NSF-CBMS Regional Conference Series
in Probability and Statistics
Volume 7

GENERALIZED LINEAR MIXED MODELS

Charles E. McCulloch
University of California, San Francisco

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: 2002116025

International Standard Book Number 0-940600-54-4

Printed in the United States of America

Contents

Preface vii

1 Introduction 1
- 1.1 Example: Chestnut leaf blight . . . 1
  - a. A model . . . 2
    - i. Basic elements of the model . . . 2
    - ii. Using the model . . . 3
  - b. Threshold model . . . 3
  - c. Correlations in the chestnut blight example . . . 4
- 1.2 Consequences of introducing random factors . . . 4
  - a. On the mean . . . 4
  - b. On the variance-covariance structure . . . 5
  - c. On the likelihood . . . 6
- 1.3 Testing for other genetic effects . . . 7
- 1.4 Summary . . . 8

2 Linear Mixed Models (LMMs) 9
- 2.1 Introduction . . . 9
- 2.2 Example: Propranolol and hypertension . . . 9
- 2.3 Fixed versus random factors . . . 10
- 2.4 Estimation and prediction . . . 11
  - a. A more general formulation . . . 12
  - b. Means and variances . . . 12
  - c. Estimation and prediction . . . 13
  - d. Best prediction of the random effects . . . 13
    - i. Prediction with a balanced data set . . . 14
- 2.5 The mixed model equations . . . 16
- 2.6 Testing fixed effects . . . 17
- 2.7 Testing random effects . . . 17
  - a. Numerical illustration for the Propranolol data . . . 18
- 2.8 Generalized estimating equations . . . 18
  - a. Example: milk, cows and diets . . . 18
  - b. Weighted estimation . . . 19
- 2.9 Summary . . . 20

3 Generalized linear models (GLMs) 21

3.1 Introduction 21
3.2 The modeling process 23
3.3 Transforming versus linking 24
3.4 Potato flour revisited 25

4 GLMMs **28**
4.1 Introduction 28
4.2 Basic idea 29
4.3 Example: Skin cancer 29
a. Covariances 30
b. Random slopes 30
c. Prediction 30
d. Unequal variances 31
4.4 Specifying GLMMs 31
4.5 A more general model 31
4.6 Inference in GLMMs 32
4.7 Further notes 33

5 Modeling and inference using GLMMs **34**
5.1 Introduction 34
5.2 Chestnut blight (gene effects) 34
5.3 Progabide and seizures 35
5.4 Chestnut blight (isolate effects) 36
5.5 Potomac River Fever 36
5.6 Neotropical migrants 37
5.7 Photosynthesis in corn relatives 38

6 Early research **40**
6.1 Introduction 40
6.2 Beta-binomial model 40
6.3 A Poisson–Gamma model 41
6.4 Marginal models 42
a. A model for binary data 42
6.5 Conditional inference 42
a. Matched pairs: maximum likelihood 43
b. Matched pairs: conditional likelihood 44
c. Between pairs: conditional likelihood 45
6.6 Summary 46
6.7 Further notes 47

7 Maximum likelihood for GLMMs **48**
7.1 Introduction 48
7.2 A simple example 48
a. Numerical evaluation of the likelihood 49
7.3 Simulation approaches to ML 50
a. Model and notation 50
b. Monte Carlo EM 51
c. Monte Carlo Newton–Raphson 52

d. Simulated maximum likelihood and moments 54
e. Stochastic approximation 54
7.4 Nonparametric maximum likelihood 55
7.5 Bayesian methods 55
7.6 Further notes 56

8 Other estimation methods **57**
8.1 Introduction 57
8.2 Generalized estimating equations 57
a. Non-normal data 59
b. Comparison of marginal versus conditional modeling 60
c. Comparison of GEE and random effects estimation methods . 61
8.3 Dispersion–mean model 62
8.4 Penalized quasi-likelihood 62
a. Higher-order Laplace approximations 66
8.5 Choosing good estimating equations 66
a. Illustration: REML 67
b. BLUP methods 68
c. Composite and working likelihoods 69
8.6 Progabide and seizures revisited 69
8.7 Further notes 70

9 Conclusions **71**
9.1 Introduction 71
9.2 Chestnut leaf blight revisited 71
9.3 Further work 72
a. Random effects distribution 72
b. Small sample distributions 73
c. Prediction error 73
9.4 Summary 73

Bibliography **74**

Index **83**

Preface

When Jim Booth from the University of Florida first approached me about being nominated as the primary lecturer for a Conference Board in the Mathematical Sciences (CBMS) summer conference I was both flattered and slightly apprehensive. When Jim called me to tell me the conference had been approved he only slightly jokingly asked if it was good news or bad. He had to set about organizing a large workshop and I had to prepare ten lectures on the state of generalized linear mixed models (GLMMs). While travelling to the conference and looking at the participant list, I realized there was likely no topic I was lecturing on for which I was the most expert participant. Fortunately, the conference participants were tolerant and engaging and Jim and I both agreed afterwards that it was a wonderful experience and well worthwhile. To a large extent the thanks go to Jim Booth, Jim Hobert, the other speakers (Xihong Lin, David Ruppert, Pat Heagerty and Jim Hobert) and participants.

I apologize for taking so long to get this monograph produced. Shortly after the conference I accepted a position at the University of California, San Francisco. Moving and adjusting to a new position simply did not allow me the time to work on the monograph.

This monograph is a fairly accurate account of the lectures I gave at the CBMS conference with a bit of updating, especially in the bibliographic notes sections that follow each of the later chapters. Compared to my other books, *Variance Components* (with Shayle Searle and George Casella) and *Generalized, Linear, and Mixed Models* (with Shayle Searle) this monograph has much more of a research focus and I have taken the liberty of covering in more depth topics I feel personally passionate about and less obligated to provide textbook-like coverage. Though the title is very similar to *Generalized, Linear, and Mixed Models* the focus is quite different: the lack of commas and the missing "and" are statistically significant!

This monograph is written assuming familiarity with linear models and matrix algebra and some exposure to mixed models and logistic regression. References are given to more standard texts that cover some of the basic material in more depth. The monograph begins with an extended example that introduces all the main ideas. Chapters 2 and 3 briefly review linear mixed and generalized linear models and Chapter 4 defines and introduces GLMMs. Chapter 5 illustrates the breadth of inferential goals possible with GLMMs. One of my main attractions in conducting research on this class of models was the wide variety of practical applications. Chapters 6 through 9 contain the "meat" and tackle the difficult aspects fitting these models to data. The monograph is organized along the lines of

the CBMS lectures.

Work on this monograph was partially supported by NSF Grants DMS-96-25476 and 01-03792.