NSF-CBMS Regional Conference Series
in Probability and Statistics
Volume 8

ANALYSIS OF LONGITUDINAL AND CLUSTER-CORRELATED DATA

Nan Laird
Harvard University

Institute of Mathematical Statistics
Beachwood, Ohio

American Statistical Association
Alexandria, Virginia

Conference Board of the Mathematical Sciences

Regional Conference Series
in Probability and Statistics

Supported by the
National Science Foundation

The production of the *NSF-CBMS Regional Conference Series in Probability and Statistics* is managed by the Institute of Mathematical Statistics: Patrick Kelly, IMS Production Editor; Julia Norton, IMS Treasurer; and Elyse Gustafson, IMS Executive Director.

Library of Congress Control Number: 2004102584

International Standard Book Number 0-940600-60-9

Printed in the United States of America

Contents

Preface i

Acknowledgments ii

1 Longitudinal Data Analysis 1

1.1 A Linear Model for Correlated Data 6

1.2 Examples and Special Cases 13

1.3 Models for the Variance/Covariance Matrix 19

1.4 Cross-sectional versus Longitudinal Effects 21

1.5 Missing Data Issues 27

2 Estimation in the LMCD Assuming Normally Distributed Errors and Complete Data 39

2.1 ML Estimation of β and Σ 39

2.2 Properties 41

2.3 (REML) Restricted Maximum Likelihood Estimation . . . 42

2.4 REML Estimation: A Bayes Approach 47

2.5 Patterned Σ 48

2.6 Closed Form Solutions for $\widehat{\beta}_{\text{ML}}$, $\widehat{\Sigma}_{\text{ML}}$ and $\widehat{\Sigma}_{\text{REML}}$. 48

3 Estimation in the LMCD Assuming Normally Distributed Errors with Unbalanced Designs/Missing Data 53

3.1 ML and REML Estimation for the Unequal n_i Case . . . 54

3.2 A General Formulation for Incomplete Data 57

3.3 Derivatives of the Log-likelihood for the Incomplete Data Model and the EM Algorithm 59

4 Semi-Parametric Estimation in the Linear Model for Correlated Data 65

4.1 Weighted Least Squares Estimators of β 66

4.2 Properties of the Weighted Least Squares Estimator . . . 67

4.3 Weighted Least Squares with Data-Dependent Weight Functions 68

4.4 Estimation of the Optimal Weight Function 70

4.5 Locally Optimal Weighted Least Squares 72

4.6 Model Based and Robust Variance Estimation 72

4.7 Joint Estimation of β and θ 74

4.8 Efficiency of OLS Estimators 75

4.9 Remarks 77

4.10 Studies with Clusters of Random Size or Missing Data . . 78

5 Random Effects and the Linear Mixed Model 81

5.1 Two-Stage Random Effects Models 82

5.2 A Linear Mixed Model (LMM) 87

5.3 ML Estimation for the LMM 89

5.4 REML Estimation in the LMM 92

5.5 Estimating the Random Effects 94

6 Longitudinal Data Analysis for Counts and Binary Outcomes: Generalized Estimating Equations (GEE) 99

6.1 The Generalized Linear Model (GLM) for Univariate Outcomes. 100

6.2 Generalized Linear Models for Longitudinal Data 104

6.3 Estimation via GEE. 106

6.4 Estimating the Correlation Matrix. 108

7 Likelihood Models for Repeated Binary Data 113

7.1 A brief overview of log-linear models 115

7.2 The Multivariate Logistic Transform (MLT) 118

7.3 A Mixed Parameter Transform 120

8 Random Effects Models for Repeated Binary Data 125

8.1 GEE approach to estimating β 127

8.2 Likelihood Approaches 129

8.3 An Approximate Likelihood Approach: PQL 130

9 Nonignorable Nonresponse 133

9.1 Terminology 135

9.2 Methodology: General Comments 136

9.3 Examples 140

Preface

The analysis of data with outcomes measured repeatedly on each subject has experienced several transforming developments in the last twenty years. This monograph presents a unified treatment of modern methods for longitudinal and/or correlated data that have developed during this period. The basic approach that we take to modeling longitudinal data is to extend familiar univariate regression models to multivariate or correlated outcomes. We deal with linear models for measured data and generalized linear models for binary and count data. We show how methods can accommodate missing outcomes and/or unbalanced designs. Both likelihood and moment methods of estimation are covered, as are random effects approaches to data modeling and parameter estimation.

The monograph assumes that the reader has a solid foundation in statistical inference, linear and generalized linear regression models, and a basic knowledge of multivariate methods. It is appropriate for second year doctoral students or postdoctoral fellows in Statistics/Biostatistics as well as researchers or faculty interested in learning about the field.

Acknowledgments

This monograph grew out of notes written for a course on longitudinal/multivariate data analysis taught by myself and others at the Harvard School of Public Health. I am grateful to many collegues and former students for their comments and contributions to the monograph. I am especially indebted to Andrea Rotnitzky, who wrote drafts of several sections of Chapters 1, 3 and 4. I am also grateful to Stuart Baker, Christl Donnelly, Garrett Fitzmaurice, Joe Hogan, Nick Lange, Stuart Lipsitz, Nick Horton and Jim Ware for collaborations over the years which have contributed greatly to my own work in this area. Roman Torgovitsky provided especially helpful comments on the monograph at a crucial point. Finally, the impetus to turn course notes into a monograph came when Paul Speckman organized a National Science Foundation workshop on longitudinal data at the University of Missouri at Columbia. I am grateful for the NSF support, and to NIGMS for supporting my research in longitudinal data during the last twenty years.