Biometrics 73, 1464–1470
December 2017

DOI: 10.1002/biom.12824

BOOK REVIEWS

EDITOR:
DONNA PAULER ANKERST

BRADLEY EFRON AND TREVOR HASTIE. **Computer Age Statistical Inference**. Cambridge: Cambridge University Press.

This book tackles the task of describing how statistical approaches to estimation and inference have changed over the last 100 years. The book's subtitle "Algorithms, Evidence, and Data Science" gives a sense of what the book is about. The premise of the book is that the standard approaches from the 1990's and earlier are no longer the only approaches, and may not even work well for the types of Big Data that is routinely collected nowadays. This change is being driven by two related facts, the first is that computation is orders of magnitude faster than it used to be, facilitating much more sophisticated and complicated approaches, and second that that other fields, notably computer science, are successfully developing methods of analyzing data that are not based on probability models and do not have inference as a goal.

The book is divided into three parts, which are titled Classic Statistical Inference, Early Computer-Age Methods, and Twenty-First-Century topics. The first section explains and contrasts Bayesian, Frequentists and Fisherian philosophies and approaches to inference. The ideas and differences are illustrated using exponential family models, and concepts such as bias, efficiency, conditioning, variance, and sufficiency are explained. The authors are clear in their preference for frequentist based inference, driven in part by the view that for statistical approaches that are based on algorithms, rather than models, the other approaches are not viable.

The second part of the book is a series of chapters, each based on an important development in statistics. These include survival analysis, the bootstrap, regression trees, generalized linear models, cross-validation, Markov chain Monte Carlo, empirical Bayes, the EM algorithm, and ridge regression. The general theme is that the new methods from this era started to required more computing for the estimation and inference, although they all have sound and principled statistical justification. It was in this section of the book that I thought the authors were at their best. They brilliantly describe these methods. For readers who want a refresher, or a not too technical introduction and discussion of these methods, these chapters are great.

The third section of the book is about 21st century topics. Here, the emphasis is on statistical methodology that is more algorithmic or designed for big n or big p situations. All the methods, here, are computer-dependent, even if they were derived from statistical principles. Topics include random forests, support-vector machines, lasso, inference after model selection, and false-discovery rate. The topics are described in sufficient detail and precision and then illustrated on real examples that readers will get a solid understanding of how the methods work and what they can achieve. Although some of the methods were developed by non-statisticians, the authors focus is on explaining them in statistical terms. The chapters cover heterogeneous topics in the sense that some are about statistical inference, whereas others are only about predictions, what connects them is that they are all about analyzing Big Data, as well as that they were recently developed.

One chapter provides a partial explanation to a question that many statisticians are currently asking "What is Deep Learning?" In some sense, it is simply a rebranding of neural networks, which statisticians think of a simply non-linear regression for prediction, with a particular structure. Deep Learning is neural networks with many more bells and whistles. It has been shown to work well for some big n problems. It involves concepts and terms such as regularization, backpropagation, pooling, convolutions, principal components, autoencoding, dropout learning, accelerated gradient descent and the list goes on. Deep Learners have perhaps dozens of tuning parameters and can take a take a considerable time to train. As the authors say, this is an area where statisticians are trying to catch up, and provide a statistical basis and understanding of a method that is advancing rapidly in other fields. This chapter is a good place to start for statisticians wanting to get going with Deep Learning. I am sure there will be many articles in the near future that provide more statistical interpretation of this complicated algorithmic approach to prediction.

The book is a good read for someone in the first or second year of a PhD program to get a nice summary of the historical development of statistics. It is also a good read for more advanced researchers both as a refresher and for the many insights it provides. The book was not intended to describe new research, although there may have been some in the final chapters on Deep Learning, Empirical Bayes, and Inference after Model Selection.

In total, the book is 475 pages long, but as the authors say, individual chapters can easily be skipped. There are words of wisdom scattered throughout, so you might miss out on the considerable statistical intuition the authors provide if you skip too much.

The authors choice of topics is heavily influenced by their own impactful contributions over 80 plus years of combined research. These include survival analysis, empirical Bayes, the bootstrap, false-discovery rate, lasso, support-vector machines, and statistical learning in general. As the authors say it is neither a catalog or an encyclopedia. They say, they could have included, to illustrate their points, but did not, topics such as estimating equations, causal inference, time series, graphical models, and experimental design. There is a chapter on objective Bayes and Markov chain Monte Carlo in the middle section of the book. The authors tend to be a bit negative about Bayesian methods and favor an empirical Bayes version, and provide a rationale for this. If I had been suggesting chapters for the third section, where the emphasis is on algorithms and Big Data and looking to the future, it would have been one that illustrated the considerable advances in Bayesian computation in the last few years, beyond MCMC. These newer approaches can, for a moderately broad set of situations, provide computationally scalable estimation and inference, while retaining their principled statistical basis.

There were two "triangles" in the book, which were fun and thought provoking. One triangle was about modes of inference, with Bayesian, Frequentist and Fisherian at the three corners, and methods such as the EM algorithm, the bootstrap, ridge regression, generalized linear models, regression trees and others placed throughout the triangle. Readers might have fun seeing where they would place these methods and their own favorite other ones too. The corners of the other triangle are Applications, Mathematics, and Computation. They use this to illustrate the development of statistics, from applications only around 1900 and earlier, then moving directly towards the mathematical corner via Fisher's work, Neyman–Pearson and decision theory in 1950, then turning towards some combination of applications and computation through Tukey's focus on data, the Cox model, the bootstrap, MCMC, penalization methods, random forests and multiple testing methods. The authors do not try to speculate on whether statistics will move more in the direction of applications or computation, but it has clearly moved away from being a branch of mathematics.

The authors note there was a coherent inferential theory based on logic, mathematics, and probability in the development of statistics. They end the book by writing "A hopeful scenario for the future is one of increasing overlap (between applications and computation) that puts data science on a solid footing while leading to a broader general formulation of statistical inference."

I thoroughly enjoyed reading the book. It was a nice refresher and it was thought provoking. I am glad, I have a copy.

Jeremy M. G. Taylor
Department of Biostatistics
Ann Arbor, Michigan, U.S.A. jmgt@umich.edu

JULIAN J. FARAWAY. **Extending the Linear Model with R: Generalized Linear, Mixed Effects, and Nonparametric Regression Models, 2nd edition**. Boca Raton: CRC Press.

It has been a great pleasure to review this book, which delivers both a readily accessible and reader-friendly account of a wide range of statistical models in the context of R software. Since the publication of the very well received first edition of the book, R has considerably expanded both in popularity and in the number of packages available. The second edition of the book takes advantage of the greater functionality available now in R, and substantially revises and adds several new topics.

Based on the preface of the book, changes compared to the first edition include the following: Coverage of binary and binomial responses was substantially expanded from one to three chapters (Ch. 2–4), including sections for proportion responses, quasibinomial and beta regression, and applied considerations regarding these models. New sections on Poisson models with dispersion and zero inflated counts were added to count regression (Ch. 5), correspondence analysis to contingency tables (Ch. 6), linear discriminant analysis to multinomial responses (Ch. 7), and sandwich, robust, and Tweedie-proposed estimation to generalized linear models (GLMs) (Ch. 8 and 9). Chapters 10 and 11 on random effects and repeated measures were substantially revised to reflect changes in the (Bates et al., 2015) lme4 package and to show how to perform hypothesis testing for the models using various methods. A new Chapter 12 on the Bayesian analysis of mixed effects models illustrates STAN (Stan Development Team, 2016) and INLA (Rue, Martino, and Chopin, 2009), which are used to compute posterior densities by employing simulation and approximation based methods, respectively. A revised Chapter 13 on GLM's reflects the much richer choice of R packages now available, Chapter 14 updates coverage of splines and confidence bands, and Chapter 16, random forests. R code throughout the book has been overhauled, in most cases with the state-of-the-art ggplot2 package (Wickham, 2009).

Every chapter is now followed by a revised and expanded set of exercises and references have been substantially updated. A well-maintained website for the book is available at `http://people.bath.ac.uk/jjf23/ELM/`, and contains information on how to install the faraway package in R to access the datasets, the errata, and the R commands used in the text. STAN and INLA implementations of the linear mixed models are also available.

Overall, this is a nicely written reference book, encompassing an overview of a wide range of several modern statistical models and their application to real data. It covers a lot of topics, including discussion of overdispersion, model diagnostics, the use of appropriate models in the context of a research question and study design. The author acknowledged that the book is not about learning R. It can be found very useful, however, for applied statisticians and for teaching purposes for graduate students in statistics. The author should be commended on what he achieved by preparing this second edition of the book.

References

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Stan Development Team (2016). *Stan Modeling Language Users Guide and Reference Manual, Version 2.12.0.* `http://mc-stan.org`

Rue, H., Martino, S., and Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* **71**, 319–392.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer.

Andrzej Galecki
Department of Biostatistics
School of Public Health
Institute of Gerontology
Medical School
University of Michigan
Ann Arbor, Michigan, U.S.A
agalecki@umich.edu

YIN BUN CHEUNG. **Statistical Analysis of Human Growth and Development**. Boca Raton: CRC Press.

The author, Yin Bun Cheung has been part of studies in human growth and development in African and Asian countries for about fifteen years. Thereby, he has recognized the lack of statistical analysis books with an emphasis on this specific field, making his book "Statistical Analysis of Human Growth and Development," a useful contribution to fill this gap. Numerous examples from different experiments and studies throughout the book help the reader to understand the various concepts and methods. Additionally, two simulated datasets are provided, which can be easily reproduced in Stata due to code in the appendix. Code and macros can also be found on the book's CRC Press Web page.

The book has 15 chapters, which consecutively build upon one another. In the beginning, the author gives a brief overview of the book and explains the content and chronology of the various chapters to orient the reader. The first chapter is an introduction to the subject of human growth and development. It delivers an extensive background and is the basis for the following sections. The second chapter deals with study design and distinguishes association and causality. Furthermore, it advises the reader over considerations before working with real data. Subsequently, basic statistical concepts and techniques are covered in Chapters 3 and 4 provides a quantification of growth and development with existing tools. Different standards are introduced and the author explains why they are defined as they are. Chapters 5 through 7 deal with regression analysis of quantitative, binary, and censored outcomes, respectively. Thereby, the author's emphasis on human growth and development becomes more focused as less common statistical analyses are employed, including quantile regression and analysis of interval-censored data. The following Chapter 8 addresses repeated measurements and clustered data. In Chapters 9 and 10, new references for quantifying growth and development, respectively, are developed. Further norms are established for longitudinal data in Chapter 11. Chapter 12 focuses on the concepts of validity and reliability of measurement tools and considers different statistical methods for assessing them. The final three chapters address challenges of real-world research practice. Chapter 13 concerns missing values and imputation, Chapter 14, problems arising from multiplicity, and Chapter 15, the case of many available variables warranting selection.

In summary, I recommend the book for anyone interested in dealing with data of human growth and development, as it extensively describes how to approach a problem in this field. The book is addressed to researchers and postgraduate students without a broad statistical education. Therefore, basic statistical concepts necessary for following subsequent chapters are explained at the beginning, and the multitude of examples facilitate understanding and application. Nevertheless, the book is an introduction to the field of human growth and development rather than an introduction to statistics. A slight drawback is that the commercial software Stata is used for the book, whereas the open source R package is more widely available.

Johanna Straubinger
Mathematics Life Science Unit
Technical University of Munich
Munich, Germany
johanna.straubinger@tum.de

C. D. COMBS, JOHN A. SOKOLOWSKI, and CATHERINE M. BANKS. **The Digital Patient: Advancing Healthcare, Research, and Education**. Hoboken: Wiley.
CHANDAN K. REDDY and CHARU C. AGGARWAL. **Healthcare Data Analytics**. Boca Raton: CRC Press.

During recent years, medical informaticians designing digital healthcare systems promise the delivery of *real world evidence which will advance healthcare, research, and education.* Both books are excellent starting points to explore what this promise is about and what it can achieve. They introduce technological advancements which have paved the way for a more systematic collection, modelling, simulation, and visualization of data related to human health outside the biostatistical community. *The Digital Patient* presents in 21 chapters over 305 pages its version of *The Vision*, the *State of the Art,* related *Challenges*, and the *Potential Impact.* The 700-page volume *Healthcare Data Analytics* is also organized into 21 chapters and three parts: *Healthcare Data Sources and Basic Analytics, Advanced Data Analytics for Healthcare, and Applications and Practical Systems for Healthcare.*

What is *The Digital Patient* about? Vanezza Diaz-Zuccarini explains this concept as the digital representation of health coupled with a sophisticated decision support system, tailored to the individual subject. This vision starts with cellular processes (chapters 7–11) and ends in public health issues by analyzing complex interventions and their impact on individual health (chapters 12–14). The editors focus on two topics: Simulation and visualization. Simulations allow integrating individual health data (sensor based, see chapter 6) into specific models and to predict individual health; visualization is seen as the main tool to inform a person on his/her health prospects. Visualization and simulation interplay by creating virtual patients for clinical training (chapter 18). Examples of personalized computational modelling are given (chapters 8, 10, and 11). In chapter 17, Joseph A. Tatman and Barry E. Ezell propose *Bayesian Networks* as the methodology of choice to operationalize complex modelling concepts. But the big challenges related to these visions are often hidden in this book and made explicit only on a few pages. They relate to the issue of how to represent health in a digital form. To approach this problem, the section *Standards of Medical Informatics* of chapter 9 is of high relevance. It describes the standardization efforts at various levels necessary to make health information linkable and computable: *open*EHR, SNOMED CT, LOINC, HL7, FIHR, and OBO. Chapter 16 is also dedicated to the problem of data integration across many disciplines. For biostatisticians who are interested to close the gap between their field and medical informatics, these two chapters provide the most relevant content. There is no chapter on data quality and quality of evidence derived from the Big Data projects presented in this book.

The volume *Healthcare Data Analytics* by Reddy and Aggarwal is more technical and gives a comprehensive introduction to fundamental principles, algorithms, and applications of health data acquisition, processing, and analysis. It starts with a survey on electronic health records (EHR), a central instrument for collecting heath data and putting these data into context. The next chapters present biomedical image data, sensor data, genomic data, and the processing of clinical text by natural language processing (NLP). Further relevant sources of health data are the biomedical literature and social media. Chapter 10 is on clinical prediction models and offers the classical biostatistical toolbox. Over the next three chapters, more complex models based on longitudinal, spatial, and high-dimensional data are discussed. The presentation uses the machine-learning perspective but offers many references from the biostatistical literature. Chapter 14 discusses *information retrieval for healthcare.* Its overall goal is to find content which meets information needs. The interplay of two processes determines the success of information retrieval: *Indexing* assigns metadata to content items, *retrieval* produces content items based on the user's query. Evaluation strategies for these processes are also discussed. My favorite part of the book is chapter 15 *privacy-preserving data publishing methods in healthcare.* The references also cite the work of Stephen E. Fienberg, a statistician who pioneered research on disclosure limitation for categorical data, and on privacy and confidentiality more generally. Steve, as known by colleagues and friends, passed away recently, leaving a legacy in this timely area among many others. The chapter gives a glimpse on a huge field of activities related to secure data sharing and secure computing: How can we improve models without getting to know the data used to build our model? Part III of the book consists of chapters 16–21 and provides visions and examples. It starts with the vision of the interconnected world of intelligent health services, followed by the issue of fraud detection in healthcare, new strategies for pharmacoepidemiology and pharmacovigilance, clinical decision systems, computer-assisted medical image analysis, mobile imaging and analytics for Biomedical Data. As in the first book, there is no chapter on data quality and quality of evidence derived from Big Data when the new IT technologies are put to work.

Both books celebrate the concept of *real world evidence.* Terms like clinical epidemiology, confounding, credibility of data, bias, and design are not used. This may be an indicator that communication between biostatistics and medical informatics communities offers many new opportunities. Both books lay open a vision of scientific activities which is beyond the framework of randomized trials and cohort studies: Innovation is coming from masses of observational data subdued by machine-learning techniques. *The Digital Patient* addresses a hopeful *Big Data* undertaking which combines systems biology, multiscale simulation, and an open sharing of data between individuals and researchers. The *Healthcare Data Analytics* analyzes recent developments in healthcare computing and discusses emerging technologies without references to biostatistics. Both books miss the disclaimer: Data taken at face value can be dangerous.

Both books are on *health care data science*, a science which is in the process of getting shaped. There should be one version of *health care data science* which is driven by at least three powerful communities: the medical informaticians, the bioinformaticians, as well as the biostatisticians. Biostatisticians should add chapters to both books which present the progress achieved to date in the analysis of large observational data. I do see in both books a great chance to rethink the cooperation with a field of scientific activities that needs our attention and contributions.

Ulrich Mansmann
Institute for Biometry and Epidemiology
University of Munich
Munich, Germany
ulrich.mansmann@lmu.de

STEPHEN W. LOONEY and JOSEPH L. HAGAN. **Analysis of Biomarker Data: A Practical Guide**. Hoboken: Wiley.

John Wiley and Sons have published another very practical and useful textbook for applied statisticians, "Analysis of Biomarker Data: A Practical Guide," by Stephen W. Looney and Joseph L. Hagan. Importantly, in this era in which biomarker research seems ubiquitous and the term "biomarker" is used in many ways, the focus of the book is on biomarkers of exposure, which are studied for their association with a clinical outcome of interest. This is in contrast to biomarkers used for disease diagnosis or screening, disease risk prediction or prognosis, or prediction of treatment response, which are not considered.

The book is targeted at applied researchers, and assumes very little statistical training. Therefore, it is most useful for the non-statistician or junior biostatistics/statistics student. The focus is on application of methods for data analysis. The concepts and methods are illustrated throughout with real data, mostly borrowed from the published literature. Problem sets are included along with solutions in SAS, and occasionally in R.

Following an introductory Chapter 1, Chapters 2 and 3 provide basic guidance on study design and data analysis for studies of association. The content is quite general, not specific to exposure biomarker studies-and covers what might be a typical Statistics 101 course. The concept of statistical distributions; measures of location and spread; basic descriptive devices; elements of inference including point estimates, confidence intervals, and p-values; methods for inference about means and correlations; linear regression; and methods for associating categorical variables, are covered. Useful recommendations, such as "designing the analysis" as part of the study design, are included. I wished that Chapter 2 would have integrated the literature on two-phase study designs, which may be particularly relevant for exposure biomarker studies in which the biomarker is difficult or expensive to measure.

Chapter 4 considers common challenges encountered in data analysis. Many topics covered are not specific to biomarker studies, such as assessing distributional assumptions and using data transformations, robust statistical procedures, accommodating heterogeneity in variance, dealing with dependent data, and detecting and contending with outliers. Especially important for exposure biomarker studies is the content on non-detected, missing, or censored biomarker measurements.

Chapter 5 is an important chapter on biomarker validation. The authors distinguish between assessing a biomarker's *reliability*, which reflects whether the results of the measurement procedure can be replicated, and a biomarker's *validity*, which reflects whether the biomarker measures what is intended. For assessing reliability, the chapter describes suitable measures of agreement for binary, nominal categorical, ordinal, and continuous measures; these are then applied to assessing intra-rater and intra-observer agreement. There is a brief discussion of study design. For assessing validity, there is a basic description of methods for assessing a biomarker's classification accuracy (true- and false-positive rates and predictive values; Receiver-Operating-Characteristic analysis). The assumption is that interest lies in what might be termed the "pure statistical" validity of the biomarker; the content does not speak to validation of the biomarker for clinical decision making. There is a discussion of methods for assessing validity in the absence of a gold standard outcome, which focuses on measuring agreement among multiple exposure measurements.

All told this book is broad in scope and will likely have practical utility for applied researchers studying exposure biomarkers.

Holly Janes
Vaccine and Infectious Disease and
Public Health Sciences Divisions
Fred Hutchinson Cancer Research Center
Seattle, Washington, U.S.A.
hjanes@fredhutch.org

RICHARD M. HEIBERGER and BURT HOLLAND. **Statistical Analysis and Data Display**. Heidelberg: Springer.

The title of the book "Statistical Analysis and Data Display–An Intermediate Course with Examples in R" immediately caught my eye as a biostatistician who uses the statistical software package R as part of daily work. For those not familiar with R, it is contributed openly by scientists, has been around for quite some time now and is a must-have for academic and practicing statisticians. Many books have only a limited number of examples or only show general functions in R. The authors of this book cover many different statistical topics and after almost every chapter, provide concrete examples of how to perform the corresponding calculations in R.

The book explains basic knowledge of R, such as "what is a package" to advanced knowledge, such as "writing and building your own packages." It even shows examples about R Shiny. It is possible with R Shiny to build interactive web pages and use R through a user-friendly interface. One does not necessarily need any knowledge about HTML/CSS or JavaScript. After some introduction and motivational examples, the book starts in Chapter 2 with a general explanation of different data types, data presentation and analysis with missing values. Chapters 3 through 9 cover an introduction to probability theory, including probability distributions, power, hypothesis testing, linear regression, and multiple regression, with nicely formatted graphics created in R. Each of these chapters includes practical working examples with the corresponding R program code.

What I especially liked about this book is that one can reproduce the examples easily by using the R package HH. The authors included code for all figures and tables in the book. To find the code for the corresponding chapter, one simply has to type the command "HHscriptnames(chapter number)" in the R console and R will return the file path to the program code. One can then use the code and change it to their needs. This makes it easy and efficient to find the specific code you are looking for.

In the Appendix, the book includes a description of how to use the HH package with the R Commander. R Commander is a platform-independent basic-statistics GUI (graphical user interface) for R. This is particularly helpful for those readers not familiar with R coding or programming in general and want to reproduce the examples. Another great part of this book is the idea of interactivity. One can explore the graphics in the book interactively in 3D with the help of the web-application framework R Shiny. I hope that more books in the future will combine the power of freely available interactive tools such as these.

Overall, the book offers a good introduction to statistical methods and is helpful for researchers to understand and pick appropriate models and methods for their analysis.

Martin Goros
Department of Epidemiology & Biostatistics
University of Texas Health Science Center at San Antonio
San Antonio, Texas, U.S.A.
goros@uthscsa.edu

SIMON MUNZERT, CHRISTIAN RUBBA, PETER MEISSNER, and DOMINIC NYHUIS. **Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining**. Hoboken: Wiley.

The amount of data produced on or for the internet every day has increased rapidly during the last decade. To meet the demand for real-time processing of such data, several techniques have been developed to extract valuable information in a time efficient way. Fortunately, many of them are implemented in my favorite programming language for statistical analysis–R. As the freeware R is also a first choice for many other scientists, the book *Automated Data Collection with R–A practical Guide to Web Scraping and Text Mining* provides very useful insight into the connection between R and web scraping.

It is structured into an introduction and three parts with 17 chapters in total. While the authors spare us another introduction to R, as this is already covered in multiple great books, they provide a good introduction to essential web-technologies in the first part. The focus, here, is on the basics needed for web scraping, and readers who are already familiar with HTTP, HTML, XML, JSON, AJAX, and SQL as well as XPath and regular expression can easily skip this part of the book or use it to refresh their knowledge. Exercises help to consolidate the fundamentals and some of the solutions are provided online. Although substantial material is covered in chapters concerning each of the technologies, it remains an introduction and further insight is often needed for particular problems. Sources for additional information are given in the book, helping one to master their own web scraping problems.

In the second part the comprehensive workflow of web scraping is covered in detail. Several retrieval scenarios and strategies for extracting information are discussed and compared, and advice for good practice is given together with an illustration of legal boundaries for scraping the web. The chapter on statistical text processing demonstrates the cleaning and analysis of textual data, however, only an overview over several learning techniques is given without a detailed description. A chapter on the management of data projects with automatization techniques and an introduction to task scheduling concludes the second part of the book.

The last part contains several case studies from different topics. These are an essential part of the book as the reader can see the introduced tools in action and get a good idea of different possibilities for web scraping. The examples provide a practical and detailed insight into web scraping tasks and the processing of textual data. An overview over the case studies and the main techniques which are used is given in the beginning, thus the reader can easily find case studies most related to their own scraping scenarios.

In general, the authors provide many examples in all chapters, which helps to understand the techniques. They also compare different approaches and emphasize disadvantages and advantages. One limitation of the book is that the R packages discussed have since publication been updated or that more advanced packages are now available. Hence, not all code snippets provided for illustration work and easier or faster implementations are available. Corrected versions can be found in the errata section of the website accompanying the book and the authors also plan to cover more recent packages, as for example *rvest* in the next edition. Nevertheless, the book is more than a collection of R code. It explains the ideas and technologies behind web scraping and thus after those are understood, the new packages can be used easily.

Finally, the book provides a good introduction that I can highly recommend to users with R experience who want to get started with web scraping. On the way users will also learn very nice features of R, which improved my programming experience beyond web scraping tasks.

Katharina Selig
Department of Mathematics
Technical University Munich
Munich, Germany
Katharina.selig@tum.de

KEN KLEINMAN and NICHOLAS J. HORTON. **SAS and R: Data Management, Statistical Analysis, and Graphics**. Boca Raton: CRC Press.

This book is not only an excellent cross-reference for SAS or R users to find the corresponding code in the opposing language, but also a useful resource for readers to learn statistical programming in both systems.

The book is organized into 12 chapters covering a wide range of programming and statistical topics, with both SAS and R code presented for all tasks. It begins with basic programming for data input, output, and management in Chapters 1 and 2. Chapter 3 reviews key probability, mathematical, and matrix functions, and Chapter 4, programming functions and interactions with the operating system. Chapter 5 describes how to generate univariate summary statistics for continuous variables, such as means, variances, and quantiles, display and analyze frequency tables and cross-tabulations of categorical variables, and perform a variety of one and two sample tests, such as the Student t-, chi-square, and Cochran–Mantel–Haenszel tests. Chapter 6 discusses ANOVA and common tasks in linear regression (model fitting, tests, contrasts, linear functions of parameters, parameters and results, and model diagnostics). Chapter 7 further discusses regression generalizations and modeling, including other types of outcome variables, longitudinal and clustered analysis, and survival methods.

After providing a compendium of graphical displays in Chapter 8, Chapter 9 reviews how to annotate graphical displays and change defaults to create publication-quality figures, as well as details regarding how to output graphics in a variety of file formats. Chapter 10 describes the simulation of data in a variety of common settings as well as their applications. Chapter 11 covers some special topics that statisticians may encounter in daily practice, including processing by group, simulation-based power calculations, reproducible analysis and output, Bayesian methods, propensity scores, bootstrapping, missing data, and finite mixture models with concomitant variables. The last chapter, case studies, demonstrates the examples of some data management tasks, reading more complex files, creating maps, scraping data from web pages, manipulating larger datasets, and an optimization problem.

This book also includes the introduction to SAS and R in Sections A and B to guide beginners to learn these two systems. The datasets and SAS and R code used in the book can be downloaded from the book website, which allows readers to try the code and obtain hands-on experience. Alternative ways are often presented for carrying out a task such that readers can learn diverse and elegant solutions.

This book is a great resource for users who have a long experience in only one system and need to use the other system. The SAS index at the end of the book is particularly of help for SAS users to look up a task for which they know the SAS code and turn to a page with that SAS code as well as the associated R code. And the R index in the book is used the same way by R users to find the corresponding SAS code for a task.

Xulei Liu
Department of Biostatistics
Vanderbilt University
Nashville, Tennessee
Xulei.liu@vanderbilt.edu