Biometrics 74, 769–778
June 2018

DOI: 10.1002/biom.12895

BOOK REVIEWS

EDITOR:
DONNA PAULER ANKERST

TREVOR HASTIE, ROBERT TIBSHIRANI, AND MARTIN WAINWRIGHT. **Statistical Learning with Sparsity: The Lasso and Generalizations**. Boca Raton: CRC Press.

Regularization has played a major role in statistics ever since the new era of data generation has reached its foothold. During this period analyzed data sets started having many more features than observations, which made standard statistical methods inconvenient. Perhaps the most widely used and well-known regularization method today is the "Least Absolute Selection and Shrinkage Operator," better known simply as the Lasso. This method was originally introduced by Robert Tibshirani in 1996 and further popularized as part of the highly accessed book "The Elements of Statistical Learning" in 2001, written by some of the same authors as the book described here. It is therefore needless to say that the authors of "Statistical Learning with Sparsity" comprise valued experts on this topic. With their most recent book, the authors aim to provide a broad overview on classical concepts of regularization in statistics as well as the newest trends in this direction. In our opinion, the book succeeds in providing a comprehensive guideline on this topic that is well-suited for teachers as well as students, more precise reasons will be given at the end of the review.

The book starts off with a general introduction to the field that describes the idea behind sparsity in modeling. The concepts are demonstrated nicely with two real-world examples that illustrate the results from applying a regularized model. The first chapter of a book is important as it often decides whether one continues interest in the book, or drops it and shops for something else. In this case, the attention of the reader is immediately captured, and even enhanced, by the well-chosen examples. The book then goes on to introduce regularization in generalized linear models in the following two chapters, where standard classical works are reviewed. What immediately jumps to attention is the completeness of discussion of single topics. Equations are not only provided and discussed, they are complemented by examples that

are relatable to many researchers. Additionally, downsides of methods in the sense of computational burden or accuracy are presented. Furthermore, the authors put effort into offering software recommendations for the methods, which is highly useful for applied scientists.

In Chapter 4, extensions of the Lasso model are discussed. Some of them, such as the Elastic Net or Grouped Lasso again appear to be well-known. But additionally, some lesser known alternatives, such as the Fused Lasso or Overlap Group Lasso, are listed. This overview of regularization-based methods is then followed by a rather technical chapter discussing different optimization strategies for the models. Subsequently, in Chapter 6 statistical inference techniques are discussed. In our opinion, this chapter presents one of the highlights of the book at least from the perspective of a statistician reader. Here, the newest developments regarding the quantification of Lasso results are discussed and the main focus is put on whether the selected covariates have also a statistical significance. This is shown to be assessed by Bayesian methods, as well as bootstrap or even theoretical derivation of p-values for Lasso coefficients. The next three chapters are mainly focused on sparse versions of popular multivariate methods, such as principal component analysis, linear discriminant analysis, deep learning or Gaussian graphical models. Again, the well-chosen and relevant examples demonstrate the power of sparsity, which these methods profit from, placing them in contrast with other standard methods, some of which produce non-interpretable results. While reading through the different methods, the reader gets the impression that almost every single existing statistical method can be regularized by the same concept. Thus the new methods are easy to follow and understand. The last two chapters of the book are then again of a more technical nature, whereby sparsity in signal approximation as well as theoretical results of the Lasso model are shown.

Overall, the book shows a great consistency throughout the different chapters (excluding Chapter 10 where even the notation is different from the rest of the book). It always first discusses regularized models based on equations, followed by example applications, before ending with a bibliography section detailing the historical development of the given method. Software recommendations (mostly open source R packages) are typically provided either in the main part or bibliography section of each chapter. And each chapter concludes with a set of selected exercises meant to deepen the gained knowledge on the given subject, which of course is of great help for teachers of statistics. For these reasons, we congratulate the authors of "Statistical Learning with Sparsity" and recommend the book to all statistically-inclined readers from intermediate to expert levels. In addition, it is worth pointing out that even for non-statisticians, the book is able to demonstrate, based on numerous real-world examples, the power of regularization.

Ivan Kondofersky and Fabian J. Theis
Institute for Computational Biology
Helmholtz Zentrum Muenchen
Munich, Germany
ivan.kondofersky@helmholtz-muenchen.de
fabian.theis@helmholtz-muenchen.de

NORMAN MATLOFF. **Parallel Computing for Data Science: With Examples in R, C++, and CUDA**. Boca Raton: CRC Press.

A very timely and engaging book, written by an active researcher in the area, Matloff's Parallel Computing for Data Science offers a highly readable and accessible introduction to many different and distinct topics relevant to bringing performance computation to the rapidly changing field of Data Science.

In a mere 300 pages, the author manages to cover thirteen distinct chapters as well as three appendices. A general introduction to parallel computing is followed by a well written and very welcome chapter discussing various obstacles and impediments to faster code. A deeper discussion of parallel loop scheduling follows, as well as three chapters on shared memory computing—a topic the author has particular expertise in, and generally prefers as a parallel computing approach—from both the R and C languages as well as on Graphics Programming Units (GPUs). This is followed by a chapter on the GPU-programming packages Thrust (by NVidia) and Rth (by the author). Next comes a brief chapter on MPI (a staple of scientific parallel programming), and another brief chapter on Hadoop which is (or maybe was?) common in industry data science installations. Next come more foundational and deeper chapters on parallel sorting and merging, prefix scans, and parallel matrix operations. The final chapter covers subset methods with a focus on chunk averaging. It is followed by three brief appendices: an introduction to matrix algebra, an introduction to R and a (very brief) introduction to C for R programmers.

The reader's head may be spinning given the general speed with which topics are introduced, briefly discussed and then mostly cast aside for the next topic, though occasional references thread back and tie chapters together. It goes without saying that basically each of these thirteen chapters has seen book-length coverage in its own right, so one cannot expect truly exhaustive coverage of all the material. The focus is therefore on general introductions, with a sprinkling of new and helpful information in several of the chapters.

Many tips and best practices suggestions are sprinkled throughout the book. The recommendation of "fast enough" is sound. Focusing on the "parallel" package (that comes bundled with R) is good, as is the focus on what it took from the snow package. Reminding the user that "hardware matters" is also important. Performance gains from "chunking" is one common thread through several chapters. For anyone familiar with Matloff's work, the preference of shared memory parallelism over, say, message passing is not surprising. The author also gives some deserved shout outs to his "partools" package.

The writing and overall tone are engaging and fluid, if overly conversational and a little too casual at times make it seems more reminiscent of writing for a blog post than a textbook.

A somewhat sad note is the lack of more thorough editing. A few simple typos persisted. The typesetting is a little surprising at times with presentation formats for actual programming code which differ by programming language. Results (for example from timing benchmarks) are most often presented in very plain tables—but graphs are used

at the very end. Why not sooner? And I was particularly disappointed by the fairly limited coverage of C and C++ programming, which these days is quite accessible from R. The overall coverage is somewhat basic, as well as dated, showing the explicit compilation and linking steps, which nowadays are completely hidden behind helper functions (when e.g., Rcpp is used), the insistence on .C() when .Call() should be used, and more. The final appendix on C for R programmers could (and maybe should) have been omitted as it adds very little.

But these minor complaints should not take away from the fine accomplishment. The book serves as a very decent introduction for those seeking first exposure to one or more of the topics. I learned something new in several of the chapters, and enjoyed reading the book in just a few sessions.

Matloff's Parallel Computing for Data Science: With Examples in R, C++ and CUDA can be recommended to colleagues and students alike, and the author is to be congratulated for taming a difficult and exhaustive body of topics via a very accessible primer.

Dirk Eddelbuettel
Debian and R Projects
Chicago, Illinois
dirk@eddelbuettel.com

JAY HERSON. **Data and Safety Monitoring Committees in Clinical Trials, 2nd edition**. Boca Raton: CRC Press.

I recently became interested in this book after I chaired my first Data Monitoring Committee (DMC) for a cancer screening trial and realized there was a process to the important work of these committees that every member should review. DMCs serve the vital role as independent safeguard for patients participating in a clinical trial, meeting periodically to review safety and efficacy endpoints, and ultimately making recommendations to the sponsor of the trial for closure or adaptation of the trial in the face of excess risk or benefit. Typically, there is only one biostatistician on a DMC so it falls on this member to explain potentially difficult statistical concepts to physician members of the committee, such as interim stopping rules for futility or benefit, as well as statistical significance of serious adverse events (SAEs).

As indicated in the preface of the first edition in 2008, DMC's, which had initially formed primarily to serve federal government funded trials in the 1980's, had reached their adolescence and become a mainstay for industry-funded trials due to a preference by the US Food and Drug Administration (FDA) for granting market approval. Now, a decade after the first edition, the book has been updated to re-visit best practices from the accumulated experience of the author and to accommodate state-of-the-art developments in clinical trial design, including novel precision medicine-, adaptive designed-, seamless Bayesian Phase II/III-, and pragmatic-trials.

The book is useful for anyone involved with a DMC, not just the committee members themselves, but also statisticians in the Data Analysis Center (DAC), as well as the study team and sponsor of the trial. Nicely written and readable cover-to-cover, the author walks through every facet of a DMC, beginning with an overview of their past and current place in drug development, to how they are organized and interfaced with other committees, and on to the specifics of how a typical meeting is split into an open and closed session. From there it moves on to clinical issues, including how SAEs are categorized, statistical methods, including Bayesian and frequentist conditional power calculations, common biases and pitfalls, and guidance for how DMC decisions are made. It concludes with emerging issues due to new clinical trial designs mandated by the FDA to speed up the drug development process.

The book is accessible to physicians, who would also benefit from reading it since the short time and typical conference call format of a DMC meeting prohibits comprehensive explanation of the issues. The chapter on statistical methods is especially nicely presented via a running example of a trial on cardiovascular events. Formulas appropriate for SAE rates dependent on the denominator of patients as opposed to person-years are clearly presented and motivated. Both Bayesian and frequentist approaches are illustrated, with the author warning that both may be used by the DAC. With tables of summaries at the end of each chapter, as well as in the Appendix, and a detailed glossary of terms, the book serves as an easy-to-use reference that can be quickly accessed during the DMC meeting.

Perhaps my favorite part of the book was the DMCounselor sections at the end of each chapter, comprising actual questions to the author and his reply as consultant to DMCs. I gained a tremendous amount of expertise reading through this wide compendium of interesting case studies, feeling as if the author was a personal mentor for my lonely job as biostatistician of a DMC. Armed with the newfound knowledge and a reference at my fingertips, I went back to my DMC meetings as a new and improved chairperson.

Donna Pauler Ankerst
Department of Mathematics
Technical University Munich
Munich, Germany
ankerst@tum.de

RICHARD M. SIMON. **Genomic Clinical Trials and Predictive Medicine**. Cambridge: Cambridge University Press.

The recent development of whole genome biotechnology platforms and advances in molecular biology have helped realize the vision of precision medicine. Clinical development of therapeutics now faces with the paradigm shift toward precision or predictive medicine through incorporating biomarkers to select appropriate treatments for individual patients. This change, however, can largely complicate the clinical development of new treatments. Thus, the avoidance of misunderstanding, confusion, and possible misconduct in clinical trials for predictive medicine is of primary importance.

One of the contributions of this concise book can be seen in this respect, through providing the essence of statistical aspects for reliable clinical trials for predictive medicine in a readable manner for broad audiences, including clinical investigators, translational scientists, and students. Introductory materials on clinical trials in Chapter 1 and statistical background in Appendix A will effectively serve as back-up for audiences who are not familiar with clinical biostatistics. Chapter 2 provides basic concepts and principles for developing and validating prognostic biomarkers. Appendix B highlights pitfalls in handling high-dimensional genomics data to develop genomic classifiers or signatures, and provides appropriate methods to this end.

Another contribution of this book is to provide the "frontier" of adaptive designs for clinical trials with predictive biomarkers/genomic signatures to identify patents who benefit from the tested treatments. This aspect is particularly useful for biostatisticians and students in biostatistics. Chapter 3 on phase II trials discusses extensions of traditional single-arm designs, such as Simon's two-stage design, and randomized phase II designs, to incorporate a candidate predictive biomarker. Bayesian adaptive designs with multiple biomarkers, including that used in a platform trial, BATTLE-I, in advanced non-small cell lung cancer, are also discussed. Chapter 4 discusses the enrichment design for phase III trials, focusing on its efficiency properties in comparison with the traditional all-comers design without biomarkers. Chapters 5–7 discuss phase III all-comers designs with one or more predictive biomarkers. In contrast to the traditional approach to confirmative trials based on strong control (for multiple hypotheses across patient subpopulations), a new framework is advocated to incorporate the development of biomarkers in phase III trials. The goal becomes to test treatment efficacy with a strict alpha control under the *global null hypothesis* that the treatment is uniformly ineffective across patients and to develop an *indication classifier* (to identify the patients for whom the new treatment should be used) after demonstrating treatment efficacy. Then, various adaptive designs, including threshold- and (cross-validated) signature-designs, with development and validation of genomic signatures are discussed. Lastly, in Chapter 8, the prospective-retrospective approach to clinical trials that archive pre-treatment specimens for marker evaluation is discussed. This second half of the book on adaptive designs can be seen as the integration of traditional statistical inference and prediction analysis for establishing the new framework for the design and analysis of clinical trials for predictive medicine.

In summary, I recommend this very unique book with both introductory and cutting-edge materials in modern clinical trials toward predictive medicine for anyone involved in clinical trials or interested in this rapidly emerging challenging field of biostatistics.

Shigeyuki Matsui
Department of Biostatistics
Graduate School of Medicine
Nagoya University
smatsui@med.nagoya-u.ac.jp

ARDO VAN DEN HOUT. **Multi-State Survival Models for Interval-Censored Data**. Boca Raton: CRC Press.

Multi-state models are stochastic processes that represent transitions of an individual between discrete states. In statistical modelling, they are used for two general kinds of data: firstly, data where the state is known at any time during an individual's follow-up, and secondly, data where the state is only known at a finite series of times. In the literature on multi-state models, it is not always instantly clear which of the two situations are presumed, but the first case has perhaps been treated more extensively in textbooks (e.g., Andersen et al. 1993 and Beyersmann et al. 2011).

To my knowledge, this book comprises the first devoted to multi-state models for intermittently-observed data. As author of the "msm" R package for this class of models and data, I have seen how common and widespread they are, with applications in areas such as medicine, epidemiology, biology, ecology, economics, finance, and engineering. These methods have only previously been covered in scattered papers, software manuals and book chapters, so it is great to finally see a full textbook.

The book is motivated by applications in demography, disease, and survival, as shown by its title and structure, which starts with standard survival analysis for times to death. This focus on one area is probably a sensible choice to keep the book clear and concise, though inexperienced practitioners in other areas may face an extra hurdle. For instance, all the examples have an absorbing state representing death, when this is not necessary in a general multi-state model.

"Interval-censored" times of disease onset or state change are introduced in the third chapter. These are also called "intermittently observed" or "panel" data in other literature. Multi-state models with general state-transition patterns are then described. An important later chapter covers estimation of expected time spent in states such as healthy life. A range of advanced topics, such as frailty models, Bayesian inference, hidden Markov models and aggregate data are introduced fairly briefly, though with enough detail to get the substance across, along with references to other resources. In the Bayesian Chapter I would have liked to see more detail of how to specify informative priors on an intuitive scale. Also the JAGS Bayesian modelling software has useful built-in functions for this class of models (an "msm" module, including a matrix exponential function) which was not acknowledged.

The applicability of the models to real data is stressed throughout. Each new model is illustrated through one of several running examples related to long-term illness or ageing. The applied interpretations of the examples are discussed clearly. Though in the results tables, I would have preferred estimates of parameters such as hazard ratios to be presented on a natural scale, as they would be in a medical journal, rather than a log scale. Likewise, transition intensities are more easily interpreted in terms of sojourn times and next-state probabilities. While less mathematically convenient, transforming results to their most interpretable scale would be a simple step to bring theory and practice closer together.

The models advocated are flexible enough to cover all typical applications. Dependence of transition rates on age or time is emphasized, and modelling this dependence is

made straightforward through a novel piecewise-constant approximation method. This is an advance over methods available in current software. Sensible modelling choices, such as parsimony constraints on model parameters, are illustrated through the examples.

The writing style strikes a good balance between readability and mathematical rigor. Each new topic is generally begun with an approachable explanation, with formal definitions following later. A helpful appendix gives some useful algebraic results and example R implementations of the non-standard methods. The level is approximately suitable for a post-graduate statistics student or applied statistician. Applied practitioners without formal statistical training should therefore be warned that the statistical modelling is treated from first mathematical principles. The intended audience seems to be people accustomed to writing custom code based on algebraic definitions. The book even presents details of an optimization algorithm used for maximum likelihood estimation when efficient general-purpose optimizers are available in most statistical software.

I will be recommending this book to users of the "msm" package and my students, especially anyone modelling chronic diseases or age-dependent conditions. I hope that the software will eventually catch up with some of the advanced methods presented here, so that they become routinely used in practice!

References

Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing Risks and Multistate Models with R.* Springer. https://www.springer.com/gb/book/9781461420347

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer. https://www.springer.com/gb/book/9780387945194#otherversion=9781461243489

Christopher Jackson
Medical Research Council Biostatistics Unit
University of Cambridge
Cambridge, England
chris.jackson@mrc-bsu.cam.ac.uk

IAN FOSTER, RAYID GHANI, RON S. JARMIN, FRAUKE KREUTER, JULIA LANE. **Big Data and Social Science: A Practical Guide to Methods and Tools**. Boca Raton: CRC Press.

This practical guide to methods and tools for working with big data in the social sciences covers an array of topics in the general realm of data science. The book is organized in three parts that follow the classical path of social science research: (i) capture and curation of big data, (ii) modeling and analysis, and (iii) inference and ethics. Part (i) includes advice on web scraping and on transferring data using web protocols, such as application program interface (API); on linking data from multiple sources; on storing, organizing, and managing data with database management systems (DBMSs); and on programming for big data using Apache Hadoop. Part (ii) introduces unsupervised and supervised methods of machine learning; text analysis; and network analysis. Part (iii) covers descriptive visualization methods with Tableau; uncertainty and statistical inference with big data; and the new flavors of privacy and confidentiality issues that arise when researchers handle large quantities of data on individuals or organizations that are collected "in the wild." Most chapters include supporting workbooks that are provided on GitHub. The workbooks make extensive use of the Python programming language and the MySQL database management system.

It is clear that the editors have put a lot of thought into the structure and organization of this book. Several of the chapters are formed around a common theme of a social science research question—the science of science policy "use study." The overarching aim of the "use study" is to offer one possible pathway for obtaining answers to science of science policy questions—such as "How much should a nation spend on science?" and "What kind of science should be funded?"—through harnessing big data for measuring the impact of research and innovation on science and society. Chapters that adhere to the "use study" then describe big data tools and methods that could be used in that context. Thus, the book demonstrates how to collect information about grants and about people funded on grants from web sources of the funding agencies and universities, how to link information coming from different sources, how to store and organize such data in ways that allow for quick summarization and data exploration, as well for convenient extraction of data for further research (e.g., for describing contents of scientific works via text analysis).

The parts of the book with cautionary tales and advice regarding limitations of data science as an approach to carry out social science research should be required reading for all those involved in data science. Chapter 1 explores the question of generalizability and discusses some classic examples, such as the failures of the Literary Digest's prediction of the 1936 U.S. presidential election outcome (Squire, 1988) and the theory-free model of Google Flu Trends (David et al, 2014). Chapter 2 explains that data found online can "provide an existence proof that something has happened—but they can not, conversely, show that it has not." Social scientists may use high numbers of downloads and online views as evidence of research impact of a relatively under-cited article, whereas they may not use the lack of tweets about an article as evidence that the article is not being discussed. Chapter 10 discusses two important sources of error specific to big data: "big data hubris" or erroneous beliefs that one can disregard scientific methods when large quantities of data are available, and "algorithm dynamics" or changes to the data-generating processes over time. Chapter 11 makes a case that same characteristics of big data that make them useful and exciting for analysis (such as the ability to detect subtle correlations) are also responsible for making big data vulnerable with respect to breaches of privacy and confidentiality.

The publisher describes this book as providing "both traditional students and working professionals" with tools for hands-on approaches to handling big data in the social sciences, however, it is likely that different audiences will benefit from different aspects and parts of the book. Thus, audi-

ences who are not familiar with either the field of computer science or the emerging field of data science may read the book and find it useful as an introduction to both the general concepts and the language used by these communities. Concrete examples dispersed throughout the book provide valuable or curious insights. For example, readers learn that Apache Hadoop was named after Hadoop the toy elephant of the creator's son. On the other hand, professionals interested in gaining practical tools for working with APIs and databases will find reading the chapters on capture and curation of big data and working through the associated workbooks useful for learning the techniques. These chapters may also serve as a hands-on instructional component in a graduate (or potentially undergraduate) class in a social science department with a primary focus on text analysis, for example. Similarly, instructors of existing courses on privacy and confidentiality may benefit from incorporating the book's material on privacy issues with big data (Chapter 11), and instructors of sample survey research may expand beyond the clean framework of total error in survey data to new errors encountered in big data analytics (Chapter 10).

As a statistician who regularly teaches students in statistics and the social sciences, I hesitate recommending the book as a solo textbook for either an undergraduate or a graduate university course. The topics of statistical modeling, machine learning, and inference with big data deserve much more attention before they can be useful for practical social science research. Some of the tips for data analyses suggested in the book seem to follow the one-size-fits-all argument without mentioning the necessary consideration of a research question. Thus, Chapter 6 suggests the following practical rules of thumb: creating lots of features to be used in models and performing model selection and regularization instead of carefully selecting a few predictors, as well as using simple models with complex features rather than complex models with simple features. While these mechanistic approaches may likely be beneficial in most cases, a suggestion to social scientists to adopt such instructions for every possible research question is unscientific.

As a statistician whose research interests lie at the intersection with the social sciences, I was very much drawn to the editors' attempt to tie the exposition of big data tools and methods to a practical question. However, after reading the book, I found the particular choice of the science of science policy example unfortunate. The complexity of this question is enormous. It should not be a surprise that the book does not actually answer any of the questions on the science of science policy that are posed at the beginning. However, because readers do not get to see inference in the context of the "use study" example, its use feels somewhat artificial at the end.

Overall, the book could be useful to researchers and data analysts who would like to understand overarching ideas of data science and big data analysis. Social scientists interested in the topic will gain knowledge of basic steps required for working with big data and will deepen their understanding of the tools and the associated language used by the data science community. The book could also be used by instructors of graduate and undergraduate courses that touch on big data.

References

Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly* **52**, 125–133.

David, L., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science* **343**, 1203–1205.

Elena A. Erosheva
Department of Statistics
University of Washington
Seattle, Washington, U.S.A.
elena@stat.washington.edu

CLAUDIO CARINI, SANDEEP M. MENON, AND MARK CHANG, EDS. **Clinical and Statistical Considerations in Personalized Medicine**. Boca Raton: CRC Press.

This book introduces the reader to issues arising in research on biomarkers, short for "biological markers," that are useful for clinical applications and "personalized medicine." Various definitions of "biomarker" have been proposed. To give one here, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Biomarkers Definitions Working Group, 2001). Biomarkers are used in basic and clinical research and are often used as primary endpoints in clinical trials, when they have been well characterized and to correctly predict relevant clinical outcomes for various treatments and populations. The use of biomarkers measured using high throughput technologies in clinical research is somewhat newer, and the best approaches to this practice are still being developed and refined. Some of the issues arising in this setting are addressed in this book.

The first four chapters of the book edited by Carini et al. discuss novel biomarkers for drug development and their use in personalized medicine, and touch on issues arising in high throughput screening settings, such as large numbers of false positive findings. Chapter 2, also with a focus on drug development, discusses RNA interference (RNAi), a biological process in which RNA molecules inhibit gene expression or translation, leading to "gene silencing," and gives details on the various assays, measurement techniques and experimental design considerations. Similarly, Chapter 3 provides an overview on epigenetics, potentially heritable changes in gene expression that do not involve changes to the underlying DNA sequence. In Chapter 4, a brief overview of biomarkers for rare diseases is given, reiterating some of the material in the earlier chapters. While chapters 1 through 4 provide much background on biology and measurement techniques, Chapters 5 and 6 are of somewhat greater relevance for the statistical reader, discussing biomarker-informed adaptive designs, and providing R and SAS code. Chapter 7 gives an overview on clinical decision and design strategies for oncology with predictive biomarkers, and in some detail discusses phase 2 proof-of-concept trials, information needed to con-

tinue investigations, and design for phase 3 clinical trials. In Chapter 8, the complex issue of identifying "predictive" sub-groups of populations based on biomarkers from clinical trials is investigated. Chapter 9 addresses issues of multiple testing in drug development research. In Chapter 10, the role of patient reported outcomes (PROs) in personalized healthcare incorporating genomic information is discussed. The book concludes with Chapter 11 on regulatory issues in the use of biomarkers in drug development, and gives several examples.

The book is clearly aimed towards a diverse and not necessarily statistical audience, and has a strong emphasis on drug discovery. Consecutive chapters of the book can be read independently of the chronology, and for those most interested in statistical concepts, starting reading at Chapter 5 is recommended. Some of the diagrams and graphs are duplicated in the book, and some more careful editing would have improved presentation and readability.

The book will serve as an introduction to some genomic biomarkers and basic statistical concepts arising in drug discovery efforts using biomarkers measured on novel technologies, which are often high-throughput. The readers will need other sources if they wish to learn about the very latest statistical developments in this rapidly evolving field.

References

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* **69**, 89–95.

Ruth Pfeiffer
Division of Cancer Epidemiology & Genetics
Biostatistics Branch, National Cancer Institute
Shady Grove, Maryland, U.S.A.
pfeiffer@mail.nih.gov

DANIEL O. STRAM. **Design, Analysis, and Interpretation of Genome-Wide Association Scans**. Heidelberg: Springer.

In "*Design, Analysis, and Interpretation of Genome-Wide Association Scans*," Dr. Stram brings his broad knowledge on quantitative genetics and practical experience in statistical methodologies to the analysis of large-scale association studies. As the title suggests, this book focuses on designing, analyzing, and interpreting the results from genome wide association studies (GWAS) while providing considerable details about all facets of GWAS. The book covers basic biology, population and quantitative genetics, single marker association and haplotype analysis, topics on design of large-scale association studies, and post-GWAS analyses. Each chapter starts with a brief summary of its contents, followed by details on statistical inference, and ends with a summary of all key points that serves as a great resource if using the book as a reference. The target audience for this book is statisticians, epidemiologists, and students with the goal to provide them essential tools for designing large-scale genetic studies and analyzing the data. The first chapter introduces a data set from the Japanese American Prostate Cancer Study (Cheng et al., 2012), a large population-based two-stage GWAS of prostate cancer with over 215,000 individuals and 600,000 single nucleotide polymorphisms (SNPs) genotyped using Illumina.Human660W_Quad_v1 SNP arrays. This data set is then referenced throughout the book as a primary example to illustrate statistical methods. At the end of each chapter, a section with 3–12 well-chosen homework problems of varying difficulty level are provided to help readers gain better understanding of the concepts and more experience with analyzing real data. I use it for a graduate course on statistical methods for association mapping taken by students from both biology and statistics. I also use it as a reference book for a big data training workshop with participants in biomedical research. Students from both groups, with and without deep quantitative and computing background, enjoy reading the book due to the illustrations with real data and companion R code. This book is organized into eight coherent chapters that are roughly partitioned into three parts. The first two chapters review basic concepts in biology, genotyping technology, and genetics. The second part of the book, comprising Chapters 3–6, focuses on association analysis. The third part, Chapters 7 and 8, leans toward issues in study design and post-GWAS analyses. After a nice introduction to some basic biology and technology in Chapter 1, including types of genetic variation with special focus on SNPs, genotyping methods and GWAS related resources, Chapter 2 provides a comprehensive overview of concepts and tools in quantitative genetics to motivate large scale genetic association scans. These include, for example, the concepts of population heterogeneity, linkage disequilibrium, recombination and haplotypes, along with corresponding tools, such as principal component analysis (PCA), methods to estimate identical by descent (IBD) probabilities, kinship matrices, and the *Balding-Nichols* beta-binomial model (1995). The author implicitly restricts to disease-common variant hypotheses, and discusses common variants of the human genome based on the HapMap Project as well as the 1,000 Genomes Project. Together, Chapters 3 and 4 present single marker association analysis of independent subjects and populations with structure, respectively. Chapter 3 dives deeper into popular methods with technical details concerning association analysis in population-based GWAS. After a brief introduction of chi-square tests for disease status and genotype association, the author relates these tests to those in generalized linear models (McCullagh and Nelder, 1989) by focusing on normal and binomial distributions to model the phenotype in genetic association studies. The equivalence of these tests is illustrated using simulated data, and R code for the corresponding functions is provided. Meanwhile, the author summarizes the mathematical formulation of likelihood-based inference, covering *Wald* tests, likelihood ratio tests, score tests and sufficient statistics. One of the advantages of generalized linear models, naturally incorporating covariates as predictors, is illustrated by many real examples where covariates and their interactions with genetic variants are important to understand the disease risk. This chapter also introduces conditional logistic regression to handle population stratifica-

tion, case-only analyses, and ordinary least squares estimation for correlated phenotypes. Other important topics for GWAS, such as the required hardware and available software, multiple comparisons, non-independence tests, and reliability of small p-values, are discussed and illustrated with simulated data and figures. Chapter 4 extends the discussion of single marker association analysis to more complicated association analysis of data from structured populations. Following an extensive review of the effects of hidden structure, such as non-mixing subpopulations, incomplete admixture and relatedness between subjects, the author introduces four different methods to adjust for the effects of population stratification, namely genomic control, leading principal components, random effects models, and retrospective approaches, along with the available software packages. This chapter ends with conclusions on the basis of comparisons of these methods through several simulated data sets. The contents evolve from single marker analysis to haplotype analysis with a larger number of SNPs in Chapter 5. Discussion of haplotype analysis, a more powerful approach for detecting causal variants than single marker analysis, centers on three topics: haplotype frequency estimation, imputation, and association analysis. Regarding haplotype frequency estimation and imputation, the author describes an Expectation–Maximization (EM) algorithm and partition-ligation EM algorithm for large numbers of SNPs. Upon general discussion of a three-step expectation-substitution method for haplotype-based regression analysis, analysis strategies for genome-wide data and with multiple populations are considered. Similarly, in Chapter 6, we see a nice extension of association analysis from single markers to imputed SNP analysis. The basic statistical method, the hidden Markov model (HMM), is first introduced and its statistical inference is illustrated with R code. Then the author describes the utility of HMM in large-scale imputation, missing genotype imputation for phased data, and prediction of haplotype phase for a large number of SNPs and individuals. Other practical issues, such as imputation accuracy, rare SNP imputation, and reference panel selection, are touched at the end of the chapter. Chapter 7 concludes association analysis with the design of large-scale association studies, and two key intertwined themes, sample size and power calculations. I appreciate the details on designing large-scale association studies that often receives surprisingly scant attention. After comparing several methods for power calculation in single marker analysis, the author moves to general statistical approaches to calculate power. In addition, one of the most popular packages for power and sample size calculations, QUANTO (Gauderman and Morrison, 2006), is introduced for case-control studies and other types of design, such as sibling-matched and parent-affected-offspring designs. The author also briefly discusses other cases related to power and sample size calculations, including two-stage designs, control of multiple comparisons and population stratification, fine mapping with multi-markers, haplotypes and imputed SNP-based analyses. The last chapter is devoted to broader topics on post-GWAS analysis. First various aspects of meta-analysis to combine results from multiple studies are discussed (Evangelou and Ioannidis, 2013), followed by considerations for multiethnic groups, fine mapping and heritability estimation, and finally, topics related to whole genome sequencing. This book strikes an excellent balance between motivating concepts using concrete examples and statistical theory, as well as analysis tools for GWAS. The format of each chapter is consistent. The author starts with motivating examples and intuition, then clearly describes the mathematical details, outlines the algorithms followed by the R code, and ends with discussions and concluding remarks. Each chapter concludes with homework problems that help students to cement understanding of the statistical methods. After discussing a large number of different methods, the book contains a list of references at the end of each chapter that makes the book useful as an additional reference on GWAS. My favorite aspect of this book is the step-by-step R code, provided for the illustrative examples throughout the book, which enables readers to implement the methods described in the book. A companion website of the book with datasets, R codes, additional material and discussion, and errata can be found at: `http://www hsc.usc.edu/ stram/strambook.html`. I believe this website will be continuously updated as new information becomes available. Another appealing aspect is that Chapters 2 and 4 conclude with "*Data and Software Exercises*," whereby readers can practice using R programs, manipulating large datasets, and performing certain functions. These exercises will reinforce the learning experience of association analysis methods and make this book suitable for a one-semester graduate level course in large-scale genetic association analysis. In summary, this book is well organized, clearly written, and easy to read. It is destined to be a textbook for graduate students interested in learning statistical methods for GWAS data, and a must-have handbook for researchers to analyze GWAS data. My only minor remark is that there is no solution to the end-of-chapter homework problems. I would recommend this book to anyone working on large-scale genetic association studies as an informative reference manual.

References

Balding, D. and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic locus and its implications for investigating identity and paternity. *Genetica* **96**, 3–12.

Cheng, I., Chen, G. K., Nakagawa, H., He, J., Wan, P., Lurie, C. et al. (2012). Evaluating genetic risk for prostate cancer among Japanese and Latinos. *Cancer Epidemiology and Prevention Biomarkers* **21**, 2048–2058.

Evangelou, E., and Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics* **14**, 379–389.

Gauderman, W. and Morrison, J. (2006). QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. `http://hydra.usc.edu/gxe`

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.* (2nd edition). Boca Raton, FL: CRC Press.

Min Zhang
Department of Statistics, Purdue University
West Lafayette, Indiana, U.S.A.
minzhang@purdue.edu

BILL SHIPLEY. **Cause and Correlation in Biology: A Users Guide to Path Analysis, Structural Equations and Causal Inference with R, 2nd ed**. United Kingdom: Cambridge University Press.

I encountered the first edition of *Cause and Correlation in Biology* when researching graph theoretic approaches for analyzing spatially correlated ecological datasets as a Statistics PhD student. I was trudging through Judea Pearl's *Causality* and struggling with how to apply the concepts and theorems when analyzing multivariate datasets. The primary statistical literature did not provide a very thorough bridge to practical application either, in my opinion. Luckily, I then read Shipley's book. Both in the first and this new edition, Shipley's presentation is much easier to understand and elegant in its motivation of causal models as a means to deepen our scientific understanding for *how* or *why* a system is responding to hypothesized ecological drivers and stressors using observational data. When fellow researchers ask me about causal models or the meaning behind a directed acyclic graph (DAG, also sometimes referred to as path analysis, structural equation model (SEM), or graphical causal models), I always have and will continue to recommend Shipley's book as a first read. The overall presentation is accessible to non-statisticians and a great starting place for applied statisticians who can then follow up with more readings, if needed. The writing style is more a stream of consciousness from an instructor than a de-personalized textbook. The examples are typically biological in nature, but there are several that are not, and one request for future editions would be to translate them all into all biological/ecological applications.

This new edition provides critical additions from the last version, including detailed information for implementing models using R software and various packages. The main motivation behind the book is still the same—to show how causal hypotheses can be conveyed as a directed acyclic graph and then how to test these hypotheses using empirical observational data. The first chapter presents the crux of the matter, which is that "the causal processes generating our observed data impose constraints on the patterns of correlation that such data display." Chapter 2, then, introduces terminology of directed graphs and some of their fundamental properties that are needed to translate a causal graph into a statistical model: d-separation, Markov condition, and probabilistic (conditional) independence. A laundry list of counter-intuitive consequences and limitations of d-separation follow within sections 2.9–2.14. Then Chapters 3 and 4 show how to test causal hypotheses with empirical data when the causal relationships can be encoded within a directed graph in which all the nodes are random variables that are directly measured-path analysis models. Chapter 5 basically introduces a standard factor analysis model with latent factors that underlie a set of attributes that are measured. This chapter provides a nice bighorn sheep example that would likely resonate with biologists better than the more common psychology examples found in statistical textbooks. Chapter 6 introduces structural equation models and contains many new sections on "representing composite variables using latents" and "reporting results in publication." Chapter 7 then tackles the more familiar data issues encountered in observational datasets (clustering and lack of independent observations). In Chapter 8, Shipley covers a very important aspect of causal modeling, which is related to "searching" for correlational structure in multivariate data to generate causal hypotheses.

Not until the end of my read, did I begin to notice a potential disconnect with my personal philosophy of how to approach causal inference in ecology. If I had to sum up my philosophy I would characterize it as a descendant (hopefully, not illegitimate) of the seminal work by Burnham and Anderson (2002), who have made a huge impact on ecological data analyses and created a whole paradigm of multi-model inference in ecology, as well as the lineage of causal modeling as portrayed in Bill Shipley and Jim Grace's collective work (e.g., Grace et al. 2012, Grace et al. 2010). *What do I mean by this?* Well, I believe that ecologists need to work harder to justify their intensions for why they want to draw an arrow or claim a variable could produce an "effect." As Burnham and Anderson challenged biologists to think harder about their a priori set of models, I contend they should substantiate the connections among variables in a causal graph, and this is where my perspective diverges from that of Shipley's text. There were some passing comments (added in this edition) about how to justify causal reasoning behind the DAG investigated (p. 86 and p. 184), but in my experience working with biologists, this is a very difficult task and worthy of a stand-alone section. In line with this, I was hoping to see a discussion of Shipley's 2013 *Ecology* paper in which he attempted to tackle the question of how to choose among competing causal graphs and developed an AIC-like criterion for model comparison, but it was not referenced. The only model comparison discussion was with regard to nested models with common versus separate path coefficients (pp. 198–200). Given the ubiquitous use of multi-model inference in Ecology and now path analysis and structural equation models, there is a need for more guidance on how the paradigms of causal and multi-model inference can complement each other (or not).

The ecological literature is peppered with inherently causal terms (e.g., driver, effect) and I have always questioned the validity of their usage. I am not a strictly "correlation does not imply causation" statistician, but I do think ecologists, in particular, have become a little lax in their use of these words in articles and in an effort to advertise the importance of their work. To investigate if I am making a mountain out of a mole hill, I decided to do my own mini-inquiry. I used the Web of Science keyword search for the term "driver" or "effect" on July 10, 2017 in the journal "Ecology" from 2016 to current and found 280 articles, and then I randomly sampled 30. Eighteen used experimental manipulations (exclosures, laboratory, mesocosms, etc.), 1 a retrospective Before-After Control-Impact design, 3 referenced Shipley or another SEM reference and performed experiments, and 8 were observational studies with the usual statistical analyses (generalized linear models, etc). Only 27% (95% confidence interval: 11% to 43%) created the error I was anticipating- exploiting causal language without employing techniques as demonstrated in Shipley's book or Grace et al. 2012, 2010- lower than I expected, but definitely

leaving room for improvement. This finding, to me, underscores that there is a continued need for discourse on when ecologists and biologists can legitimately use these causal terms in manuscripts. My big criticism of the book is related to Chapter 8 entitled: "Exploration, discovery and equivalence." Shipley appears to give biologists the license to "search" using complex computer algorithms and then "correct" the edges found based on their expert opinion. In my experience, this type of exercise could quickly become data-snooping that is not acknowledged in the presentation of the final result. To be fair, Shipley does address confirmatory versus exploratory analyses and emphasizes repeatedly whether he is testing or searching for causal hypotheses throughout Chapter 8, but I still was not satisfied. I think, in general, more discussion on how to use causal inference appropriately in situations with multiple competing causal graphs (confirmatory) versus those that search for the causal graph (exploratory) and the blurry lines that divide them would be fruitful. In summary, I would like to commend Bill Shipley on his excellent resource for biologists and ecologists; it was a pleasure to review. The fun part for me was it rekindled my interest in exploring the use (and abuse) of causal inference and language in ecology.

References

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Second edition. New York, NY: Springer-Verlag.

Grace, J. B., Anderson, T. M., Olff, H., and Scheiner, S. M. (2010). On the specification of structural equation models for ecological systems. *Ecological Monographs* **80**, 67–87.

Grace, J. B., Schoolmaster Jr., D. R., Guntenspergen, G. R. , Little, A. M., Mitchell, B. R., Miller, K. M., et al. (2012). Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere* **3**, 73.

Pearl, J. (2009). *Causality, Models, Reasoning, and Inference*. second edition. New York, NY: Cambridge University Press.

Shipley, B. (2013). The AIC model selection method applied to path analytic models compared using a d-separation test. *Ecology* **94**, 560–564.

Disclaimer: Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Kathryn M. Irvine
Research Statistician Northern Rocky
Mountain Science Center, US Geological Survey
Bozeman, Montana, U.S.A.
kirvine@usgs.gov