

Preface

Before the beginning of years
There came to the making of man
Time with a gift of tears,

–Algernon Charles Swinburne

If we offend, it is with our good will.
That you should think, we come not to offend,
But with good will. To show our simple skill,
That is the true beginning of our end.

– William Shakespeare

Longitudinal data occurs when we repeatedly take the same type of measurement across time on the subjects in a study. My purpose in writing this textbook is to teach you how to think about and analyze longitudinal data.

As a graduate student, I joined the American Statistical Association and began to subscribe to professional journals. I was aware that most people did not read their journals, and in the natural exuberance of early graduate-student-hood I vowed to be different. I opened my first journal with the express intent to read it cover to cover; and quickly discovered not every article was interesting. However, I did read one article thoroughly. Ware (1985) had published an article titled “Linear Models for the Analysis of Longitudinal Studies.” I spent a lot of time trying to understand that article, and in a real sense I am still working on it today. This book is the outcome of my interest in longitudinal data, that began with that article.

Why This Book?

This is a textbook, not a monograph. Included material must be directly helpful when analyzing longitudinal data. Mathematical presentation is kept to a minimum although not eliminated, and statistical computing is not covered.

This book has several key features that other books on longitudinal data do not have. First of all, this book has chapter-length treatments of graphical methods, covariance modeling, and modeling the effects of covariates. These chapters are often only a small section in most other texts currently on the market. The effects of covariates requires at least one full chapter on top of what students have learned about covariates from their linear regression courses.

Many current texts are unbalanced in their coverage of this material. Many texts spend a lot of space on discrete data analysis—an entertaining and important topic. However, like courses on linear regression and generalized linear regression, students should cover linear regression in depth before moving on to logistic and Poisson regression. One book spends more than 25% of its space on missing data modeling. Understanding missing data and bias is an important part of statistical data analysis of longitudinal data. I do provide an introduction to missing data here, but first students need to know how to model regular longitudinal data before spending time learning about missing data.

Texts on longitudinal data from the 1980s and even 1990s are already out of date, usually concentrating on generalizations of analysis of variance rather than on generalizations of regression. The techniques they cover are often archaic. There are also several doctoral-level monographs on longitudinal data that cover multivariate analysis at a more advanced mathematical level, usually including substantial effort on computation and inference, but this is at the expense of not covering the nuts and bolts of data analysis, and those books cannot be read by master's-level students.

A number of texts treat longitudinal data as a special case of repeated measures or hierarchical or multi-level data. Those books emphasize the random effects approach to modeling to the detriment of other covariance models. Random effects models are powerful and flexible, and several sections of this text are devoted to random effects models. However, polynomial random effects models often do not provide the best fit to longitudinal data. Consequently, I treat random effects models as one of several covariance models to be considered when modeling the covariance matrix of longitudinal data.

Computation

I assume that computation will be handled by a software package. Statistical textbooks at the master's level typically do not cover statistical computation, and this book is no exception. My discussion of computation

tries to aid the data analyst in understanding what the software does, why it may or may not work, and what implications this has for their own data analysis. I do not discuss code for particular packages because software changes too rapidly over time. It is altered, often improved, and eventually replaced. I am thankful to the vendors that supply programs for analyzing longitudinal data, and I wish them a long and successful run. Extensive software examples will be available on the book's Web site. A link to the book Web site will be located at <http://www.biostat.ucla.edu/books/mld>. You will find data sets, example code, example homework problem sets, computer labs, and useful longitudinal links.

Initially, example code for fitting these models in SAS[®] Proc Mixed[®] and Proc Nlmixed[®] will be available on the course Web site. Sets of computer labs will also be available for teaching longitudinal data analysis using SAS.

Mathematical Background

I have kept the mathematical level of the text as low as I could. Students really should be comfortable with the vector form of linear regression $Y = X\alpha + \delta$ where X is a matrix of known covariates with n rows and K columns, α is a K -vector of coefficients, and Y and δ are n -vectors of observations and residual errors, respectively. I use α rather than the more common β for the regression coefficients. Linear algebra beyond $X\alpha$ is rarely required, and those spots can be readily skipped. In chapters 5 and 6, I write down some likelihoods and the weighted least squares estimator for the regression coefficients in longitudinal data. This requires a few matrix inverses. This material is partly included to assuage my guilt had it been omitted and to provide hooks into future mathematical material should the reader cover more advanced material elsewhere. But this material is not central to the main theme. If the students do not swallow that material whole, it should not impede understanding elsewhere. I do review linear regression briefly, to remind the reader of what they learned before; one can't learn regression fresh from the review, but hopefully it will serve to exercise any neurons that need strengthening.

Multivariate Data and Multivariate Data Courses

Because longitudinal data is multivariate, you will learn something about multivariate data when you read this book. Longitudinal data is not the only type of multivariate data, although it is perhaps the most common type of multivariate data. One of the (dirty little?) secrets of statistical research in classical multivariate data methods is that many methods, while purporting to be multivariate, are actually illustrated on, and mainly useful for, longitudinal data.

Many statistics and biostatistics departments have courses in multivariate data analysis aimed at master's-level students and quantitative

graduate students from other departments. These courses cover multivariate analysis of variance (MANOVA) and multivariate regression, among other things. I strongly recommend replacing such a course with a course in longitudinal data analysis using this book. The value of longitudinal data analysis to the student will be much greater than the value of MANOVA or multivariate regression. I often think of this course as a “money course.” Take this course, earn a living. I hire many students to analyze data on different projects; it used to be that I required familiarity with regression analysis. Now familiarity with longitudinal data analysis is the most usual prerequisite.

Target Audience

Graduating master’s students in statistics and biostatistics are very likely to be analyzing longitudinal data at least some of the time, particularly if they go into academia, the biotech/pharmaceutical industry, or other research environment. Doctoral students and researchers in many disciplines routinely collect longitudinal data. All of these people need to know about analyzing longitudinal data.

This book is aimed at master’s and doctoral students in statistics and biostatistics and quantitative doctoral students from disciplines such as psychology, education, economics, sociology, business, epidemiology, sociology and engineering among many other disciplines. These are two different audiences. The common background must be a good course in linear regression. A course at the level of Kutner, Nachtsheim, and Neter (2004), Fox (1997), Weisberg (2004) or Cook and Weisberg (1999) is a necessary prerequisite to reading this book. The seasoning provided by an additional statistics or biostatistics course at this level will be exceedingly helpful. I have taught this material to students from other disciplines whose mathematical background was not up to this level. They found this course rewarding but challenging.

The statistics and biostatistics students bring a deeper knowledge of mathematics and statistics to the course, but often little knowledge of longitudinal data other than perhaps knowledge that longitudinal data is likely to be in their future or on their comprehensive exam. Students from outside stat/biostat tend to have much less mathematical and statistical background. Instead, they bring with them the motivation that comes from having data in hand and needing to analyze it, often for their dissertation. The two different backgrounds can both lead to success in learning this material.

Applied researchers with a good regression course under their belt and some added statistical sophistication should be able to read this book as well. For anyone reading this book, the single best supplemental activity when reading the text would be to have your own data set and to draw all

the relevant plots and fit all the relevant models you read about to your own data.

An Overview

This overview is for anyone; but I'm writing it as if I were talking to another teacher.

Chapter 1, *Introduction*, introduces longitudinal data, gives examples, talks about time, discusses how longitudinal data is different from linear regression data, why analyzing longitudinal data is more difficult than analyzing linear regression data and defines notation.

Chapter 2, *Plots*, discusses the plotting of longitudinal data. Intertwined with the plots are ways of thinking about longitudinal data, issues that are naturally part of longitudinal data analysis. Even if you do not wish to cover every last piece of this material in a course, I recommend that the students read the whole chapter.

Chapter 3, *Simple Analyses*, discusses things like paired t -tests and two-sample t -tests and the two-sample t -test on paired differences, called the difference of differences, (DoD) design. These simple analyses are done on various subsets of the data or on summaries of the data. The ideas are re-used in the chapter on specifying covariates. Chapter 4, *Critiques of Simple Analyses*, complains about these analyses and explains some of the problems. Perhaps the real cost of simple analyses is the loss of the richness of multivariate data.

Chapter 5, the *Multivariate Normal Linear Model*, starts with the iid multivariate normal model for data, then introduces parameterized covariance matrices and covariates and the basic aspects of and techniques for drawing conclusions.

Chapter 6, *Tools and Concepts*, contains a grab-bag of useful tools (likelihood ratio tests, model selection, maximum likelihood and restricted maximum likelihood, back-transforming a transformed response, an introduction to design) and discussions about issues with longitudinal data analysis (assuming normality, computation). These tools may be skipped at first reading. However, my suspicion is that those readers who only read a section or two out of the entire book are most likely to dip into this chapter or into one of the topics chapters at the end. Many readers will come back to the various sections of chapter 6 when needed or interested. Most readers will continue on to chapters 7 and 8, coming back to pick up material on model selection, computation, inference as needed.

Chapters 7 and 8, *Specifying Covariates* and *Modeling the Covariance Matrix*, respectively, are the chapters that allow the flavor and beauty of longitudinal data analysis to come to full bloom. As best as possible, I have tried to write these chapters so they could be read in either order. I have tried both orders; my preference is to study covariates first. Covariate specification in longitudinal data analysis requires additional modeling

skills beyond what is taught in linear regression and is where the science usually comes in when analyzing longitudinal data. I prefer to have that as early as possible so students can start thinking about their own longitudinal data problems and how to specify the scientific questions. Another reason is that otherwise we are well past the mid-quarter mark before having talked about covariates and that is too long in the quarter to put off talking about covariates. Because there are many short references to covariance matrix specification in chapters 5 and 7, it allows for a softer introduction to the material on covariance models. The downside of this order is that students tend to ask a lot of questions about covariance models before you are ready to discuss them.

Chapter 9, *Random Effects Models*, discusses the random effects model as a hierarchical model, with discussions of random effects estimation and shrinkage. Longitudinal data sets frequently have subjects nested inside larger groups, for example students in classrooms or children in families. We explain how to model this data as well.

Chapter 10, *Residuals and Case Diagnostics*, presents current knowledge about residuals and case diagnostics with emphasis on residuals in random effects models as more is known (by me at any rate) about residuals there than in the general multivariate linear regression model.

Chapter 11, *Discrete Longitudinal Data* introduces discrete longitudinal data models. I discuss the random intercept model for binary data and for count data.

Chapter 12, *Missing Data*, is an introduction to issues surrounding missing data in longitudinal data. We talk about intermittently observed data and dropout and missing at random and variants.

Finally, chapter 13, *Analyzing Two Longitudinal Variables*, introduces bivariate longitudinal data, when you measure two variables repeatedly over time on subjects and wish to understand the interrelationship of the two variables over time.

Teaching from This Book

I teach this book as a quarter course, covering essentially the entire text. Lectures are supplemented with a computer lab that covers the use of a computer program for analyzing longitudinal data.

I have also taught precursors of this material as a subset of a quarter course on multivariate analysis for biostatistics doctoral students. In this course, I cover material from chapters 1, 2, 7, 8, and 9 in three to four weeks, concentrating on the mathematical presentation. I replace chapter 5 with a substantially higher level of mathematical rigor. Chapters 1 and 2 are shortened and the material tightly compacted. Next time I teach that course, I plan to require that students read the entire book and may add parts from chapter 11 and 13 to lectures as well.

A number of homework problems are included. That is how you can tell this is a textbook and not a monograph. The most important homework problems should lead students through a complete analysis of a simple data set. I use the Dental data for this first set of homework problems, which is why it does not appear in the text. Students should first plot and summarize the data, then explore the fixed effects, model the covariance matrix, look at the residuals and finally put their results all together in a report. This can be over a set of three homework assignments. The next assignment(s) can either be a report on the complete analysis of a somewhat more complicated longitudinal data set or another three homework assignments analyzing a data set with unbalanced or random times and more covariates. The last project should be the analysis of a still more complex data set supplied by the teacher or a data set supplied by the student. I do not give exams when I teach this material as a stand-alone course. Report writing supplies a useful form of training that was often historically lacking in statistical training. Ironically, the initial motivation for chapter 7 came from observing the difficulty that many very smart biostatistics doctoral students had in setting up even simple covariate matrices for longitudinal data during comprehensive exams. The Web site has homework assignments that I have used.

Feedback

Comments are actively solicited; especially comments that will help me make the reading and learning experience more helpful for future readers.

Acknowledgments

Many people have provided assistance in the writing of this book, in ways large and small. A number of colleagues have helped indirectly by talking with me about longitudinal data and directly with information for and comments on various drafts of the book. My apologies for omitting way too many of them from these acknowledgments. My thanks to my colleagues at UCLA both inside the Department of Biostatistics and outside for putting up with, encouraging, ignoring, and abetting. Particular thanks to Robert Elashoff, Bill Cumberland, and Abdelmonem Afifi for early encouragement.

Students in the courses Biostat 236 and Biostat 251 at UCLA have sat through many presentations of this material over a number of years and have contributed much through their questions, enthusiasms, homework answers, report writing, yawns, laughter, and typo reports. I'd like to thank them for being willing participants in reading early versions as they were written and in particular for letting me read the notes to them and for helping me catch the typos on the fly. A number of students have written master's papers and doctoral dissertations with me on longitudinal data analysis; every one has helped me understand the subject better.

I have had tons of assistance in data management and in writing code using SAS, R[®]/Splus[®], and ARC[®]/xlpstat[®]. Thanks to Charlie Zhang, Zhishen Ye, Yunda Huang, Susan Alber, Leanne Streja, Lijung Liang, Luohua Jiang, Jim Sayre, John Boscardin, Scott Comulada, Zhen Qian, Wenhua Hu, and others who I have unfortunately omitted.

I'd like to thank Sandy Weisberg for L^AT_EX[®] help and for always answering my questions about almost anything; John Boscardin for programming, L^AT_EX, and longitudinal help; Marc Suchard for detailed comments, and many many discussions, breakfasts, bagels, and pushes; Steve West; Eric Bradlow; Bill Rosenberger; Billy Crystal in *Throw Momma from the Train*: "A writer writes: always!"; Lynn Eberly for particularly helpful comments and for encouragement; several anonymous reviewers for comments both general and detailed; Susan Alber for comments on the writing, help with SAS, and teaching me about longitudinal data analysis. A big thanks to John Kimmel for his patience, encouragement, stewardship, and for finding the right answers when it mattered.

I have gotten data sets from a number of places. I'd like to thank Dr. Lonnie Zeltzer for the Pediatric Pain and Vagal Tone data; Dr. Mary Jane Rotheram for the BSI and other data sets; Dr. Charlotte Neumann for the Kenya data; Robert Elashoff for the Weight Loss data.

Finally, I would like to thank my multi-generational family for putting up with me while I worked on this. You have often asked when this book would be done. If you are reading this for the first time, check your watch and you will have your answer.

Robert Weiss
Los Angeles
2005



<http://www.springer.com/978-0-387-40271-0>

Modeling Longitudinal Data

Weiss, R.E.

2005, XXII, 432 p., Hardcover

ISBN: 978-0-387-40271-0