

Book Reviews

Editor: Ananda Sen

Estimation and Testing under Sparsity

Sara van de Geer

Springer, 2016, xiii+274 pages, \$59.99, softback

ISBN: 978-3-319-32773-0

Sparsity in the title refers to statistical models with a large number of parameters most of which are essentially zero. This monograph is based on lectures to a mathematical audience. The author is explicit that no computational aspects are discussed nor are specific applications given. The issue under discussion is how to exploit sparsity in the analysis of regression-type models in which the number of explanatory variables is large relative to the number of observations. More specifically, the objective is in the light of a sparsity assumption to choose a specific model with a small number of parameters so as to produce satisfactory prediction.

The book explores these issues with elegance and remarkable lucidity given the demanding mathematical level. It will surely generate further interest in what is a very topical theme and be a major source for new work in mathematical statistics.

To my regret, I am ill-equipped to comment in depth on the mathematical aspects of the book. I do, however, have a concern over formulation at least in some contexts. If the objective is interpretation rather than prediction, the prime issue is: have the right variables been chosen?

D.R. Cox: david.cox@nuffield.ox.ac.uk

Nuffield College

Oxford OX1 1NF, UK

Ordered Regression Models: Parallel, Partial, and Non-Parallel Alternatives

Andrew S. Fullerton, Jun Xu

CRC Press, 2016, xvi + 172 pages, £49.99/\$79.95, hardcover

ISBN: 978-1-4665-6973-7

Readership: Senior and graduate students, researchers and practitioners, working with ordinal data.

Ordinal regression may be appropriate when observations are outcomes that are ranked on an ordered scale. Five-point Likert scales, with responses that range from ‘Strongly dislike’ (or disagree) to ‘Strongly like’ (or agree) are common in survey questions. Ordinal regression models aim to respect the ordering of categories, while not assuming, as when the ordinal values are used as values of the outcome variable in a linear regression that cut-points are equally spaced. Ordinal regression models combine in one model the binary regression models for the individual cut-points.

The book contains five chapters. Chapter 1 gives a brief history, describes the different types of model, and describes the data sets used in later chapters. Cumulative models compare, for each cut-point, the total count below with the total above. Self-rated health data

is used in Chapters 2–4 to illustrate models of this type. Continuation ratio models, also called ‘stage’ models, compare the count for each category with the count for that and all higher categories. Educational attainment data, where later categories can be reached only after proceeding through earlier categories, are used for illustration. Adjacent category models direct attention to the relative numbers in adjacent categories. Data on attitudes to welfare spending are used for illustration.

In the fully parallel models that are the subject of Chapter 2, covariate effects move all cut-points, on the scale of the linear predictor, by the same amount. Chapter 3 relaxes the parallel assumption for a subset of the covariates, while Chapter 4 treats non-parallel models. Chapter 5 discusses formal methods for testing assumptions of parallelism. Chapter 6 discusses extensions of the models in earlier chapters that can account for more complex patterns and data structures. These are as follows: heterogeneous choice models, multilevel ordered response regression, and a Bayesian approach to ordered response regression. There are extensive comments on reasons for choosing one model rather than another, on interpretation, and on specific issues and difficulties for some types of model.

The analyses in Chapter 5 place much reliance on multiple tests, with a combination of composite tests and tests for individual variables. Subsection 5.4.5 discusses the use of p -value based sequential testing to choose the ‘best’ model. Even for the authors’ stated purpose of identifying variables that should be the focus of subsequent tests based on predicted probabilities, this has serious implications. Biases, not the expressed concern about overfitting, are the major concern. In particular, subsequent p -values are too small. More to the point is the advice in Section 5.9 that ‘In general, we recommend using the model with the fewest parameters that yields approximately the same predicted values as more complex models’.

Here, note the warning in Harrell (2015, Section 4.3), speaking of p -value based variable selection more generally, that this procedure ‘violates every principle of statistical estimation and hypothesis testing’. Harrell gives a summary of the problems, commenting that a reasonable alpha that allows for deletion of some variables may be $\alpha = 0.5$.

Appendices to Chapters 2, 3 and 4 give STATA and R codes for the main analyses. More detailed code and examples, for STATA and R, are available (or for R will be available) from the website <http://sociology.okstate.edu/people/directories/faculty-directory/dr-andrew-fullerton>. The R code uses functions from Thomas Yee’s vector generalized linear and additive models package.

There is much useful general advice on the rationale for choosing one model or a class of models over another. While the mathematics of ordinal regression models as presented here is relatively straightforward, the conceptual demands and the practicalities of their use are not. It would have been helpful to start the discussion with simple examples with just one factor or covariate. Chapters 2–5 would benefit, and be simplified, from a rewriting that uses as the starting point the most complex model that makes good sense, with simplifications limited to those that do not change predicted values much.

John H. Maindonald: john.maindonald@anu.edu.au
40 Futuna Close, Wellington 6012
New Zealand

References

Harrell, F. E. (2015). *Regression Modelling Strategies, with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed. New York: Springer.

Handbook of Health Survey Methods

Timothy P. Johnson, Editor

John Wiley and Sons, Inc., 2015, xxii + 816 pages, £100.00/€125.00, hardcover

ISBN: 978-1-118-00232-2

Readership: Statisticians, survey methodologists and health researchers.

Major scientific programmes of health survey data collection date back to the 1960s and 1970s, but the past 25 years have clearly produced the most important advances in research and publication on the important methodological issues faced in health survey design, measurement, data collection methods and analysis. Advances towards a comprehensive body of literature on the methodology of health surveys have faced several impediments. The first is the heterogeneity of health survey programmes and their scientific objectives. Health survey programmes have many aims—describing population health status; investigating disease epidemiology, health disparities and vulnerabilities; assessing treatment outcomes; and evaluating the efficacy of health systems and healthcare costs. A second barrier has been the disciplinary differences in where and how survey methodologists, statisticians and health scientists published or otherwise communicated knowledge, ideas and experiences related to health studies.

Handbook of Health Survey Methods is a massive volume that brings together most of what is known from health survey practice. Comprehensive in its breadth of coverage and detailed in its treatment of each topic, the volume is a compendium of 29 monograph chapters prepared by 54 authors with wide-ranging expertise and practical experience in the world of health survey research. The many chapters are effectively organised as five sections or parts that span major topical areas including Part I: Design and Sampling Issues; Part II: Design and Measurement; Part III: Field Issues; Part IV: Health Surveys of Special Populations; and Part V: Data Management and Analysis. The choice of chapter topics spans the important methodological domains from design and planning through analysis of the survey data. In addition to the expected chapters on design, measurement, data collection, weighting and analysis, the volume includes important guidance on ethical issues, administrative and spatial data linkage, mode effects and the use of new data collection technologies.

The volume is carefully edited and formatted. Each topical chapter provides the reader with a review of the literature and appropriate theoretical background. Most chapters include relevant applications or empirical results from actual health survey programmes. A strength of the *Handbook of Health Survey Methods* is that the editor and authors explicitly recognise the globalisation of health survey research programmes and the cross-national sharing of designs and methodology. Worldwide, many health survey programmes that study household populations, patients under care, physicians and providers and healthcare networks share common designs and methods. While country-specific survey conditions and objectives certainly still mandate local adaptations in these programmes of health survey research, there are many shared features such as sample designs and eligibility rules, measurement constructs, questionnaire content, data collection management, data processing, survey weighting and analytic techniques.

Purchasing an edited volume can be analogous to sitting down with friends to enjoy a meal of dim sum—there will be individual differences in preferences for the various dishes. Some dishes will appeal to everyone present while others may have few ‘takers’. Following this analogy, the *Handbook of Health Survey Methods* holds something for everyone. Each reader is likely to find their own set of favourite selected chapters in this edited volume. As a survey statistician and practitioner who has struggled with the design and methodology of health surveys for almost 40 years, I particularly appreciated the Part I material on sampling and design

and in particular, the coverage in Chapter 3 of case-control designs for health surveys. As a reader less familiar with the complexities of developing or selecting physical and health measurement scales including indices and scales, the eight chapters included in Part II provided me a comprehensive, up-to-date review of the literature and best practices for item pretesting and adapting measures to new data collection technologies. Part III coverage of Field Methods goes into depth on proxy reporting, measurement of sensitive behaviours and contextual and ethical considerations in health survey measurement. Chapter 15 in this section addresses an important, emerging area of health survey methodology—the collection of biospecimens from survey participants. To meet policy-related or humanitarian goals, health survey programmes must study many different populations—medical care providers, clinics and hospital systems, patients and the general public. Part IV of the handbook includes chapters that focus on design and methodology for surveys of these varied populations as well as specialised approaches that are needed in the study of vulnerable or at risk populations. As web surveys are becoming more commonplace in health surveys, Chapter 24 addresses the challenge of ensuring that persons with disabilities are able to access and complete self-administered, online interviews.

The final chapters (Part V) of the volume cover statistical treatments for health survey data including such important constructs as validity and dimensionality, design-based weighting for inference to survey populations, survey data linkage to ancillary data and general methods of analysis of data from complex samples. Linkage of survey data to administrative data systems and techniques for merging the data with spatial and contextual information are increasingly common and important in the analysis of health survey data. Chapters 27 and 28 of the handbook discuss both the opportunities and common pitfalls in these forms of augmented analysis of the basic health survey data. Readers seeking in-depth coverage of analytic methods for health survey data should supplement the material in this volume with one of several texts that deal specifically with health survey data analysis (Korn and Graubard, 1999; Heeringa, West and Berglund, 2010).

In summary, health researchers, survey statisticians and survey methodologists will find that the *Handbook of Health Survey Methods* is a very accessible and comprehensive edited volume that will enable them to learn the basics or to update their knowledge of theory and current best practices for health survey research. The volume is not specifically designed as a teaching text; however, with careful selection of chapter readings and modest augmentation from external publications, it could well serve as the primary text for a one or two-semester course in health survey methods. This text is a ‘must’ on the bookshelf of anyone who is engaged on a daily basis in designing, conducting or analysing surveys of health topics.

Steven G. Heeringa: sheering@umich.edu
Institute for Social Research
University of Michigan
Ann Arbor, MI 48109, USA

References

- Korn, E.L. & Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley and Sons.
Heeringa, S.G., West, B.T. & Berglund, P.A. (2010). *Applied Survey Data Analysis*. Boca Raton, FL: Chapman and Hall.

Generalised Principal Component Analysis

Rene Vidal, Yi Ma, S. Shankar Sastry
Springer, 2016, 553 pages, €72.79, hardcover

Readership: Statistics graduate students, researchers and computer scientists.

This is an attractive and useful book, which provides theory and methods (with computational details) for modelling high dimensional data with one or more low-dimensional subspaces and manifolds. This book generalises the classical principal components analysis (PCA) to problems with missing data, outliers and non-linearity.

The book is very well structured and starts with a glossary that explains the mathematical symbols used throughout the book. The book has three parts. Part I focuses on data in a single space. It starts with a review of the classical PCA and then extends it to robust and non-linear PCA. In Part II, it addresses more complex problems that are modelled in multiple subspaces. In both parts, the dimension reduction problem is formulated both from a geometric and a statistical point of view, which provides a comprehensive understanding of the PCA.

The last part (Part III) talks about various applications related to pattern recognition and image analysis. However, the material is challenging for readers who do not have a working knowledge of the field. The long Appendix (about 70 pages) covers various background concepts used in this book, including optimisation methods, mathematical statistics and algebraic geometry. The Appendix is very useful for readers who are not so familiar with certain mathematical facts used in the book, especially in the early chapters. Additionally, the book provides algorithms for implementing the methods, where each algorithm contains a step-by-step set of instructions.

Principal components analysis is a very broad, active research area. The technique is becoming increasingly important, and a first line of exploration due to the advent of large-scale problems. This book is a welcome addition to the literature, complementing the existing texts in this topic by providing a comprehensive generalisation.

Lili Zhao: zhaolili@umich.edu

Department of Biostatistics, University of Michigan
1415 Washington Heights, Ann Arbor, MI 48109, USA

Branching Process Models of Cancer

Richard Durrett

Springer International Publishing, 2015, vii + 63 pages, €31.79, softcover
ISBN: 978-3-319-16064-1

Readership: Students and researchers in a broad range of sciences at the intersection of applied probability and biological sciences.

This book stems from a lecture series organised by the Mathematical Biosciences Institute. It makes an excellent introduction to the branching processes in general and their applications in cancer in particular. The style of the book, while mathematically rigorous, is accessible to students and researchers in a broad range of sciences at the intersection of applied probability and

biological sciences. Knowledge of basic probability and Poisson processes should be a sufficient prerequisite. More complicated foundations of the theory such as martingales are mentioned, so an interested reader could follow with a more comprehensive mathematical reading. The book is sufficiently short and concise, so it will actually be read in full with high likelihood.

The book also represents an excellent module for a course on mathematical or statistical methods in cancer research. These are graduate-level courses that introduce students to methods that, while important, are typically missed by traditional mainstream courses in applied stochastic processes and probability. The book provides sufficient exposition of basic cancer biology to serve as an introduction to cancer for a quantitatively oriented student or a researcher who has never applied mathematics in this field.

The theory of multi-stage models of carcinogenesis that occupies a central place in the book has so far been represented mainly in scientific papers and a handful of review articles in professional mathematical biosciences journals. Chapters 1–3 of the book introduce cancer as a multi-stage process going back to the original Armitage and Doll model and providing some basic facts from branching process theory. Chapters 3–6 provide a link to the cancer development and progression events as the time/size of the tumour reaches a certain threshold.

Mapping biological processes to abstract unobserved mathematical objects is an inherently fuzzy business. However, mathematical models that establish such links are useful as they capture the salient feature of cancer development and progression phenomena, and serve as the foundation for statistical models that connect the mathematics to real data and provide an important feedback through model fitting and calibration. One example of such linking is a definition of stages of carcinogenesis through different types of mutations and linking them to the observable processes such as cancer detection. Mathematical theory and the biological model co-evolve in the book in Chapters 3–15 where mutations, numbers of cells of certain type and the detection process are carefully characterised while the necessary mathematics of branching processes is gradually introduced in a to-the-point fashion supporting the model development and supplying the necessary motivation for mathematical development. This style of exposition maintains the reader's attention and interest and avoids leaving a non-specialist reader wondering about the need of mathematical abstraction.

In Chapters 16–18, the theory is brought to bear on real-life applications. These include a study of tumour metastases (Chapter 16), ovarian cancer, the problem of cancer screening and size-metastases relationship (Chapter 17) and within-tumour heterogeneity (Chapter 18).

In summary, this is an interesting and entertaining reading that I would put first if I were a quantitatively oriented scientist or student interested in getting into applications of probability and stochastic processes to study cancer. The book makes a captivating introduction to the field and fosters enthusiasm to study the field further by following this text with more comprehensive branching processes books by Harris (1963) and Jagers (1975), and professional scientific journal literature such as *Mathematical Biosciences* (Elsevier).

Alexander Tsodikov: tsodikov@umich.edu

Department of Biostatistics, School of Public Health
University of Michigan, Ann Arbor, MI 48109, USA

References

- Harris, T.E. (1963). *The Theory of Branching Processes*. Berlin: Springer.
Jagers, P. (1975). *Branching Processes with Biological Applications*. London: Wiley.

Exposure-Response Modelling: Methods and Practical Implementation

Jixian Wang

Chapman & Hall/CRC, 2015, xv + 331 pages, £61.99, hardback

ISBN: 978-14665-7320-8

Readership: Biostatisticians and applied statisticians interested in exposure-response modelling in a pharmaceutical context.

In the preface of the book, the author states: ‘*This book covers a wide range of topics and new developments starting from traditional pharmacokinetic-pharmacodynamic (PKPD) modelling, and progressing to using measurement error models to treat sequential modelling, fitting models with exposure and response driven complex dynamics, survival analysis with dynamic exposure history, Bayesian analysis and model-based Bayesian decision analytics, causal inference to eliminate confounding biases, and exposure-response modelling with response-dependent dose/treatment adjustments, known as dynamic treatment regimes, for personalised medicine and treatment adaptation*’. I think that such a description gives a good overview of the contents but to accomplish such goals in a single book is a challenge. Such an objective necessitates the reader to have a strong background knowledge in statistical concepts and applied statistical modelling. So I would suggest that a potential reader take a look first at Appendix A that summarises some essential statistical theory. Chapter 1 is an introduction to the exposure-response (ER) relationship with description of some practical scenarios in ER modelling. I like how the applicability of the concept is linked to many different areas even though the four concrete examples are all from pharmaceuticals.

Chapters 2–9 deal with models of ER. The basic models are described in Chapter 2, whereas models for longitudinal data are discussed in Chapter 3, followed by sequential and simultaneous modelling in Chapter 4. Survival models for time-to-event data are dealt with in Chapter 5. Dynamic relationships are described in Chapter 6 that make segue to Bayesian modelling and decision analysis in Chapter 7. Causal modelling are described in Chapter 8 with Chapter 9 devoted to advanced dose-response modelling. Although some of the topics may sound specific to ER modelling, in practice those turn out to be just applications of standard statistical models in the ER context.

The extent of the different models and related statistical issues presented is impressive. It feels like you come to a table that is adorned with a wide variety of dishes — some everyday ones, some a bit more exotic, all spiced using the ER flavour. Everything is elegantly written with concise mathematical notation giving a good overview and reminder of essentials pertinent to those topics. I could easily give this book to someone who wanted to become an applied statistician and say that here you can see a good deal of things you should be familiar with. However, in my opinion, this book alone is not enough to adopt those ideas from scratch. There are separate more detailed books for all the topics. Indeed, the book provides bibliographic notes and suggestions for further reading at the end of each chapter.

Chapter 10 discusses implementation of the models using statistical software. Three types of software are used in the examples of the book: SAS, R and NONMEM. Although the details of using different software packages are different, it is argued in the book that there are two key elements needed for modelling in the software, namely the model and the data. This chapter focuses on the specification and fitting of linear and non-linear mixed models in the different software platforms. This chapter stands out from the rest and could well serve as an Appendix that should be read, if necessary, before Chapters 2–9.

Chapters 2–9 also contain examples and code scripts for selected models or tasks. This certainly is a step towards an easy application of the models as it allows for concrete testing of the approaches. However, no supplementary material containing the scripts or data were available online. This is an unfortunate hindrance to a seamless implementation of the methodology.

In principle, I liked the variety of material included in the book. However, this also made the overall presentation perhaps a bit scattered. There are also many examples including novel applications of the methods in the ER context, but for my personal taste, applications and examples from areas other than traditional pharmaceuticals would have been nice. In any case, this is probably the most comprehensive book on the topic including everything you need as long as your breadth of knowledge in applied statistics is substantive.

Reijo Sund: reijo.sund@helsinki.fi

Centre for Research Methods, Department of Social Research
University of Helsinki, P.O. Box 18, FI-00014, Finland &
Kuopio Musculoskeletal Research Unit (KMRU), Institute of Clinical Medicine
University of Eastern Finland (UEF), Kuopio, Finland