Biometrics 65, 326–339
March 2009

BOOK REVIEWS

EDITOR: THOMAS M. LOUGHIN

Semiparametric Theory and Missing Data
(A. A. Tsiatis) *Andrea Rotnitzky*

Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis
(M. J. Daniels and J. W. Hogan) *Daniel F. Heitjan*

Bayesian Biostatistics and Diagnostic Medicine
(L. D. Broemeling) *Paul Gustafson*

Statistics in the Pharmaceutical Industry, 3rd edition
(C. R. Buncher and J.-Y. Tsay, Editors) *Ralph B. D'Agostino Jr.*

Introduction to Machine Learning and Bioinformatics
(S. Mitra, S. Datta, T. Perkins, and G. Michailidis) *Yulan Liang*

The Statistics of Gene Mapping
(D. Siegmund and B. Yakir) *Hongyu Zhao*

DNA Methylation Microarrays: Experimental Design and Statistical Analysis
(S.-C. Wang and A. Petronis) *Kimberly D. Siegmund*

Multiple Testing Procedures with Applications to Genomics
(S. Dudoit and M. J. van der Laan) *Ruth Heller*

The Statistical Analysis of Functional MRI Data
(N. A. Lazar) *Wesley K. Thompson*

Simulation and Inference for Stochastic Differential Equations with R Examples
(S. M. Iacus) *Dave Campbell*

Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
(N. Markovich) *M. Ivette Gomes*

Time Series Analysis with Applications in R, 2nd edition
(J. D. Cryer and K.-S. Chan) *Timothy D. Johnson*

Brief Reports by the Editor

Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences
(C. P. Doncaster and A. J. H. Davey)

Computational Statistics Handbook with MATLAB®, 2nd edition
(W. L. Martinez and A. R. Martinez)

Models for Probability and Statistical Inference: Theory and Applications
(J. H. Stapleton)

Medical Biostatistics, 2nd edition
(A. Indrayan)

Computational Methods in Biomedical Research
(R. Khattree and D. N. Naik, Editors)

TSIATIS, A. A. **Semiparametric Theory and Missing Data**. Springer, New York, 2006. xvi + 383 pp. $95.00/£74.85. ISBN 9780387324487.

In most studies, the intended (full) data, i.e., the data that the study investigators wish to collect, are inevitably incompletely observed. In modern studies, the full data are typically high dimensional, usually comprising many baseline and time-varying variables. Scientific interest, however, often focuses on some low-dimensional parameter of the distribution of the full data. Specification of realistic parametric models for the mechanism generating high-dimensional data is most often very challenging, if not impossible. Nonparametric and semiparametric models, i.e., models in which the data-generating process is characterized by parameters ranging over a large, non-Euclidean, space and, possibly, also a few meaningful real-valued parameters, meet the analytic challenge posed by these high-dimensional data because they do not make assumptions about the components of the full data distribution that are of little scientific interest. Analytic strategies based on semiparametric models avoid the possibility of incorrect inferences due to misspecification of models for the secondary parts of the full data law.

Drawing from the modern theory of semiparametric efficient inference developed since the 1980s, Robins and Rotnitzky (1992) derived a general estimating equations methodology in coarsened, i.e., incompletely observed, data models under non- or semiparametric models for arbitrary full data configurations. This methodology, based on the geometry of scores and influence functions, applies when the even serve as a nice introduction to more advanced books in the general theory of efficient estimation in semiparametric models. The material covered in the book will appeal to statisticians who wish to pursue research in methods for missing and censored data but it will also be valuable for anyone wishing to sample the area. It will be a good textbook for a one-semester introductory course in semiparametric estimating function methodology for missing data. In conclusion, the book is highly valuable.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 305–334.

Kosorok, M. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Liang, K. Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology—Methodological Issues*, N. Jewell, K. Dietz, and V. Farewell (eds), 297–331. Boston: Birkhäuser.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87,** 113–124.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer Verlag.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85,** 935–948.

Andrea Rotnitzky
Department of Economics
Universidad Di Tella
Buenos Aires, Argentina
Department of Biostatistics
Harvard School of Public Health
Boston, U.S.A.

DANIELS, M. J. and HOGAN, J. W. **Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis**. Chapman & Hall/CRC, Boca Raton, Florida, 2008. xx + 303 pp. $79.95/£42.99. ISBN 9781584886099 (Hardcover).

The introduction of SAS Proc Mixed in the early 1990s sparked a revolution in statistical practice. For the first time, statisticians could readily estimate a wide range of models from longitudinal data sets, even those beset with arbitrary patterns of missingness. The price of admission to this world is the assumption (typically untestable) that the dropout mechanism is missing at random (MAR), failure of which could cause likelihood-based inferences to go badly wrong. Thus a major thrust of research in the intervening years has been to develop methods that accommodate departures from MAR.

The new monograph by Daniels and Hogan offers a timely and thorough review of this maturing research area. The book comprises three parts: An introductory section describes a Bayesian strategy for longitudinal modeling with complete data; the second part adapts the modeling to MAR incomplete data; and the final section offers a paradigm for analysis of sensitivity to departures from MAR. The authors illustrate the methods in an array of examples drawn from recent clinical trials and observational studies in psychiatry, geriatrics, smoking cessation research, and HIV medicine.

The book is comprehensive in covering models for both continuous and discrete outcomes from both the pattern mixture and selection modeling perspectives. Analyses of sensitivity to nonignorability are organized around the identification of a sensitivity parameter—i.e., a parameter whose value affects the prediction of missing observations but not the fit of the model. Thus a complete expression of the authors' perspective in the analysis of a particular set of data would involve positing a model for the underlying complete data and dropout process; estimating it under MAR; then isolating the sensitivity parameter and constructing a prior for it, executing an adjusted analysis by averaging over this source of uncertainty.

The book's composition offers much to admire. The writing is clear and direct, the notation is sensible and consistent, and tables and figures are simple and uncluttered. Typos are mercifully rare; I estimate about one every eight pages.

I have only two criticisms worth mentioning: First, I am less sanguine than the authors about the possibility of using "expert opinion" and "historical data" to inform the prior distributions of nonignorability parameters in Bayesian sensitivity analysis. The problem is that relevant "historical data" seldom exist, and without such data, it is difficult to regard the opinions of even renowned experts as anything more than conjecture. Second, with the methodology's reliance on sampling-based inference procedures, readers may question the authors' decision to limit discussion of computing to an 11-page overview in Chapter 3. By the time one reaches the novel data analyses, there is seldom more than a perfunctory reference to the lengths of the Markov chains and the burn-in periods. Moreover, I was unable to find a link to the example WinBUGS code that the preface claims is posted on Professor Hogan's website.

Because this is a research monograph, the descriptions of background material take the form of brief but thoughtful synopses, and there are no exercises. The book would be ideal for self-study or as the text for an advanced course in longitudinal modeling. A more applied or basic course could use it as a supplement.

Dropout of uncertain provenance evidently is no less common a feature of longitudinal studies today than it was in the early 1990s. Fortunately we now possess a much better understanding of how to extract valid statistical inferences from such data. Biostatisticians who seek a clear and thorough overview of the state of knowledge in this area would do well to make this excellent book their first stop.

Daniel F. Heitjan
Department of Biostatistics & Epidemiology
University of Pennsylvania
Philadelphia, Pennsylvania, U.S.A.

BROEMELING, L. D. **Bayesian Biostatistics and Diagnostic Medicine**. Chapman and Hall/CRC Press, Boca Raton, Florida, 2007. xii + 198 pp. $79.95/£42.99. ISBN 9781584887676.