Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Book Reviews

To cite this article: (2018) Book Reviews, Journal of the American Statistical Association, 113:521, 480-485, DOI: 10.1080/01621459.2018.1444852

To link to this article: https://doi.org/10.1080/01621459.2018.1444852

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 521, 480–485, Book Reviews
https://doi.org/10.1080/01621459.2018.1444852

Taylor & Francis
Taylor & Francis Group

Absolute Risk: Methods and Applications in Clinical Management and Public Health. Ruth M. Pfeiffer and Mitchell H. Gail. Boca Raton, FL: Chapman & Hall/CRC Press, 2017, xxiv + 202 pp., $69.95 (H), ISBN: 978-1-46-656165-6.

Absolute risk of an event, defined as the probability of occurrence of the event over a stated time period, is an important concept in patient counseling, clinical management, and public health. Written by two leading experts in the field, this book provides a comprehensive overview of absolute risk, including both theoretical basis and clinical implications before and after the disease diagnosis. Equipped with sufficient technical details on the estimation and inference of absolute risk as well as a range of real examples, this book is targeted toward a broad audience, including epidemiologists, clinicians, and statisticians.

While a few other books on theoretical aspects of absolute risk are available in the literature, the book by Pfeiffer and Gail treats absolute risk from several new angles including (1) the estimation and inference of absolute risk under various sampling designs, (2) the use of disease registry data, (3) criteria for assessing the performance of risk models and for comparing different risk models, and (4) a collection of applications including examples in genetic studies. The book consists of 10 chapters. Through some real stories in patient counseling and examples of risk models for incidence of diseases including breast cancer, prostate cancer, and coronary heart disease, Chapter 1 provides an overview of absolute risk and discusses its importance in making clinical decisions or establishing policies for disease prevention. This chapter also discusses the clear distinction between absolute risk and relative risk.

Chapters 2 and 3 describe basic concepts for survival data and competing risks data. Various censoring and truncation schemes, such as right censoring and left truncation, are described. In the presence of competing risks, Chapter 2 describes the concept of cause-specific hazard, cumulative incidence function or absolute risk, as well as their relationship. The definition of absolute risk of an event from a particular cause within a time interval is formally provided in Equation (3.5). Nonparametric estimation and inference of absolute risk under right censoring are provided. The remarks on the difference between the pure probability of an event and its absolute risk are particularly important as they establish why absolute risk is more relevant to clinical management. These two chapters will benefit readers with limited experience in survival analysis.

Chapter 4, upon which Chapters 5–9 are built, is the most important and technical chapter of the book, in my opinion. This chapter discusses the regression models for absolute risk estimated from cohort data under various designs, including cohort design, case-cohort design, nested case–control design, and complex survey designs. Two types of semiparametric estimation methods of absolute risk are described. One is based on cause-specific hazard regression and the other is based on cumulative incidence regression. Readers will particularly appreciate the detailed discussion on the pros and cons of each method. While cause-specific proportional hazards model and proportional sub-distribution hazard model receive the most attention, alternative hazard models and cumulative incidence regression models are discussed. Technical details of variance estimation for the absolute risk estimators are provided.

Chapter 5 extends the cause-specific hazard regression methods described in Chapter 4 by combining case–control or cohort data with disease registry data; for example, the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) data. The strengths and limitations of using registry data are described. This chapter discusses the relationship between attributable risk, composite age-specific incidence (obtained from registry data), and baseline hazard as well as the estimation and inference of absolute risk using the combined data.

Chapters 6 and 7 are devoted to assessing and comparing the performance of risk models. Several criteria for assessing the performance of a risk model are described. These include methods using external validation data and criteria such as calibration, predictive accuracy, classification accuracy, discriminatory accuracy, and reduction in expected loss or increase in expected utility. Methods comparing two risk models based on the receiver operating characteristic (ROC) curve and its summary measures such as the area under the curve (AUC) or partial area under the curve (pAUC) are discussed. A related topic not covered in these two chapters is model diagnosis. Discussing methods for checking assumptions of the risk models under consideration as well as the impact of violation of model assumptions on the estimation of absolute risk can potentially improve the presentation of the book.

Chapter 8 focuses on how to build and update relative risk models. Specifically, this chapter covers three issues for modeling absolute risk: covariate selection, missing covariate data, and updating previously well-established risk models by adding new covariates. One method is particularly of interest, especially in the era of "big data." The so-called constrained maximum likelihood estimation (CML) method (see Qin 2000; Chatterjee et al. 2016) builds regression models based on individual-level data from an internal study while using information extracted from an external big-data source.

Chapter 9 discusses risk estimates from family-based studies concerning genetic effects on the disease of interest. Kin-cohort studies and studies of families with one or more diseased members are considered. Issues arising from these family studies including residual correlations and ascertainment bias are discussed. This chapter also describes how to combine risk estimates from family-based case–control studies with population-based incidence rates. Finally, Chapter 10 presents several related topics such as missing or misclassified information on prognosis following disease onset, event type, and time-varying covariates. Practitioners will find Chapter 10.5 particularly useful: it describes the application of risk models for counseling individuals and for public health strategies for disease prevention.

One appealing feature of the book is that each chapter contains a range of real applications. Although links to software are provided in several instances, it would help the readers to understand the methodologies further and to apply them in their own research if some of the datasets and code for analyzing the data were available. Overall, this well-written monograph provides a thorough introduction to methods for estimation and inference of absolute risk under various designs and its applications

in clinical management and public health. The book is a welcome addition to the literature of survival analysis and is useful for both statisticians and practitioners.

References

Chatterjee, N., Chen, Y. H., Maas, P., and Carroll, R. J. (2016), "Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information from External Big Data Sources," *Journal of the American Statistical Association*, 111, 107–117. [481]

Qin, J. (2000), "Miscellanea. Combining Parametric and Empirical Likelihoods," *Biometrika*, 87, 484–490. [481]

Guoqing Diao
George Mason University
http://orcid.org/0000-0001-7304-9591

Handbook of Big Data. Peter Bühlmann, Petros Drineas, Michael Kane, and Mark van der Laan, eds. Boca Raton, FL: Chapman & Hall/CRC Press, 2016, xvi + 464 pp., $129.95 (H), ISBN: 978-1-48-224907-1.

This book describes modern approaches to the analysis and understanding of big data. It covers many important aspects of big data analysis with a good balance between statistical methodology, theory, and application. The book consists of eight sections. The first section covers general perspectives on big data. Some exploratory methods are presented in Section II. Section III introduces some efficient algorithms to deal with the computational challenges. Graph and regularization approaches are discussed in Sections IV and V, respectively. The book also briefly touches ensemble methods, causal inference, and targeted learning in the last three sections.

The book strikes a great balance between the breadth and depth of recent research-active topics. It is an excellent reference book to keep for both academic researchers and industrial practitioners. It is also a good reference book for whoever teaches in the area of big data analysis.

Yichao Wu
University of Illinois at Chicago

Handbook of Neuroimaging Data Analysis. Hernando Ombao, Martin Lindquist, Wesley Thompson, and John Aston, eds. Boca Raton, FL: Chapman & Hall/CRC Press, 2016, xxxix + 661 pp., $139.95 (H), ISBN: 978-1-48-222097-1.

There have been enormous advances in the field of brain research in recent years, and technological developments have created a need for new statistical methodology. The *Handbook of Neuroimaging Data Analysis* is a timely and thoughtful collection of introductory material and advanced methods, which addresses this need. It contains 21 chapters from approximately 50 authors. These chapters are primarily written by statisticians, but the book is nicely balanced by contributions from biomedical engineers, psychologists, and cognitive scientists. Similar to other works in this CRC series—*Handbooks of Modern Statistical Methods*—this book is an assemblage of mostly self-contained chapters rather than a textbook in which the chapters build upon each other. However, it could be used in a special topics course for biostatistics and statistics graduate students. A reader interested in a particular imaging modality or analytical method can quickly delve into the topic.

The intended audience is statisticians and quantitative researchers interested in methodological development. Researchers in psychology and the behavioral sciences with less training in statistics may find some chapters challenging. However, I feel that a number of chapters are accessible beyond the book's intended audience, including but not limited to the introductory chapters on structural magnetic resonance imaging (MRI), clinical MRI, functional MRI (fMRI), and electroencephalography (EEG; Chapters 3, 5–7). A number of chapters list gaps in the current methodology and suggest future research directions, which is helpful for researchers and graduate students. For the most part, the *Handbook* does not provide instructions for using software to conduct neuroimaging data analysis, instead focusing on the methodological literature. Thus, I would not recommend it to users interested in hands-on instruction into statistical and neuroimaging software. There are some exceptions to this: for example, Chapter 5 includes commands for R as well as the neuroimaging programs FSL and ANTS, and other chapters do note software resources.

The *Handbook*'s main strengths are that it provides a broader treatment of different imaging modalities with greater rigor than other books on similar topics, and thus fills an important gap. Generally speaking, other books in this field focus on a single modality, often fMRI (e.g., Poldrack, Mumford, and Nichols 2011). In contrast, the *Handbook* covers different types of MRI, including structural (Chapter 3) and diffusion (Chapter 4), as well as other technologies such as positron emission tomography (PET) (Chapter 2) and EEG (Chapter 7). Textbooks in fMRI are generally written for behavioral scientists with less mathematical training and thus tend to shy away from mathematical notation. In contrast, the *Handbook* presents material at a level commensurate with its intended audience. I should note that there are no theorems or proofs, which seems appropriate given its emphasis on statistical practice. The chapters vary somewhat in their quality, where some are carefully edited without errors, whereas there are minor typos or undefined notation in some parts, and occasionally chapters contain overlapping content. Overall, I found the presentation to be clear and well-executed.

Chapters 1–7 introduce a number of popular imaging modalities, while later chapters contain more detailed information. Rather than provide an overview of the many topics, I will highlight a few. Chapter 7 provides a nice overview of EEG data, including a description of the neurophysics of volume conduction and the electrical signals from synaptic potentials. Chapter 8 provides background on MRI physics and details

how the original observations are in fact complex-valued data measured at spatial frequencies. Compared to other textbooks, Chapter 11 provides a more thorough review of methods to estimate the hemodynamic response function in fMRI, including parametric and nonparametric approaches, nonlinear models motivated by physiology, and frequency-domain approaches.

The neuroimaging community widely recognizes multiple testing problems, which are particularly severe in MRI studies in which tests are often conducted at hundreds of thousands of locations. Chapter 13 provides an overview of controlling the family-wise error rate and false discovery rate, and it includes high-level descriptions of random field theory (RFT) and permutation tests. I personally would have liked to see details on the mathematical underpinnings of RFT-based inference and more information on permutation tests, although some information is provided in other chapters (Chapters 9 and 12). Recently (after this chapter was written), cluster-based inference using RFT has drawn a lot of attention with evidence of up to 70% false positives (Eklund, Nichols, and Knutsson 2016), and permutation tests are arguably becoming the standard approach. This is a rapidly changing area in the practice of neuroimaging data analysis.

An important question is how to model changes in the brain. Chapter 18 provides an excellent overview of methods for estimating trajectories, including tools from functional data analysis. There is a clear and illuminating discussion on when cross-sectional data can mischaracterize mean trajectories. This chapter also nicely illustrates some statistical complications that can arise when estimating peaks in growth trajectories, for example, instances where the standard second derivative penalty when using B-splines appears to over-smooth towards linear fits.

Readers interested in nonstationarity and frequency-domain analyses will find Chapter 21 particularly interesting. The authors describe advanced methods for analyzing nonstationary EEG, and the ideas are well-motivated by data from an epileptic subject. Whereas many change-point methods focus on the mean or variance, this chapter also develops sophisticated measures based on changes in spectral properties. The chapter includes a nonparametric frequency-specific change-point detection method, which appears to detect changes in cross-coherence shortly preceding the onset of a seizure.

There is a quote in the chapter "Neuroimage Preprocessing," written by the scientists Stephen Strother and Nathan Churchill, that bears repeating: *It seems likely that this research field would benefit significantly from insights provided by an influx of researchers from the broader statistical community. In this view, we agree with Keith Worsley that neuroimaging may be statistics' "agricultural field trials of the 21st century."* Indeed, the *Handbook* reveals that this is a fertile area, and I recommend this book to statisticians interested in learning about neuroimaging and contributing to its growth.

References

Eklund, A., Nichols, T. E., and Knutsson, H. (2016), "Cluster Failure: Why fMRI Inferences for Spatial Extent have Inflated False-Positive Rates," *Proceedings of the National Academy of Sciences*, 113, 7900–7905. [483]

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011), *Handbook of Functional MRI Data Analysis*, Cambridge: Cambridge University Press. [482]

Benjamin B. Risk
Emory University
http://orcid.org/0000-0003-1090-0777

Modeling Discrete Time-to-Event Data. Gerhard Tutz and Matthias Schmid. Switzerland: Springer International Publishing, 2016, x + 247 pp., $89.99 (H), ISBN: 978-3-31-928156-8.

Multiple texts covering methods for the analysis of continuous censored survival data have been written. Seminal books in the field, ranging from most applied to most rigorous, include Hosmer, Lemeshow, and May (2008), Klein and Moeschberger (2003), Kalbfleisch and Prentice (2002), and Fleming and Harrington (2011). In most cases these texts do not go into great depth regarding the analysis of discrete survival outcomes despite such data commonly arising in public health, economics, political science, and the social sciences.

Modeling Discrete Time-to-Event Data provides further exposure to the area. The book is quite clear throughout and considers topics at roughly the level of Klein and Moeschberger (2003). However, to get the most out of the text, the reader needs a solid foundation in continuous-time survival methods, generalized linear models, and statistical learning methods. Generally speaking, the text would be suitable for second year or above masters students in statistics or biostatistics. In addition to model development and description, the text has numerous applied examples ranging from economic data to clinical applications and each chapter contains exercises for further enforcing the material presented. If used in a classroom setting, the exercises would be a good supplement to additional problem sets that further emphasized theoretical results and data analysis strategies.

Chapters 1 and 2 provide a general background of time-to-event data, censoring, and nonparametric estimation of the survival function. In these chapters, a general comparison of discrete versus continuous-time data is presented to motivate the topic. A good deal of space is rightly devoted to life table estimates of the survival function as this sets the stage for discrete hazard estimation. If the text were a first exposure to censored time-to-event data, I would like to see further discussion of censoring types, mechanisms, and informative censoring. For a far more in-depth treatment of the topic, I would refer a reader to Klein and Moeschberger (2003).

Chapter 3 introduces the primary regression modeling framework used throughout the text, including the proportional odds model and the discrete time proportional hazards model. There is limited development of theoretical results for estimators in the context of these models, but this is acceptable given