
Contents

List of Examples	xv
List of Figures	xvii
List of Tables	xxvii
Preface	xxxi
Acknowledgments	xxxvii
I Mixed Linear Models: Syntax, Theory, and Methods	1
1 An Opinionated Survey of Methods for Mixed Linear Models	5
1.1 Mixed Linear Models in the Standard Formulation	5
1.2 Conventional Analysis of the Mixed Linear Model	16
1.2.1 Overview	16
1.2.2 Mean Structure Estimates	17
1.2.3 Estimating ϕ , the Unknowns in G and R	19
1.2.3.1 Maximizing the, or rather a, likelihood	19
1.2.3.2 Maximizing the Restricted (Residual) Likelihood	20
1.2.4 Other Machinery of Conventional Statistical Inference	21
1.2.4.1 Standard Errors for Fixed and Random Effects	22
1.2.4.2 Testing and Intervals for Fixed and Random Effects	23
1.2.4.3 Testing and Intervals for ϕ	24
1.3 Bayesian Analysis of the Mixed Linear Model	26
1.3.1 A Very Brief Review of Bayesian Analysis in General	26
1.3.2 Bayesian Analysis of Mixed Linear Models	29
1.3.2.1 Tests and Intervals for Unknowns	29
1.3.2.2 Some Comments on the Bayesian Analysis	31
1.3.2.3 Prior Distributions for Variances	33
1.3.2.4 Prior Distributions for Covariance Matrices	35
1.3.3 Computing for Bayesian Analyses	37
1.4 Conventional and Bayesian Approaches Compared	40
1.4.1 Advantages and Disadvantages of the Two Approaches	40

1.4.2	Conventional and Bayesian Analyses for the Viral-Structure Example	43
1.5	A Few Words about Computing	46
	Exercises	48
2	Two More Tools: Alternative Formulation, Measures of Complexity	51
2.1	Alternative Formulation: The “Constraint-Case” Formulation	51
2.1.1	The Constraint-Case Formulation	51
2.1.2	A Brief History of the Constraint-Case Formulation	55
2.2	Measuring the Complexity of a Mixed Linear Model Fit	55
2.2.1	DF in the Whole Fit	56
2.2.2	Partitioning a Fit’s DF into Components	58
2.2.2.1	Notation, Motivation for the Partition of DF	58
2.2.2.2	Partitioning DF	59
2.2.2.3	Properties of the Definition	59
2.2.2.4	Example of Partitioning DF	60
2.2.3	Uses of DF for Mixed Linear Models	63
2.2.3.1	Using DF to Specify F -Tests	63
2.2.3.2	Using DF to Describe Model Size, for Model-Selection Criteria	65
2.2.3.3	Prior Distributions on DF	65
2.2.4	DF Compared to p_D of Spiegelhalter et al. (2002)	68
2.2.5	Some Intuition for Fractional Degrees of Freedom	70
2.2.5.1	Using the Constraint-Case Formulation	70
2.2.5.2	Using the Mixed Model Formulation	72
	Exercises	73
II	Richly Parameterized Models as Mixed Linear Models	75
3	Penalized Splines as Mixed Linear Models	79
3.1	Penalized Splines: Basis, Knots, and Penalty	79
3.2	More on Basis, Knots, and Penalty	83
3.2.1	A Few More Bases	83
3.2.2	A Bit More on Penalty Functions	85
3.2.3	Brief Comments on Some Operational Matters	87
3.3	Mixed Linear Model Representation	88
3.3.1	Applying the Mixed-Linear-Model Approach	90
3.3.2	Brief Comments on Other Aspects of the Mixed-Linear-Model Analysis	98
	Exercises	99

Contents	ix
4 Additive Models and Models with Interactions	101
4.1 Additive Models as Mixed Linear Models	102
4.1.1 The Pig Jawbone Example	102
4.1.2 Additive Models Defined	103
4.1.3 Additive Models Fit to the Pig Jawbone Data	104
4.2 Models with Interactions	109
4.2.1 Categorical-by-Continuous Interactions	110
4.2.2 Categorical-by-Categorical Interactions (Smoothed ANOVA)	114
4.2.3 Smoothed ANOVA, Balanced Design with a Single Error Term	115
4.2.3.1 Notation and Other Machinery	116
4.2.3.2 DF in Effects, Prior Distributions, SANOVA Table	119
4.2.3.3 SANOVA Applied to the Denture-Liner Example	121
4.2.4 Smoothed ANOVA for More General Designs	126
Exercises	126
5 Spatial Models as Mixed Linear Models	129
5.1 Geostatistical Models	130
5.2 Models for Areal Data	132
5.2.1 Common Areal Models: SAR, CAR, and ICAR	133
5.2.2 More on the ICAR Model/Prior	136
5.2.3 Smoothed ANOVA with Spatial Smoothing	141
5.3 Two-Dimensional Penalized Splines	144
5.3.1 Tensor-Product Basis	144
5.3.2 Radial Basis	145
5.3.3 A Comment on Tensor-Product vs. Radial Basis	149
Exercises	149
6 Time-Series Models as Mixed Linear Models	151
6.1 Example: Linear Growth Model	151
6.2 Dynamic Linear Models in Some Generality	154
6.3 Example of a Multi-component DLM	156
Exercises	160
7 Two Other Syntaxes for Richly Parameterized Models	163
7.1 Schematic Comparison of the Syntaxes	163
7.2 Gaussian Markov Random Fields (Rue & Held 2005)	164
7.3 Likelihood Inference for Models with Unobservables (Lee et al. 2006)	168
Exercises	171

III From Linear Models to Richly Parameterized Models: Mean Structure	173
8 Adapting Diagnostics from Linear Models	177
8.1 Preliminaries	178
8.2 Added-Variable Plots	181
8.3 Transforming Variables	183
8.4 Case Influence	184
8.5 Residuals	192
Exercises	197
9 Puzzles from Analyzing Real Datasets	201
9.1 Four Puzzles	201
9.1.1 Introducing Spatially Correlated Errors Makes a Fixed Effect Disappear	201
9.1.2 Adding a Clustering Effect Changes One Fixed Effect but Not Another	203
9.1.3 Differential Shrinkage of Effects with Roughly Equal Estimates and Standard Errors	205
9.1.4 Two Random Effects Obliterated by Adding an Apparently Unrelated Random Effect	206
9.2 Overview of the Next Three Chapters	210
Exercises	210
10 A Random Effect Competing with a Fixed Effect	211
10.1 Slovenia Data: Spatial Confounding	211
10.1.1 The Mechanics of Spatial Confounding	211
10.1.1.1 The Model, Re-written as a Mixed Linear Model	211
10.1.1.2 Spatial Confounding Explained in Linear-Model Terms	213
10.1.1.3 Spatial Confounding Explained in a More Spatial- Statistics Style	214
10.1.2 Avoiding Spatial Confounding: Restricted Spatial Regres- sion	215
10.1.3 Spatial Confounding Is Not an Artifact of the ICAR Model	216
10.1.4 Five Interpretations, with Implications for Practice	217
10.1.4.1 S Is a New-Style Random Effect, a Formal Device to Implement a Smoother	218
10.1.4.2 S Is an Old-Style Random Effect	221
10.1.5 Concluding Thoughts on Spatial Confounding	223
10.2 Kids and Crowns: Informative Cluster Size	224
10.2.1 Mechanics of Informative Cluster Size	224
10.2.2 Illustrative Examples	230
10.2.3 Some Simple Analyses of the Kids-and-Crowns Data	233
Exercises	235

Contents	xi
11 Differential Shrinkage	237
11.1 The Simplified Model and an Overview of the Results	237
11.1.1 The Simplified Model	238
11.1.2 Overview of the Results	239
11.2 Details of Derivations	240
11.2.1 Centered Predictors and Outcome and No Intercept	240
11.2.1.1 Expressions for the Inferential Summaries	240
11.2.1.2 Proofs of the Claims	242
11.2.2 Centered Predictors with an Intercept	245
11.2.3 Predictors Not Centered	246
11.3 Conclusion: What Might Cause Differential Shrinkage?	247
Exercises	248
12 Competition between Random Effects	251
12.1 Collinearity between Random Effects in Three Simpler Models	252
12.1.1 Model with Clustering and Heterogeneity	252
12.1.1.1 Theory	252
12.1.1.2 Illustrations Using Simulated Data	255
12.1.2 Two Crossed Random Effects	262
12.1.2.1 Theory	262
12.1.2.2 Illustrations Using Simulated Data	264
12.1.3 Three Predictors, Two with Shrunk Coefficients	267
12.2 Testing Hypotheses on the Optical-Imaging Data and DLM Models	270
12.2.1 The Hypothesis about Collinearity of the Design Matrices	271
12.2.2 The Hypothesis about Lack of Fit	275
12.3 Discussion	282
Exercises	282
13 Random Effects Old and New	285
13.1 Old-Style Random Effects	285
13.2 New-Style Random Effects	286
13.2.1 There Is No Population	287
13.2.2 The Effect's Levels Are the Whole Population	290
13.2.3 A Sample Has Been Drawn, but . . .	292
13.2.4 Comments on New-Style Random Effects	293
13.3 Practical Consequences	294
13.3.1 Inference and Prediction	294
13.3.1.1 Inference	294
13.3.1.2 Prediction	297
13.3.2 Interpretation of Analytical Artifacts	298
13.3.3 Simulation Experiments to Evaluate Statistical Methods	298
13.4 Conclusion	302
Exercises	302

IV Beyond Linear Models: Variance Structure	303
14 Mysterious, Inconvenient, or Wrong Results from Real Datasets	307
14.1 Periodontal Data and the ICAR Model	307
14.2 Periodontal Data and the ICAR with Two Classes of Neighbor Pairs	309
14.3 Two Very Different Smooths of the Same Data	312
14.4 Misleading Zero Variance Estimates	312
14.5 Multiple Maxima in Posteriors and Restricted Likelihoods	315
14.6 Overview of the Remaining Chapters	316
Exercises	318
15 Re-expressing the Restricted Likelihood: Two-Variance Models	319
15.1 The Re-expression	319
15.2 Examples	323
15.2.1 Balanced One-Way Random Effect Model	324
15.2.2 Penalized Spline	324
15.2.3 ICAR Model; Spatial Confounding Revisited	329
15.2.3.1 Attachment-Loss Data: Simple ICAR Model	329
15.2.3.2 Spatial Confounding: ICAR with Predictors	334
15.2.4 Dynamic Linear Model with One Quasi-Cyclic Component	338
15.3 A Tentative Collection of Tools	343
Exercises	345
16 Exploring the Restricted Likelihood for Two-Variance Models	347
16.1 Which \hat{v}_j Tell Us about Which Variance?	347
16.1.1 Some Heuristics	347
16.1.2 Case-Deletion Using the Gamma GLM Interpretation	348
16.1.3 Modified Restricted Likelihood	354
16.1.4 Summary and an Important Corollary	358
16.2 Two Mysteries Explained	358
Exercises	365
17 Extending the Re-expressed Restricted Likelihood	367
17.1 Restricted Likelihoods That Can and Can't Be Re-expressed	367
17.1.1 Two Restricted Likelihoods That Can't Be Re-expressed	368
17.1.2 Balanced Designs	370
17.1.2.1 Derivations Regarding Balance	373
17.1.3 Gaussian Processes Using the Spectral Approximation	376
17.1.4 Separable Models	380
17.1.5 Miscellaneous Other Models	381
17.2 Expedients for Restricted Likelihoods That Can't Be Re-expressed	382
17.2.1 Expedient 1: Ignore the Error Variance	382
17.2.2 Expedient 2: Ignore the Non-zero Off-Diagonals	385
Exercises	395

Contents	xiii
18 Zero Variance Estimates	397
18.1 Some Observations about Zero Variance Estimates	397
18.1.1 Balanced One-Way Random Effects Model	397
18.1.2 Balanced ANOVA for the Nerve-Density Example	398
18.2 Some Thoughts about Tools	401
Exercises	403
19 Multiple Maxima in the Restricted Likelihood and Posterior	405
19.1 Restricted Likelihoods with Multiple Local Maxima	405
19.2 Posteriors with Multiple Modes	406
19.2.1 Balanced One-Way Random Effect Model	407
19.2.2 Two-Level Model: The HMO Data (Example 9) Revisited	409
Exercises	411
References	413
Author Index	425
Subject Index	429

Preface

If you believe in things that you don't understand, then you suffer (Wonder 1973).

When I was in graduate school in the early 1980s, linear model theory had just been perfected and I studied with some of the people who had perfected it. With its combination of simplicity and near-total explanatory power, linear model theory is like nothing else in statistics and by the early 1980s, this theory had been honed to the point where it could be taught almost whole to people taking their first serious regression course (e.g., Weisberg 1980).

At about the same time, however, the main thrust of statistical research turned in a different direction, emphasizing breadth over depth by producing methods for specifying and fitting models of greater generality and with weaker assumptions: generalized linear models, which dropped normality; additive models, which dropped linearity; generalized estimating equations (GEE), which dropped independence; hierarchical models, which added layers of structure; mixed models, which added random effects; Markov chain Monte Carlo, with the great flowering of Bayesian methods it enabled; the modeling syntax of the S and then R systems; structural equations models; spatial models and smoothers; dynamic linear models (state-space models); and penalized fits, among many others. Any statistician with a pulse has to love and be impressed by this explosion of sheer modeling power. I do, and I am. If you read something in this book that seems to contradict that statement, go back and read it again.

My admiration notwithstanding, I am mostly an applied statistician and when I use these new methods to analyze my collaborators' data, I routinely get results that are mysterious, inconvenient, or plainly wrong. For mixed linear models, these unhappy results include zero variance estimates, multiple maxima, counterintuitive outlier effects, odd fits (e.g., a wiggly smooth with one smoother but not with another apparently similar smoother), big changes in fit from apparently modest changes in the model or data, and unpredictable convergence of numerical routines, among other things. When my collaborators' datasets produce such puzzles, I urgently need something like the powerful theory of linear models so I can explain them and figure out what to do. As far as I can tell, however, little if anything is known about most of these puzzles. I see very few mentions of them in the listservs I peruse, in the statistical literature, or in talks. There's no reason to think I am a magnet for freak problems, so I suspect that many, perhaps most, working statisticians encounter the same puzzles. When I publish a paper or do a talk about some of these things, I usually hear from people who have had the same problem and are relieved that it

wasn't just a programming error, as a referee or their thesis advisor had insisted. For some reason, however, we don't talk or write much about these puzzles.

It seems as if research in statistics has come to mean promoting new methods, as opposed to understanding methods, old or new. Obviously, it would be inaccurate and unfair to say that *nobody* tries to understand existing or new methods; counterexamples include the trace plots used to describe LASSO results or the separation plots used to explain the support-vector machine. But when a classmate and I tried to assemble a catalog of the mysteries and puzzles we've found as a rationale for more investment in understanding these new methods, we found it impossible to formulate anything that looked like a contemporary journal article. These days, a journal article needs to end in a triumph — “Behold! We have conquered this messy dataset/new class of models/previously intractable computing problem/[etc.]!” — and it is hard to make a mystery or puzzle look like a triumph. The triumphant narrative style is so embedded in today's conception of a journal article that even not-too-thoughtful extensions of linear model theory are cast as new methods. If you doubt this, try doing a literature search for outlier-detection methods for hierarchical or other random-effect models. With two exceptions that I know of (one that I wrote and one that I recently refereed), every such paper ends with a standard triumph in which the new method is shown to identify outliers, but none of these papers is based on a systematic understanding of models with random effects.

It's not hard to imagine why the literature looks this way. There are so many new models, where do you begin? With the present-day emphasis on generality, how can you do anything general enough to interest a good journal? The new methods are so complex! And so on. The solution to this quandary, it seems to me, is to stop trying to learn something about all or even a large fraction of the wonderful new methods developed in the last 30 years. They are too disparate and they are developing too quickly.

Having made that unsexy concession, there does seem to be a good place to start. Many of the new methods are now undergoing a process of unification analogous to the unifications that produced generalized linear models and, even earlier, the projection theory of linear models. The unification of models with random effects — so far — consists of a few competing syntaxes for expressing a large class of models and a method for fitting models expressed in each syntax. Parts I and II of this book emphasize one such syntax, mixed linear models using normal distributions, and some of the great variety of models, which I call richly parameterized models, that can be expressed this way and analyzed using conventional and Bayesian methods for mixed linear models. Examples of this unification include Robinson (1991) and Ruppert et al. (2003). This class of models is rich enough to be interesting and close enough to single-error-term linear models to allow many insights and methods to be borrowed or adapted.

A theory of richly parameterized linear models needs more than a syntax and a computing method. It needs to explain things that happen when these models are used to analyze data, to provide ways to detect problems, and when possible, to show how to mitigate or avoid those problems. Part III takes a step from the theory of ordinary linear models toward a theory of richly parameterized models by adapting ideas

central to linear model theory. Part IV then takes a step beyond linear model theory by examining the information in the data about the mixed linear model's covariance matrices, which are the difference between ordinary and mixed linear models.

Parts III and IV are founded on two key premises. The first premise comes from linear model theory: Writing down a model and using it to analyze a dataset is equivalent to specifying a function from the data to the inferential or predictive summaries. However you rationalize or interpret that model, it is essential to understand, in a purely mechanical sense, the function from data to summaries that is implied by the model. Often when I present material from this book, people respond with things that are *non sequiturs* in these terms, for example: This model estimates a causal effect while this other model does not; or a Gaussian process probability model with such-and-such covariance function has realizations with such-and-such properties; or this posterior distribution or likelihood is by definition what the data have to say about the model's unknown parameters. I don't dispute such assertions but for the purposes of this book, they are irrelevant. The question asked here is: When I fit *this* model to *this* dataset, why do I get *this* result, and how much would the result change if I made *this* change to the data or model?

The second premise is that we must distinguish between the model *we choose* to analyze a given dataset and the process that *we imagine* produced the data or that, in rare cases, we know actually did produce the data. Our choice to use a model with random effects does not imply that those random effects correspond to any random mechanism out there in the world, and that fact has practical implications. These implications are a recurring theme of this book's first three parts and are summarized in Chapter 13.

Building on these two premises, Parts III and IV are organized around mysterious, inconvenient, or plainly wrong results that turned up in real problems. Most of these are from my collaborative work but some have come from colleagues, for example Michael Lavine's dynamic linear model puzzle (Chapters 6, 9, 12, and 17). Some of these puzzles are now understood to a greater or lesser extent, while others are barely understood at all. In that sense, Parts III and IV are not quite a catalog of unsolved problems — who knows how many puzzles are as yet unreported or undiscovered? — and their theory is grossly incomplete. I will apologize for that once, now. I hope Parts III and IV stimulate enough research so that the second edition of this book, if there is one, can report fewer mysteries and, yes, more triumphs.

Most statistics texts emphasize what we know. This book emphasizes what we don't know. Judging from reviews of my proposals, a lot of academics think mixed linear models are completely understood, when in fact they are still largely not understood. But this is good news: What a bounty of unsolved problems, and for a heavily used class of models! Graduate students, start your engines!

In emphasizing what we don't know, Parts III and IV consider open problems, which leads to a stylistic quandary. Traditionally, statistical theory follows a mathematical style emphasizing results that can be packaged as theorems. I have observed this tradition whenever possible, because you just can't do better than a relevant theorem. Unfortunately, I could not always produce theorem-like material in a reasonable amount of time. In situations like this, statistical theorists usually present

nothing at all or present something they *can* package as a theorem, most often some kind of large-sample result. That seemed unproductive, so when I haven't been able to produce relevant theorem-like results, I have instead followed a style used by my scientific (as opposed to statistical) colleagues. They work by posing hypotheses and gathering a variety of evidence, often indirect, so that their hypotheses are either refuted or accumulate credibility while becoming more refined. Obviously accumulated credibility can't replace the iron-clad certainty of a theorem when a relevant theorem can be proven, but the new methods of the last three decades are so complex that it may never be possible to prove relevant theorems about them. We can, however, make progress by approaching our black-box methods in the same way our scientific colleagues approach nature's black boxes, by prying them open gradually and indirectly if necessary. Chapters 11, 12, and 17 are examples of this style, with Chapter 11 refuting a hypothesis and Chapters 12 and 17 developing some hypotheses and producing a first increment of credibility for each.

Along with a sometimes non-mathematical style of inquiry, I've written in a relatively informal narrative style. I've done so because it's friendlier in two senses: It's easier to understand on a first reading and it doesn't hide my opinions and ignorance behind the passive voice and calculated omissions. Students and other readers *should* find fault with the current state of this field, including the things I've contributed to it, and I want them to see those faults and decide they can do better. I will be delighted if this book attracts the interest of young people with better math and computing skills than I have, who can change this area of study from the backwater it is into the thriving area it can and should be.

Some Guidance about Using This Book

The object of Part I is to present a survey of essentials and a particular point of view about them. The object of Parts II, III, and IV is to present the beginnings of a theory of richly parameterized linear models. This book is not intended to be a magisterial overview of everything known about mixed linear models. It is rather intended to present a point of view about what we do and do not understand about mixed linear models and to identify research opportunities. Overviews of mixed linear models include Searle et al. (1992); Ruppert et al. (2003), which focuses on penalized splines represented as mixed linear models; Verbeke & Molenberghs (1997), which focuses on SAS's MIXED procedure; Diggle et al. (1994), which focuses on longitudinal models; Snijders & Bosker (2012), a thorough treatment of hierarchical (multi-level) models obviously based on a lot of experience fitting them and explaining the fits; and Fahrmeir & Tutz (2001), which catalogs models with exponential-family error distributions and linearity in the mean structure.

The hazard of writing a book like this is that I have to write short chapters about sub-fields of statistics with huge literatures. Even D.R. Cox might not be able to master all those sub-fields; I certainly haven't. Academics tend to be territorial and to view the world through a microscope, so whatever I write is guaranteed to offend specialists in each sub-field even if I say nothing that is factually incorrect. Also, the literature and folklore of mixed linear models are gigantic and I know less than

I probably should about them. Therefore I ask your indulgence: I have tried to be nice to everybody and in return, if I've said something factually incorrect or wrong-headed, please tell me and provide detailed citations. If I decide you're right, I'll post a suitable piece on the book's web site with credit to you and replace the relevant passages in the next edition, if there is one.

I've tried simultaneously to give results about both Bayesian and conventional (non-Bayesian) analyses which, these days, mostly revolve around the restricted or residual likelihood. I've done this because Leo Breiman (2001) was right: The two approaches really aren't much different in practice, at least in this area. Chapter 1 is my argument for that claim. I had the good fortune to study in a department where I could become fluent in both languages but most people aren't so fortunate and thus might find it difficult to switch back and forth between Bayesian and conventional language. I've tried to make this as clear as I can and I apologize for any failures of clarity.

I wrote this book from classroom overheads for a one-semester course that I teach for advanced PhD students in the Division of Biostatistics at the University of Minnesota. I use *Semiparametric Regression* by Ruppert, Wand, & Carroll (2003) as a textbook for that course and Parts I and II of the present book refer to it frequently. *Semiparametric Regression* is simply lovely. Among statistical books with hard technical content, it is the friendliest I've ever read and I only disagree with three or four things in the whole book. I recommend it without reservation. If the present book is written half as well, I will be happy.

Each of the present book's chapters ends with exercises, which are of two types. The first type is standard results that PhD students should be able to derive, which are intended to provide practice with the mostly algebraic methods used in this book. Most chapters also include exercises that are, as far as I know, open research questions. Often these include suggestions about where to start, but of course you should feel free to ignore my suggestions.

The book's web site includes datasets analyzed as examples, when I could get permission to include them. I will be happy if you find these datasets useful and horrified if any of them ends up being pawed over eternally like the stack-loss data or the Scottish lip-cancer data. Publish your own datasets! The world will be richer for it.

When I have used one of my published analyses as an example, I have presented it the way it was published. I figured it would be both dishonest and hazardous to make myself look smarter than I actually was and hope nobody checked the original papers. This also gave me an incentive to be nicer to other researchers than I might otherwise be. In each such case I point out what I now believe is wrong with the analysis I published and when it seems worth the effort and space I give a better analysis. If you identify blunders I haven't mentioned and I agree they are blunders, I will post your attempts to alleviate my ignorance on the book's web site.

Finally, Part I refers to SAS frequently because I am in a biostatistics department and even though cognoscenti are obligated to sneer at SAS, we teach it to our students (along with R and WinBUGS) because it helps them find jobs. I don't mean to single

out SAS for criticism but it *is* the Microsoft Office of statistical software and that makes it a good example for many purposes.

May 2013