Preface

0.1 Overview

The generalized estimating equation (GEE) approach (Liang and Zeger, 1986) has become one of the most popular methods for modeling longitudinal and clustered data, particularly in the area of biostatistics. Its popularity stems from its usefulness. GEE extends generalized linear models (GLM) so that correlated data might be appropriately modeled. This is accomplished by relaxing the assumption of independence of measurements, which is violated when the data are longitudinal or in a clustered or panel form.

Quasi-least squares (SQLS) is a two stage computational approach for estimation of the correlation parameters that is in the framework of GEE. Stage one of QLS was proposed for balanced data by Chaganty (1997) and for unbalanced and unequally spaced data by Shults and Chaganty (1998), while stage two was proposed by Chaganty and Shults (1999). As we shall see in this monograph, QLS extends the application of GEE by allowing for easier consideration of new patterns of association, improved estimation of the regression and correlation parameters in some situations, and an alternative option for analysis should GEE fail to converge.

We first describe GLM and GEE, which comprise the foundation for QLS. We then discuss various limitations of GEE and provide a thorough description of the theoretical underpinnings of the QLS approach that highlights the relative advantages of QLS. We describe how to implement the standard correlation structures that are available for GEE, followed by newer structures that are plausible for unequally or equally spaced longitudinal data, familial data, and data with multiple sources of correlation. The latter structures are not currently available in the major software implementations of GEE.

Chapter 7 considers a recent topic of debate in the statistical literature, which is whether semi-parametric methods such as QLS and GEE are even appropriate for the analysis of discrete data. The problem is that although semi-parametric approaches only require specification of the first two moments of the underlying distribution, for discrete data, it is possible to obtain estimates for which there can be no valid distribution with the estimated means and correlations. We discuss the theoretical and practical implications of this issue in the context of logistic regression for longitudinal data. Our assessment includes a comparison between QLS and the recently developed first-order Markov

maximum-likelihood (MARK1ML) approach for analysis of longitudinal binary data.

Chapter 8 offers an evaluation of methods for selection of the best-fitting correlation structure. This is important because the enhanced ability of QLS to implement new patterns of association requires accompanying approaches to choose between the structures. We implement working correlation structures that were not considered in prior comparisons of methods for selection of a working structure. Our comparisons utilize a recently developed approach for simulation of discrete data with decaying product correlations.

The final chapter provides a discussion of sample size calculation, and a worked example demonstrating the application of QLS for estimation of model parameters, assessment of the fit of several candidate working correlation structures, and sample size calculation for planning a future study.

0.2 Intended Audience

This text provides the means to conduct an improved GEE analysis for study designs that are commonly encountered in research, and should therefore be a valuable resource for researchers who would like to appropriately account for the correlation in their data using an approach that is straightforward to apply. However, we also provide a thorough description of the theoretical underpinnings of QLS, so that our intended audience for this book includes both statisticians (and perhaps statistically sophisticated non-statisticians) who are interested in applying QLS in their analyses, in addition to statisticians who might be interested in working on some of the open research problems in this area. Those who are primarily interested in implementing QLS and GEE could skim over the theoretical developments, and head straight to the worked examples that are offered at the end of each chapter.

With the approval of our editor Rob Calver, the first-author used the manuscript-version of this book as the text for an Advanced Elective course for students in the Ph.D. program in Biostatistics at the University of Pennsylvania. Her positive experience using the manuscript for the course indicates that this monograph may be suitable as a text for advanced graduate students who are enrolled in a Ph.D. program in Biostatistics, or in Statistics. Suggested prerequisites for the course include Linear Models, with some knowledge of longitudinal data analysis (and in particular GEE) being helpful. However, if the instructor focuses on the worked examples in each chapter, this monograph could serve as a supplemental, or even primary, text for a course on analysis of correlated data, or perhaps for a course on improved GEE based-analyses of longitudinal and clustered data. Problems are included at the end of each chapter focusing on theory and applications. The instructor may select particular problems that are most appropriate for his or her students.

0.3 Software

We have primarily used Stata statistical software, version 13.0 (2013) for demonstrating example modeling output. User-authored software using Stata's proprietary programming language was developed by Shults et al. (2007) for general QLS estimation; extensions of this software to R, SAS, and MATLAB are also available (Xie and Shults, 2009; Kim and Shults, 2010; Ratcliffe and Shults, 2008).

Readers are also encouraged to refer to the web-site for this monograph, where we will post updates and accompanying code and guidelines to replicate many of the examples in this book in Stata, R, SAS, and MATLAB, `https://dbe.med.upenn.edu/biostat-research/Book-QLS`. Additional information and errata will also be available on Professor Shult's BePress Selected Works Site, `http://works.bepress.com/justine_shults`. Resources will also be provided on the publisher's web-site for the book, `http://www.crcpress.com/product/isbn/9781420099935`.

0.4 Acknowledgments

We gratefully acknowledge the National Cancer Institute of the National Institutes of Health, which provided the funding for Professor Shults, for the Longitudinal Analysis for Diverse Populations Project (LADP Project R01-CA096885). The goal of the LADP project was to develop more efficient and cost-effective methods for analysis of longitudinal studies using quasi-least squares (QLS), with special emphasis on studies in diverse populations that included community based interventions. Many of the results in this text stem from the LADP project that is described on the web-site for the project, `https://dbe.med.upenn.edu/biostat-research/ladp`.

We also wish to acknowledge the assistance and insight from a number of individuals who have been associated with our work. J. Shults is extremely grateful to Ardythe Morrow for first encouraging her and her dissertation advisor, N. Rao Chaganty, to learn more about GEE. She also thanks Professor Chaganty and his former students Genming Shih, Deepak Mav, Yihao Deng, and Roy Sabo, for their research on QLS and GEE. She thanks Professor Joe Gastwirth for being a wonderful mentor over the years. J. Shults also thanks James Hardin for developing the Stata command **xtgee**; she used **xtgee** in the **xtqls** command, to solve the GEE estimating equation for the regression parameter at the current QLS estimate of the correlation parameter.

J. Shults is also extremely grateful to her many colleagues at the University of Pennsylvania, and in particular those who worked with her on the LADP project and on research concerning QLS, including her close colleague and good friend Sarah Ratcliffe, who developed MATLAB software for QLS for longitudinal data and for data with multiple sources of correlation. She is also grateful to Mary Leonard, a pediatric nephrologist and clinical epidemiologist

with whom she has worked closely on statistical issues involved in the longitudinal analysis of bone density, structure, and strength since 1999. She is also grateful to her fellow Co-Director of the Pediatric Section in the Department of Biostatistics, Russell Localio, for his generous advice and expertise on methods for analysis of longitudinal data. She is also very grateful to Jimbo Chen, Scarlett Bellamy, Carissa A. Mazurick, and Richard Landis.

As a faculty member in the Department of Biostatistics and Epidemiology at the University of Pennsylvania, J. Shults has had the honor of working with many excellent graduate students. She thanks Jichun Xie for her work on the development of the R package **qlspack** and on the implementation of QLS for familial data that is featured in chapter 6; Matthew White, for his work on comparison of methods for selection of a correlation structure for GEE and QLS; Seunghee Beck, Chia-Hao Wang, Xiaoying Wu, Jiwei He, Arwin Thomassen, and Yimei Li for exploring issues related to QLS in their Master's thesis in Biostatistics at Penn. She also thanks former student Wenguang Sun for his work on comparison of QLS with other methods that utilize unbiased estimating equations for the regression parameter. She also thanks Qian Wu (Vicky) for useful comments she provided on this text when she was enrolled in the Advanced Elective course on QLS at Penn.

J. Shults is also very grateful to her first PhD student, Hanjoo Kim, for his work on extending QLS for unbalanced data with multiple sources of correlation and for his development of software for QLS in SAS for data with one and multiple sources of correlation. She is thankful for Hanjoo's continued enthusiasm to collaborate on research, some of which is cited in this text. She also thanks her second PhD student, Matthew W. Guerra, in particular for his input regarding the comparisons with MARK1ML, a recently developed approach that will be featured in their forthcoming book on logistic regression for correlated data (Guerra and Shults, 2014).

J. Shults would also like to acknowledge two faculty members who sadly, were lost recently. She is extremely grateful to former Professor Dayanand Naik for his ground-breaking work on multivariate analysis, and for setting a wonderful example in terms of kindness and strength of character. She is also exceedingly grateful to former Professor Thomas Ten Have for his guidance and collaboration on issues related to QLS, and for also setting a wonderful example as an outstanding researcher with integrity.

J. Hilbe wishes to acknowledge his long working relationship and friendship with Professor James Hardin, with whom he has co-authored 5 texts on GLM and GEE, as well as a number of encyclopedia and journal articles. Texts include Hardin & Hilbe, *Generalized Estimating Equations* (2003, 2013), Chapman Hall/CRC, and *Generalized Linear Models and Extensions* (2001, 2007, 2012), Stata Press- CRC. Professor Hardin and Hilbe co-authored the current version of Stata's **glm** command in 2001, which was a revision of Hilbe's initial version **glm**, first published in January 1993.

Both authors would also like to thank several additional people, whose valuable insights are much appreciated. We are extremely grateful to our patient and supportive editor Rob Calver, who has provided excellent advice and guidance throughout the writing process. We really appreciate all the time he put into working with us on this monograph, and enthusiastically recommend him as an editor to other statisticians who might contemplate writing a text.

In addition, special thanks are due to our external reviewers. Professor Preisser's detailed comments led to many improvements throughout the monograph, in addition to a new section on the characteristics of clustered and longitudinal data (Section 5.1). Overall, his reviews were extremely detailed and thoughtful, and we are very grateful for his efforts on behalf of the text. Professor Adrian Barnett also provided thoughtful and detailed reviews regarding the content that led to substantial improvements in the discussions and to important changes in the text, e.g. to improve the clarity of explanations. Thanks are also due to Shashi Kumar of Glyph International, who provided thoughtful advice regarding LaTeX.

We also wish to acknowledge and thank our spouses and children. The first author's primary responsibility for writing the chapters came at the expense of time spent with her family. She is therefore exceedingly grateful to her wonderful husband and children, Chuck, Chucky, and Erika, in addition to her former dogs Bud (who almost made it to the age of 16), and Noble (a sweet pit bull, originally from the Philadelphia SPCA, who recently lost his battle with lymphoma). The second author also wishes to thank the members of his family for their patience while he once again worked on a book. Such an effort takes away time that he would otherwise have spent with them. His dedication and appreciation goes to his wife, Cheryl, and to Sirr, a small white Maltese dog who would have preferred that his owner spend time with him instead of working at the computer.