Springer Series in Statistics

Springer Series in Statistics

Alho/Spencer: Statistical Demography and Forecasting
Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes
Atkinson/Riani: Robust Diagnostic Regression Analysis
Atkinson/Riani/Ceriloi: Exploring Multivariate Data with the Forward Search
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications, 2nd edition
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition
Bucklew: Introduction to Rare Event Simulation
Cappé/Moulines/Rydén: Inference in Hidden Markov Models
Chan/Tong: Chaos: A Statistical Perspective
Chen/Shao/Ibrahim: Monte Carlo Methods in Bayesian Computation
Coles: An Introduction to Statistical Modeling of Extreme Values
Devroye/Lugosi: Combinatorial Methods in Density Estimation
Diggle/Ribeiro: Model-based Geostatistics
Dudoit/Van der Laan: Multiple Testing Procedures with Applications to Genomics
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications
Eggermont/LaRiccia: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation
Fahrmeir/Tutz: Multivariate Statistical Modeling Based on Generalized Linear Models, 2nd edition
Fan/Yao: Nonlinear Time Series: Nonparametric and Parametric Methods
Ferraty/Vieu: Nonparametric Functional Data Analysis: Theory and Practice
Ferreira/Lee: Multiscale Modeling: A Bayesian Perspective
Fienberg/Hoaglin: Selected Papers of Frederick Mosteller
Frühwirth-Schnatter: Finite Mixture and Markov Switching Models
Ghosh/Ramamoorthi: Bayesian Nonparametrics
Glaz/Naus/Wallenstein: Scan Statistics
Good: Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition
Gouriéroux: ARCH Models and Financial Applications
Gu: Smoothing Spline ANOVA Models
Gyöfi/Kohler/Krzyźak/Walk: A Distribution-Free Theory of Nonparametric Regression
Haberman: Advanced Statistics, Volume I: Description of Populations
Hall: The Bootstrap and Edgeworth Expansion
Härdle: Smoothing Techniques: With Implementation in S
Harrell: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis
Hart: Nonparametric Smoothing and Lack-of-Fit Tests
Hastie/Tibshirani/Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications
Heyde: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation
Huet/Bouvier/Poursat/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples, 2nd edition
Ibrahim/Chen/Sinha: Bayesian Survival Analysis
Jiang: Linear and Generalized Linear Mixed Models and Their Applications
Jolliffe: Principal Component Analysis, 2nd edition
Knottnerus: Sample Survey Theory: Some Pythagorean Perspectives

(continued after index)

Peter X.-K. Song

Correlated Data Analysis: Modeling, Analytics, and Applications

Springer

Peter X.-K. Song
Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West
Waterloo, Ontario, Canada N2L 3G1
song@uwaterloo.ca

Library of Congress Control Number: 2007929730

ISBN 978-0-387-71392-2 e-ISBN 978-0-387-71393-9

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To Ru and Angela

Preface

This book, like many other books, was delivered under tremendous inspiration and encouragement from my teachers, research collaborators, and students. My interest in longitudinal data analysis began with a short course taught jointly by K.Y. Liang and S.L. Zeger at the Statistical Society of Canada Conference in Acadia University, Nova Scotia, in the spring of 1993. At that time, I was a first-year PhD student in the Department of Statistics at the University of British Columbia, and was eagerly seeking potential topics for my PhD dissertation. It was my curiosity (driven largely by my terrible confusion) with the generalized estimating equations (GEEs) introduced in the short course that attracted me to the field of correlated data analysis. I hope that my experience in learning about it has enabled me to make this book an enjoyable intellectual journey for new researchers entering the field. Thus, the book aims at graduate students and methodology researchers in statistics or biostatistics who are interested in learning the theory and methods of correlated data analysis.

I have attempted to give a systematic account of regression models and their applications to the modeling and analysis of correlated data. Longitudinal data, as an important type of correlated data, has been used as a main venue for motivation, methodological development, and illustration throughout the book. Given the many applied books on longitudinal data analysis already available, this book is inclined more towards technical details regarding the underlying theory and methodology used in software-based applications. I hope the book will serve as a useful reference for those who want theoretical explanations to puzzles arising from data analyses or deeper understanding of underlying theory related to analyses. This book has evolved from lecture notes on longitudinal data analysis, and may be considered suitable as a textbook for a graduate course on correlated data analysis.

This book emphasizes some recent developments in correlated data analysis.

First, it takes the perspective of Jørgensen's theory of dispersion models for the discussion of generalized linear models (GLMs) in Chapter 2. It

is known that the class of generalized linear models plays a central role in the regression analysis of nonnormal data. In the context of correlated data analysis, these models constitute marginal components in a joint model formulation. One benefit from such a treatment is that it enables this book to cover a broader range of data types than the traditional GLMs. Two types that are of particular interest and discussed in detail in the book are compositional (or continuous proportional) data and directional (or circular) data.

Second, it gives a systematic treatment for the theory of inference functions (or estimating functions) in Chapter 3. The popular GEE methods presented in Chapter 5 are then easily introduced and studied as a special class of inference functions. Building upon Chapter 3, some alternative estimating function methods can be readily discussed. Recent work on quadratic inference functions (QIF) is an example that benefits from Chapter 3.

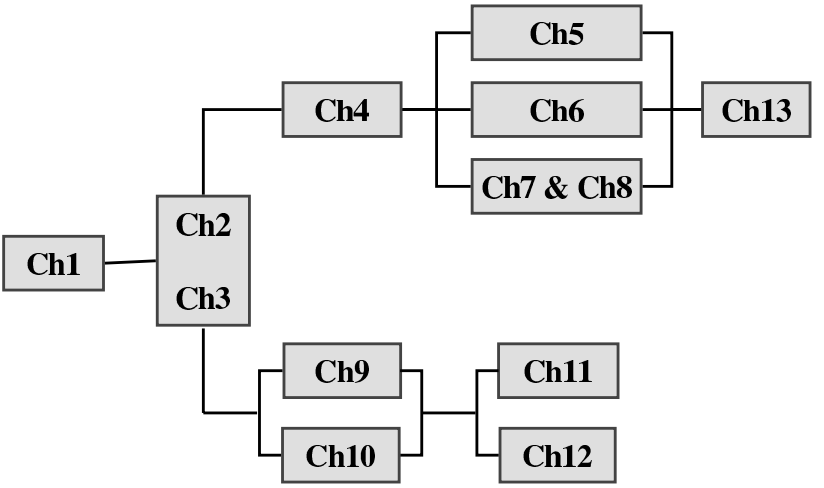
Third, it presents a joint modeling approach to regression analysis of correlated data via the technique of parametric copulas. Copulas are becoming increasingly popular in the analysis of correlated data, and Chapter 6 focuses on Gaussian copulas, for which both theory and numerical examples are illustrated.

Fourth, it deals with state space models for longitudinal data from long time series. In contrast to longitudinal data from short time series, modeling stochastic patterns or transitional behaviors becomes a primary task. In such a setting, asymptotics may be established by letting the length of the time series tend to ∞, as opposed to letting the number of subjects tend to ∞, as in the case of data consisting of many short time series. Chapters 10, 11, and 12 are devoted to this topic.

Fifth, this book covers two kinds of statistical inferences in generalized linear mixed effects models (GLMMs): maximum likelihood inference in Chapter 7 and Bayesian inference based on Markov Chain Monte Carlo (MCMC) in Chapter 8. In Chapter 8, the analysis of multi-level data is also discussed in the framework of hierarchical models. Inference can be dealt with easily by the MCMC method, as an extension from the GLMMs with little extra technical difficulty.

The book contains some other topics that are highly relevant to the analysis of correlated data. For example, Chapter 13 concerns missing data problems arising particularly from longitudinal data.

The presentation of some material in the book is a little technical in order to achieve rigor of exposition. Readers' backgrounds should include mathematical statistics, generalized linear models, and some knowledge of statistical computing, such as represented R and SAS software. The following chart displays the relationship among the thirteen chapters, and readers can follow a particular path to reach a topic of interest.



A webpage has been created to provide some supplementary material for the book. The URL address is

`http://www.stats.uwaterloo.ca/~song/BOOKLDA.html`

All data sets used in the book are available. A SAS Macro QIF is available for a secured download; that is, an interested user needs to submit an online request for permission in order to download this software package. In addition, some figures that are printed in reduced size in the book are supplied in their full sizes. Exercise problems for some of the thirteen chapters are posted, which may be useful when the book is used as a text for a course.

I would like to acknowledge my debt to many people who have helped me to prepare the book. I was fortunate to begin my research in this field under the supervision of Bent Jørgensen, who taught me his beautiful theory of dispersion models. At UBC, I learned the theory of copulas from Harry Joe. This book has benefited from some of the PhD theses that I supervised in the past ten years or so, including Zhenguo (Winston) Qiu, Dingan Feng, Baifang Xing, and Peng Zhang, as well as from a few data analysis projects that graduate students did in my longitudinal data analysis course; thanks go to Eric Bingshu Chen, Wenyu Jiang, David Tolusso, and Wanhua Su. Many graduate students in my course pointed out errors in an early draft of the book. Qian Zhou helped me to draw some figures in the book, and Zichang Jiang worked with me to develop SAS MACRO QIF, which is a software package to fit marginal models for correlated data.

I am very grateful to my research collaborators for their constant inspiration and valuable discussions on almost every topic presented in the book. My great appreciation goes to Annie Qu, Jack Kalbfleisch, Ming Tan, Claudia Czado, Søren Lundbye-Christensen, Jianguo (Tony) Sun, and Mingyao Li. I would also like to express my sincere gratitude to people who generously provided and allowed me to analyze their datasets in the book, including John

Petkau and Angela D'Elia. Zhenguo Qiu, Grace Yi, and Jerry Lawless provided with me their valuable comments on drafts of the book. My research in the field of correlated data analysis has been constantly supported by grants from the Natural Sciences and Engineering Research Council of Canada. I thank John Kimmel and Frank Ganz from Springer for their patience and editorial assistance.

I take full responsibility for all errors and omissions in the book. Finally, I would like to say that given the vast amount of published material in the field of correlated data analysis, the criterion that I adopted for the selection of topics for the book was really my own familiarity. Because of this and space limitations, some worthwhile topics have no doubt been excluded. Research in this field remains very active with many new developments. I would be grateful to readers for their critical comments and suggestions for improvement, as well as corrections.

Waterloo, Ontario, Canada *P.X.-K. Song*
December 2006

Contents