predictions of batting and pitching measures. *Journal of Quantitative Analysis in Sports*, 12(2):73–85, 2016.

Albert, J. and J. Bennett. *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. Copernicus Books, New York, NY, 2003.

Appelman, D. Get to know: Re24, March 2008. http://www.fangraphs.com/blogs/get-to-know-re24/.

Baumer, B. Marcel the matrix, June 2014. https://baseballwithr.wordpress.com/2014/06/25/marcel-the-matrix/, accessed July 1, 2016.

Baumer, B. and P. Terlecky. Improved estimates for the impact of baserunning in baseball. In *JSM Proceedings, Statistics in Sports, ASA*, Alexandria, VA, 2010.

Baumer, B. S. Why on-base percentage is a better indicator of future performance than batting average: An algebraic proof. *Journal of Quantitative Analysis in Sports*, 4(2):1–11, 2008.

Baumer, B. S. Using simulation to estimate the impact of baserunning ability in baseball. *Journal of Quantitative Analysis in Sports*, 5(2):1–16, 2009.

Baumer, B. S., S. T. Jensen, and G. J. Matthews. openWAR: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports*, 11(2):69–84, 2015.

Baumer, B. S., J. Piette, and B. Null. Parsing the relationship between baserunning and batting abilities within lineups. *Journal of Quantitative Analysis in Sports*, 8(2):1–17, 2012.

Bennett, J. M. and J. A. Flueck. An evaluation of major league baseball offensive performance models. *The American Statistician*, 37(1):76–82, 1983.

Boswell, T. (ed.). Welcome to the world of total average where a walk is as good as a hit. In *How Life Imitates the World Series*, pp. 137–144. Penguin Books, New York, 1982.

Brown, L. D. In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics*, 2(1):113–152, 2008.

Bukiet, B., E. Harold, and J. Palacios. A markov chain approach to baseball. *Operations Research*, 45(1):14–23, 1997.

Click, J. Station to station: The expensive art of baserunning. In *Baseball Prospectus 2005*, pp. 511–519. Workman Publishing, New York, 2005.

Cockcroft, T. H. Inside the projections process, February 2015. http://espn.go.com/fantasy/baseball/story/_/page/mlbdk2k15_projectionstalk/how-fantasy-baseball-projections-calculated-how-best-use-them, accessed July 1, 2016.

Cook, E. *Percentage Baseball*. Waverly Press, Baltimore, MD, 1964.

Cover, T. M. and C. W. Keilers. An offensive earned-run average for baseball. *Operations Research*, 25(5):729–740, 1977.

Cramer, R. D. and P. Palmer. The batter's run average (B.R.A.), 1974. http://research.sabr.org/journals/batter-run-average, accessed July 1, 2016.

Cross, J. Steamer percentile projections, August 2014. https://steamerprojections.shinyapps.io/steamer_error_bars/error_bars.Rmd, accessed July 1, 2016.

D'Angelo, J. P. Baseball and markov chains: Power hitting and power series. *Notices of the AMS*, 57(4):490–495, 2010.

Eddy, M. Sabr analytics: New takes on batted-ball profiles, pitch framing, March 2015. http://www.baseballamerica.com/majors/sabr-analytics-bis-offers-new-takes-batted-ball-profiles-pitch-framing-day-three/, accessed July 1, 2016.

Efron, B. and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Fangraphs Staff. Projection systems, 2015a. http://www.fangraphs.com/library/principles/projections/, accessed July 1, 2016.

Fangraphs Staff. wSB, 2015b. http://www.fangraphs.com/library/offense/wsb/, accessed July 1, 2016.

Fangraphs Staff. 2016. Re24. http://www.fangraphs.com/library/misc/re24/, accessed July 1, 2016.

Fox, D. Circle the wagons: Running the bases part iii, August 2005. http://www.hardballtimes.com/circle-the-wagons-running-the-bases-part-iii/, accessed July 1, 2016.

Furtado, J. Introducing XR, 1999. http://www.baseballthinkfactory.org/btf/scholars/furtado/articles/IntroducingXR.htm, accessed July 1, 2016.

Guzzo, G. *Strat-O-Matic Fanatics: The Unlikely Success Story of a Game That Became an American Passion*. ACTA Sports, Skokie, IL, 2005.

Highley, T., R. Gore, and C. Snapp. Granularity of weighted averages and use of rate statistics in AggPro. In *Proceedings of the 2010 Winter Simulation Conference (WSC 2010)*, Baltimore, MD, December 5–8, 2010, pp. 1318–1329. WSC, 2010.

James, B. *The Bill James Historical Baseball Abstract*. Random House Inc., New York, NY, 1986.

James, B. and J. Henzler. *Win Shares*. STATS Pub., Northbrook, IL, 2002.

James, G., D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, New York, NY, 2013. http://www-bcf.usc.edu/ gareth/ISL/, accessed July 1, 2016.

Jensen, S. T., B. B. McShane, A. J. Wyner, et al. Hierarchical Bayesian modeling of hitting performance in baseball. *Bayesian Analysis*, 4(4):631–652, 2009. http://projecteuclid.org/euclid.ba/1340369815, accessed July 1, 2016.

Johnson, P. and B. James. Estimated runs produced, 1985. http://www.baseballthinkfactory.org/btf/pages/essays/jameserp.htm, accessed July 1, 2016.

Katsunori, A. Modified offensive earned-run average with steal effect for baseball. *Applied Mathematics and Computation*, 120(1):279–288, 2001.

Keene, C. A. Statistician steps up to the plate for red sox. The Boston Globe, March 2016. https://www.bostonglobe.com/business/2016/03/04/baseball-numbers/rfdabpRE7zJObT7Ozvc8bP/story.html, accessed July 1, 2016.

Keri, J. Q&A: MLB advanced media's Bob Bowman discusses revolutionary new play-tracking system, March 2014. URL http://grantland.com/the-triangle/mlb-advanced-media-play-tracking-bob-bowman-interview/, accessed July 1, 2016.

Larson, W. The baseball projection project, 2015a. http://www.bbprojectionproject.com/, accessed July 1, 2016.

Larson, W. 2014 projection review (updated), March 2015b. http://www.fangraphs.com/community/2014-projection-review-updated/, accessed July 1, 2016.

Lewis, M. *Moneyball: The Art of Winning an Unfair Game*. WW Norton & Company, New York, NY, 2003.

Lindbergh, B. Before beane, July 2015. http://grantland.com/features/2015-mlb-avm-systems-ken-mauriello-jack-armbruster-moneyball-sabermetrics/, accessed July 1, 2016.

Lindsey, G. R. An investigation of strategies in baseball. *Operations Research*, 11(4):477–501, 1963.

Marchi, M. and J. Albert. *Analyzing Baseball Data with R*. CRC Press, Boca Raton, FL, 2013.

McCracken, V. Pitching and defense: How much control do hurlers have? 2001. http://baseballprospectus.com/article.php?articleid=878, accessed July 1, 2016.

McShane, B. B., A. Braunstein, J. Piette, and S. T. Jensen. A hierarchical Bayesian variable selection approach to major league baseball hitting metrics. *Journal of Quantitative Analysis in Sports*, 7(4):1–24, 2011.

Meyer, D. Evaluating the 2014 projection systems, December 2014. http://www.hardballtimes.com/evaluating-the-2014-projection-systems/, accessed July 1, 2016.

MLBAM, GameDay files. http://gd2.mlb.com/components/game/mlb/, accessed July 1, 2016.

Neal, D., J. Tan, F. Hao, and S. S. Wu. Simply better: Using regression models to estimate major league batting averages. *Journal of Quantitative Analysis in Sports*, 6(3):1–12, 2010. http://www.degruyter.com/view/j/jqas.2010.6.3/jqas.2010.6.3.1229/jqas.2010.6.3.1229.xml, accessed July 1, 2016.

Null, B. Modeling baseball player ability with a nested Dirichlet distribution. *Journal of Quantitative Analysis in Sports*, 5(2):1–36, 2009. http://www.degruyter.com/dg/viewarticle/j$002fjqas.2009.5.2$002fjqas.2009.5.2.1175$002fjqas.2009.5.2.1175.xml, accessed July 1, 2016.

Pankin, M. Evaluating offensive performance in baseball. *Operations Research*, 26(4):610–619, 1978.

Pankin, M. D. Baseball as a Markov chain. In *The Great American Stat Book*, James, B. (ed.), pp. 520–524. Ballantine Books, 1987.

Pankin, M. D. Markov chain models: Theoretical background, pp. 11–26, 2007. http://www.pankin.com/markov/theory.htm, accessed July 1, 2016.

Schwarz, A. *The Numbers Game: Baseball's Lifelong Fascination with Statistics*. Thomas Dunne Books, New York, NY, 2005a.

Schwarz, A. Predicting futures in baseball, and the downside of damon, November 2005b. http://www.nytimes.com/2005/11/13/sports/baseball/predicting-futures-in-baseball-and-the-downside-of-damon.html, accessed July 1, 2016.

Silver, N. Introducing pecota. In *Baseball Prospectus 2003*, pp. 507–514. Brassey's Publishers (Dulles, VA), 2003.

Skoog, G. R. Measuring runs created: The value added approach. In *The Bill James Baseball Abstract*, James, B. (ed.). Ballantine Books, 1987.

Sportvision. HiTf/x, 2009. https://www.sportvision.com/baseball/hitfx%C2%AE, accessed July 1, 2016.

Szymborski, D. Zips Q&A, 2015. http://www.baseballthinkfactory.org/szymborski/zipsqa.rtf, accessed July 1, 2016.

Tango, T. Marcel 2012, 2012. http://www.tangotiger.net/marcel/, accessed July 1, 2016.

Tango, T. M., Lichtman, and A. Dolphin. *The Book: Playing the Percentages in Baseball*. Potomac Books, Dulles, VA, 2007.