Journal of Statistical Software

August 2019, Volume 90, Book Review 1. *doi: 10.18637/jss.v090.b01*

Reviewer: Tim Downie
Beuth University of Applied Sciences Berlin

Analyzing Baseball Data with R (2nd Edition)

Max Marchi, Jim Albert, Benjamin S. Baumer
Chapman & Hall/CRC, Boca Raton, 2019.
ISBN 978-0-8153-5351-5. 342 pp. USD 61.95.
https://www.crcpress.com/9780815353515

This second edition of *Analyzing Baseball Data with R* is a heavily revised and updated version of the first edition by Marchi and Albert (2013). In this second edition a few more chapters have been added, including some new baseball topics. The data examples have been updated, to include Major League Baseball (MLB) data from the 2016 or 2017 seasons. A major change is to the style of the R code (R Core Team 2019), which has been modernized to be consistent with the syntax style of the **tidyverse** and **dplyr** packages, for example making extensive use of the pipe operator `%>%`, and working with `tibbles` and **ggplot2** graphics.

The book is presented in the style of a course book, which could accompany an applied statistics course at advanced bachelor's or master's level. It is also well suited to self-study for people with the appropriate background. All the R code is printed in clearly identifiable light grey text-boxes and all source code and data files are available at the accompanying *GitHub* website (Marchi, Albert, and Baumer 2019). At the end of each chapter there are many exercises, allowing the reader to reinforce the methods learned in the main text.

The book however has a limited potential readership. It is a prerequisite that the reader likes and is knowledgeable about baseball, and desirable is a familiarity with basic *sabermetrics* (the name for quantitative analysis of baseball). Some baseball expressions are explained well such as *catcher framing*, but others are not, for example the relevance of a runner being hit by a batted ball. If you are planning to learn or update your R skills using practical examples but only have a moderate interest in baseball, then this book is not for you. The benefit of this specialism is that, those who do belong to this particular group can quickly get to grips with analyzing interesting and complex baseball problems. If your interest is more oriented towards the sabermetric results rather than data analysis procedures, then two other text books by Jim Albert would be more appropriate (Albert and Bennett 2001; Albert 2017).

The book divides into three main parts: the first part introduces the reader to the baseball data used in this book, the **tidyverse** and **dplyr** style of R code and gets the reader started using R Graphics. As these chapters focus on computing aspects, the baseball themes have been chosen to be relatively well known ones, which should be familiar to the readers. An example is to compare graphically the career performances of four famous home run sluggers Babe Ruth, Hank Aaron, Billy Bonds and Alex Rodriguez. The main part of the book consists of nine chapters each one concentrating on a different aspect of analyzing Baseball. The self explanatory chapter names are, in order: *The Relation Between Runs and Wins*, the *Value of Plays Using Run Expectancy*, *Balls and Strikes Effects*, *Catcher Framing*, *Career Trajectories*, *Simulation*, *Exploring Streaky Performances*, *Using a Database to Compute Park Factors*, and *Batted Ball Data from Statcast.* Readers familiar with the other works by these authors will be familiar with many of these subjects. The final part consists of three Appendices, containing detailed guides to the various data sources and a description of the variables found in these databases.

The strength of this book lies in its practical use of R. All data sets are downloaded from publicly available sources, if need be scraping data from websites. Often data from different sources are combined and where appropriate data is stored and accessed via an SQL database. A wide variety of statistical analysis techniques and visualization methods are taught and the style is consistent throughout the book. Within each chapter the code is self contained. In a couple of cases a data frame is used, that had been constructed in a previous chapter. Where this occurs, the location of the code is clearly indicated. This enables a reader to jump straight to a section of particular interest to her or him, if so desired, without having to work through all the previous chapters beforehand.

The authors claim in the preface that *the purpose of this book is to introduce R to sabermetricians, baseball enthusiasts and students interested in exploring baseball data* and this book is certainly well oriented to this purpose. However the level of statistical education expected is considerably higher than the above sentence suggests. A reader without the skills learned in a statistics bachelor's degree or equivalent will find many sections in this book difficult. Most statistical methods are "pulled off the shelf" with little or no explanation. It is not reasonable to expect that a baseball enthusiast knows what a GAM, a mixed effects model or bootstrapping is and these methods are explained in just a couple of sentences. Where such methods are new to the reader, he or she should consult other sources to avoid the danger, that a model is applied without understanding the principles behind it, resulting in an inappropriate analysis or conclusion.

In several places a statistical method or baseball term is first used with no explanation but is explained in a later chapter. For example, the *OPS* measure is used in Chapter 4, but is first defined in Chapter 6. Many similar examples suggest that the new chapters have resulted in a slight loss of continuity. The terms *launch speed*, *launch velocity*, *exit speed* and *exit velocity* are all used to mean the same thing in different places in the book. One can expect the reader to know that speed and velocity are commonly conflated, but only by comparing code between chapters, can one see that launch speed and exit speed are actually the same. The explanation of this recently developed metric is hidden away in an Appendix. Furthermore, the units of this metric are not mentioned at all in the first chapter in which it is used. Only in a diagram in a later chapter does one learn that the units are miles per hour.

The chapter on simulation could be considerably better. Half innings are simulated using a Markov chain model, using a constant transition matrix obtained from 2016 season averages. An obvious deficiency to this model is that in real baseball the batting skills decrease considerably further down the batting line up. The transition probabilities are structurally dependent on the batting position, e.g., the lead of batter has a high probability of getting on base but a relatively low probability of a home run. The second part of the simulation chapter simulates a complete 1968 season including playoffs using a simulated fixture list and a simulated, one-dimensional variable called *talent* to represent the strength of each team. Particularly confusing, is that the team names are real team names, but the talent variable bears no resemblance to the actual teams. Stating that the New York Yankees won the World Series in this simulated version is meaningless, since each of the 20 teams had an equal probability of winning. There is clearly much scope in extending these simulation models based on actual player and team data, as is done through the rest of the book.

There are a few weak points from the viewpoint of a statistician, which seem minor, but are mistakes one would expect a statistics text book to avoid. Winning percentages are given on a scale from 0 to 1. Journalists mix up the terms proportion, percent and permill, but a statistics text book should not. In several places the graphics have not been adequately proofread. In many instances the R output runs over two pages, which results in poor presentation of the results. An example is a frequency table relating to the number of games in a row without a hit for a given player. The frequencies for 3 to 12 matches appear on a different page to the frequencies for 1 and 2 matches, making the overall pattern of the frequencies difficult to perceive.

Overall, the book meets its main aim of teaching the reader to analyze real data using R. It is well suited to baseball fans, who have a solid statistical background, and want to learn R or modernize their style of R programming. Baseball fans with a more basic statistical education will also learn from this book, although they should be prepared to read more into some of the statistical methods used.

References

Albert J (2017). *Visualizing Baseball*. Chapman & Hall/CRC, Boca Raton.

Albert J, Bennett J (2001). *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. Springer-Verlag, New York.

Marchi M, Albert J (2013). *Analyzing Baseball Data with R*. 1st edition. Chapman & Hall/CRC, Boca Raton.

Marchi M, Albert J, Baumer BS (2019). "Analyzing Baseball Data with R (2nd Edition)." URL https://github.com/beanumber/baseball_R.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reviewer:

Tim Downie
Beuth University of Applied Sciences Berlin
Department II
D-13353 Berlin, Germany
Email: tim.downie@beuth-hochschule.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
August 2019, Volume 90, Book Review 1 *Published:* 2019-08-24
doi:10.18637/jss.v090.b01