DOI: 10.1111/biom.13227

Multistate models for the analysis of life history data

Richard J. Cook | Jerald F. Lawless

Boca Raton, FL : Chapman and Hall/CRC, 2020. Hard cover. pp. 441. CDN$ 121.00

Survival analysis (or lifetime data analysis) has been an important area in biostatistics, and relevant ideas and techniques are widely used in many research fields including cancer research, clinical trials, epidemiological studies, and actuarial science. Multistate processes, one of important topics and branches in lifetime data analysis or clinical trials, also attract researchers' attentions in recent years. Different from conventional lifetime data analysis with only one failure time, difficulties and challenges of multistate data come from transition of states/diseases. Although many research papers contribute estimation methods to analyze multistate models, there is no reference book that systematically discusses estimation methods for multistate models and relevant data analysis. Fortunately, the authors Cook and Lawless, who both are experts of lifetime data analysis, published a book entitled *Multistate Models for the Analysis of Life History Data* in 2018. Personally, this book has several advantages that attract my attention. For example, this book provides detailed descriptions of data structures and comprehensive discussions of different estimation methods on multistate models and data analysis. In addition to rigorously mathematical formulations, this book also contains many detailed explanations so that readers can easily understand interpretations, meanings of model formulations, and analytical procedures. Furthermore, this book demonstrates numerical studies including simulations and analysis of several real datasets, which makes readers clearly understand methodologies and their applications in real datasets. Many figures and tables are available and well summarized in this book so that readers can easily understand relevant materials with visualization. Many "exercise" problems are available in the end of each chapter, which stimulate readers to review key materials in each chapter and explore some new research ideas. Most importantly, the authors kindly share and summarize software packages, simulation processes, programming code, and datasets in appendices so that the readers can follow the authors' ideas to develop new methods and study data analysis.

For the contents in detail, in Chapter 1, the authors provide general introductions of multistate models, including life history analysis and methods of multistate processes. Several real datasets and relevant software packages are also outlined in this chapter.

Chapter 2 starts by introducing event history processes and multistate models. To characterize event history and multistate processes, the authors first introduce intensity functions and counting processes that essentially govern the stochastic movement between states. To deal with estimation, the authors consider product integration and then derive the likelihood function. Their methods are also extended to deal with time-dependent covariates and informative censoring. After that, the authors introduce some important types of multistate models, including *modulated Markov models* and *modulated semi-Markov models*. In the end of this chapter, the authors discuss how to characterize and obtain process features such as transition probabilities and sojourn time distributions.

Chapter 3 discusses multistate models and aims to analyze data from individuals who are observed continuously over a period of time. Markov models and semi-Markov models are primarily interested in this chapter. The authors first consider parametric models and then apply the likelihood methods proposed in Chapter 2 to obtain the estimators. Moreover, the asymptotic distribution of the estimators is also derived. After that, the authors extend parametric models and consider the nonparametric estimation of cumulative transition intensities. The nonparametric Nelson-Aalen estimator is proposed to estimate the cumulative intensities. In the presence of covariates, the authors then consider semiparametric regression models, specifically focusing on modulated Markov models, additive Markov and semi-Markov models. Finally, to see the adequacy of specifications for the transition intensity functions, the authors discuss model assessment that aims to check parametric and semiparametric model assumptions, predictive performance of models, consequences of model misspecification, and robustness.

In Chapter 4, the authors aim to adapt key ideas in Chapters 2 and 3 to several specific applications. The first application is competing risk analysis. The authors clearly introduce features of competing risk and related intensity-based analysis, and then summarize several estimation methods such as cumulative incidence functions and direct binomial regressions. Another interesting application in this chapter is the estimation of state occupancy probabilities. In addition to binomial estimating functions and pseudo-values methods, the authors also suggest to estimate occupancy probabilities via distributions for entry and exit times for each state. The last application in this chapter is analysis of state sojourn time distributions.

In Chapter 5, the authors focus on a setting in which subjects' states are known only at intermittent observation times. This type of data is common in cohort studies. A main challenge of this setting is that the exact transition times and the number of transitions between successive observation times are unknown. In this chapter, the authors suggest to use conditionally independent visit process (CIVP) to characterize multistate response and time-dependent covariates in this setting. Under the CIVP and Markov assumptions, the authors derive the partial likelihood function and then estimate associated parameters. Parallel with discussions in Chapter 3, nonparametric estimation and method of model checking are also carefully explored. Furthermore, the authors present several interesting extensions on this setting, such as analysis with non-Markov models (which typically extend model formulation in previous section) and mixed observation schemes, which is the case that certain transition times are unobservable but the exact times of others may be observed, subject only to right censoring. Finally, several models, such as illness-death models, general models, and progression-free survival in cancer trials, are carefully discussed.

Chapter 6 discusses heterogeneity and dependence in multistate processes. Heterogeneity typically comes from variation between processes that cannot be explained by available covariates, while dependence represents correlated multistate processes. The authors present heterogeneity in the beginning of this chapter. To characterize heterogeneity, the authors first employ frailty models and introduce latent variables in survival analysis, and then adopt similar ideas to analyze multistate models with latent variables treated as random effected variables. For discussions of dependence in multistate processes, the authors provide several approaches, including formulation using shared or correlated random effects, intensity-based models for local dependence, and the use of copula models for constructing flexible multivariate random effect distributions. In the last two sections of this chapter, the authors introduce finite mixture models (where the random effect variables are discrete) and hidden Markov models to further analyze heterogeneity and dependence in multistate processes.

Sampling of specific individuals whose observed process history satisfies particular conditions is a crucial issue in clinical trials. To provide valid sampling methods in multistate processes, the authors discuss process-dependent sampling schemes in Chapter 7. Two sections are included in this chapter: *history- and state-dependent selection* and *outcome-dependent subsampling and two-phase studies*. In the first section, the authors introduce several types of selection schemes and likelihoods. In addition, some sampling schemes are also discussed, including prevalent cohort sampling, design based on probabilistic state-dependent sampling, and selection and initial conditions with heterogeneous processes. In usual sampling, it is difficult or costly to measure certain variables for everyone in a population. To improve shortcoming of usual sampling methods in multistate processes, the authors consider the two-phase studies in the second section. In this section, the authors first provide general introductions on two-phase studies. After that, the authors focus on two-phase studies with multistate processes and discuss inference for models with semiparametric multiplicative intensities. In general, this chapter not only contains introductions of sampling in multistate processes, but also demonstrates many applications of real data analysis.

Finally, the authors provide several additionally interesting and important topics related to multistate models in Chapter 8. The first topic is analysis of process-related costs and benefits in clinical trials or experiments. Individual-level models and population-level cost analysis are comprehensively discussed. Another crucial topic is prediction. That is, based on multistate models and a given individual's observed history of states, one may be interested in predicting an individual's future life history. As suggested by the authors, the Brier score and the logarithmic score are useful tools to measure the performance of prediction. The authors further take one dataset as an example to demonstrate an application of prediction. Moreover, other complex settings, such as joint modeling of dynamic covariates and multistate processes and causal inference with life history processes, are also clearly presented in the last two sections of this chapter.

In general, I really enjoy reading this book. Many important materials are clearly presented, so that readers easily understand key ideas of multistate models and analysis. Besides, I think this book is a useful reference and I highly recommend this book to researchers who are interested in multistate models and related data analysis.

Li-Pang Chen

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada
Email: L358CHEN@uwaterloo.ca

ORCID

Li-Pang Chen https://orcid.org/0000-0001-5440-5036

DOI: 10.1111/biom.13224

Deep learning with R

François Chollet | Joseph J. Allaire

Shelter Island, NY: Manning

The steep evolution of deep learning in the early 2010s has been driven by the enormous amount of data and advances in computing power. Deep learning is an engineering science, built up on empirical findings based on multiple layers of artificial neurons rather than by theory and assumptions. Deep learning has led to major breakthroughs, which could not be achieved through previous approaches in machine learning, including natural language processing, image analysis, image/text generation, AI in games, and autonomous driving. Consequently, there is substantial interest in deep learning in the statistics community, and this is a very useful introductory book.

This book mainly introduces Keras (a Python library developed by the author of this book, François Chollet) and how to use Keras for various deep learning models through an R interface. Keras is known to be easy to use and user friendly. The R-version of Keras will be especially useful in the statistics community.

In the introductory chapter, the authors provide a broad overview of deep learning and its relationship to machine learning and artificial intelligence. Chapter 2 explains basic mathematical concepts such as tensors and backpropagation. Basic data structure is expressed through tensors. For example, time series data are encoded as 3-dimensional tensors (*samples*, *timesteps*, and *features*), image data are encoded as 4-dimensional tensors (*samples*, *height*, *width*, and *color depth*), and video data are encoded as 5-dimensional tensors (*samples*, *frames*, *height*, *width*, and *color depth*). Backpropagation is a gradient-based optimization algorithm that computes the gradient of the loss function backward from the last layer to the first layer applying the chain rule. These concepts are essential to understand how neural networks work for the practical examples presented in the following chapters.

Chapter 3 offers some hands-on experience that helps understand various types of network architectures, the right learning configuration, how to train models, and how to know which model gives you the right result. This was the chapter that made me open my computer. After a few hassles to install the most updated "keras" package in R (which was not described in the book), I started running the keras and sample code provided in the book. I then enjoyed reading the book even more. Chapter 4 provides a conceptual framework for general machine-learning problem solving (not limited to deep learning) focusing on generalizing the algorithm to new data. A blueprint described in Chapter 4.5 is an excellent high-level summary, covering all steps to follow, with helpful tips and key choices needed to be made at each step. A common way of handling missing values is also described in this chapter.

Chapters 5 and 6 introduce convolutional neural networks (*ConNets*) and recurrent neural networks (*RNNs*) as deep learning models for computer vision/image and general sequence data, respectively. The sequence data include text data (viewed as sequences of words and characters) and time series data (eg, weather and stock). The first half of the chapters well describes the entire process of using the *ConvNets* and *RNNs* in the keras package, including unique preprocess of data (eg, convolution of image data and tokenization of text data) to validating and visualizing the trained representations. However, as a novice to deep learning, I wish these chapters had begun with the architectural foundations of the *ConvNets* and *RNNs*.

I personally was fascinated by two aspects of deep learning: pretrained networks and use of *ConvNets* for sequencing data. The prelearned features on a large dataset can be used across different problems, which makes deep learning very effective with even small datasets. The second half of Chapters 5 and 6 focuses on how to use a pretrained network. As a faster alternative to *RNNs* for the sequencing data, 1-dimensional *ConvNets* are presented in Chapter 6.4, given that the natural language process can be viewed as pattern recognition similar to the pattern recognition in pixels.

Combining different types of neural networks is introduced in Chapter 7 as an advanced technique to build models with complex structures such as multi-input models, multi-output models, graphical models, layer sharing, and model sharing. This chapter also adds visualization tools to monitor models during training and to make appropriate adjustments. Current and future directions for deep learning, including generative models that create new work, are discussed in Chapters 8 and 9.

To sum up, this is an excellent introductory textbook for statisticians, data scientists, and graduate students. The book covers most fundamental concepts of deep learning, while focusing on their implementation. As the chapters are logically connected, the book flows easily. Lots of diagrams, pictures, and annotations embedded right next to sample code improve the clarity and intuition, which makes it straightforward to use. If you need one book to get you started with deep learning, and you would like to take advantage of easy accessible R interface, you will enjoy this book.

Sehee Kim

Department of Biostatistics, School of Public Health,
University of Michigan, Ann Arbor, Michigan, USA
Email: seheek@umich.edu

ORCID

Sehee Kim https://orcid.org/0000-0002-1815-6957

DOI: 10.1111/biom.13225

Spatial data analysis in ecology and agriculture using R

Richard E. Plant

Boca Raton, FL: CRC Press, 2019.

In the Anthropocene, landscapes are shaped by environmental conditions, land-use intensification, and climate change resulting in a wide range of spatial patterns (eg, gradients and patches) according to the spatial scale(s) at which ecological data are studied. To characterize, quantify, test, and ultimately model these various spatial patterns, there is now a comprehensive body of spatial analytical tools available to researchers ranging from spatial statistics, spatial regressions to spatial modeling. Hence, nowadays researchers are poised with a wide array of spatial and spatiotemporal methods to assess the spatial dynamic of ecological data. Yet, to determine which spatial methods to use requires understanding the specifics and nuances of all these methods and the appropriate R code packages to analyze ecological data.

The second edition (2019) of the book comes only a few years after the first edition (2012). Yet, it is well known that by the time a book is printed it is already outdated. This is especially the case with books providing R code as the field of spatial analysis is growing at an exponential pace and R code at quantum speed. Hence, I applause Dr. Plante for offering us a second edition of his book with updated material and R code.

The second edition of *Spatial Data Analysis in Ecology and Agriculture Using R* is offering an elegant balance between providing the essentials of the key spatial methods and practical insights about the current R code available needed to perform these analyses. Additionally, each chapter ends with a series of exercises and suggestions for further readings. The chapters cover the same material in both editions with, of course, some improvements and additions in the second edition. Yet, the former chapter entitled *Spatial Data Exploration via Multiple Regression* was reorganized into two new chapters aiming to better present how first to model spatial data using parametric nonspatial models (*Data Exploration Using Non-Spatial Methods: The Linear Model*) and then using nonparametric nonspatial models (*Data Exploration Using Non-Spatial Methods: Nonparametric Methods*). These new chapters are making a comprehensive presentation of the state-of-the-art in the ways to model spatial data. Then, Dr. Plante had to make some strategic decisions and as such he added a very valuable new chapter entitled *Analysis of Spatiotemporal Data*. Although spatiotemporal analysis of ecological data deserves a book by itself as the one by Wikle *et al.* (2019), Dr. Plante is presenting the key spatiotemporal steps and methods to analyze ecological data ranging from spatiotemporal variograms, spatiotemporal interpolation, spatiotemporal clustering, spatiotemporal process models, and Bayesian spatiotemporal analysis. Unfortunately, to fit within the constraints of pages limit, he had to remove a chapter and he decided that it would be the one entitled *Multivariate Methods for Spatial Data Exploration*, dealing principal components analysis and recursive partitioning methods (eg, classification trees, regression trees, and random forest). To render the removed material, as well as new material on other methods, Dr. Plante makes them available on a website.

In short, the second edition of Dr. Plante's book is well written, informative, and useful. The textbook allows the researchers to learn spatial analysis through a series of examples and exercises. Also, I strongly recommend it as a textbook for spatial statistical courses.

Marie-Josée Fortin

Department of Ecology and Evolutionary Biology,
University of Toronto, Toronto, Ontario, Canada
Email: mariejosee.fortin@utoronto.ca

REFERENCE

Wikle, C.K., Zammit-Mangion, A. and Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Boca Raton, FL: CRC Press.

DOI: 10.1111/biom.13223

Nonparametric models for longitudinal data: With implementation in R

Colin O. Wu | Xin Tian

Boca Raton, FL: CRC Press

The book *Nonparametric Models for Longitudinal Data with Implementation in R* written by Dr. Colin Wu and Dr. Xin Tian contains 15 Chapters categorized in five Parts. The book cites 164 research works in its bibliography. It contains 38 figures and 7 tables, and used four famous longitudinal data sets for practical implementation of various methods discussed throughout the book. Among these data sets, two are epidemiological studies, and remaining are longitudinal clinical trials. At the end of each chapter, unsolved research questions and directions for new research are mentioned for graduate students and researchers.

Part I has two chapters (1 and 2), which provides the synopsis of the remaining chapters and sections of the book. It starts discussion about the data sets, mathematical notations, and structure of the longitudinal data and variables. It also provides basic notion and mathematical formulations of the parametric, semi-parametric, unstructured nonparametric, structured nonparametric, and mixed effect models for longitudinal data. Each model specified in Part I is derived theoretically together with asymptotic results and applications in the subsequent chapters and sections in details. Practical implications for each model are demonstrated by statistical software R. All R codes are provided in GITHUB.

Part 2 consists of three chapters (3-5) and introduces smoothing methods such as kernel, local polynomial, basis approximation, and spline. “Leave-one-subject-out” cross-validation method has been introduced in this section to choose smoothing parameter, known as bandwidth. Bootstrap confidence interval, simultaneous confidence bands, rate of convergence, hypothesis testing, and asymptotic properties of the smoothing estimators are also derived explicitly in this section.

Part 3 (Chapters 6-9) discusses smoothing methods under time-varying coefficient models. The authors systematically discuss the one-step kernel and local polynomial smoothing methods, and then they introduce two-step smoothing methods and provide the rationales for two-step smoothing methods. In Chapter 9, they discuss global smoothing methods via basis approximation and B-spline. In all cases, they derive the relevant asymptotic results, and show R implementation of these methods. The authors can add “one-step kernel log likelihood method” for smoothing estimation of time-variant parametric models between Chapters 7 and 8.

Part 4 discusses the shared-parameter and nonparametric mixed effect models in Chapters 10 and 11. With the inclusion of the concomitant intervention, a substantial extension of the structured modeling for the effects of outcome-adaptive covariates is addressed in Chapter 10. The readers have been familiarized with a class of nonparametric mixed-effects models in Chapter 11 to overcome the modeling assumptions and limitations of the parametric mixed effect models.

Finally, in Part 5, which includes Chapters 12 through 15, cover methods for estimation of conditional distribution function, conditional quantiles, rank-tracking probabilities, and others. Rank-tracking probability of the longitudinal variables helps to track individuals with high risk of developing some adverse health conditions such as high blood pressure, high cholesterol, and high Triglyceride, to name a few.

In conclusion, each chapter of the book has been formatted by introducing some general discussion and need for mathematical models, mathematical definition and elaboration of different types of models, R implementation of different methods on real-life data, and finally mathematical derivation of asymptotic results. Chapters of the book are arranged in such a way that each topic is linked with topics from the previous chapters and also they are independently readable. The examples provided in the chapters also perfectly align with the theory discussed in them. For this reason, the concepts of the book prove to be comprehensive for its target demography of readers. As the contents of the books are from very high level research publications, it could be hard for applied researchers to entirely follow the contents of the book without some assistance from software packages. To relieve the workloads of the applied researchers, the authors graciously provided an R package for easier implementation of the discussed methods in the book, which makes it a good resource for applied research. The book can be an excellent reference guide for students who look to carry out research in subjects such as epidemiology, biostatistics, and for applied practitioners but might be a hard read for general readers. So, the authors could take another initiative for writing a simpler and applied version of the book where all mathematical derivation could be ignored, and more applications on real-life data can be provided.

Mohammed Chowdhury

Department of Statistics and Analytical Sciences, Kennesaw State University, Kennesaw, Georgia, USA
Email: mchowd10@kennesaw.edu

ORCID

Mohammed Chowdhury
https://orcid.org/0000-0001-8330-8438