

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI[®]

**Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

**NONLINEAR MIXED EFFECTS MODELS
FOR META-ANALYSIS**

By

Nicholas J. Barrowman

**SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
AUGUST 2000**



© Copyright by Nicholas J. Barrowman, 2000



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*

Our file *Notre référence*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-57342-7

Canada

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

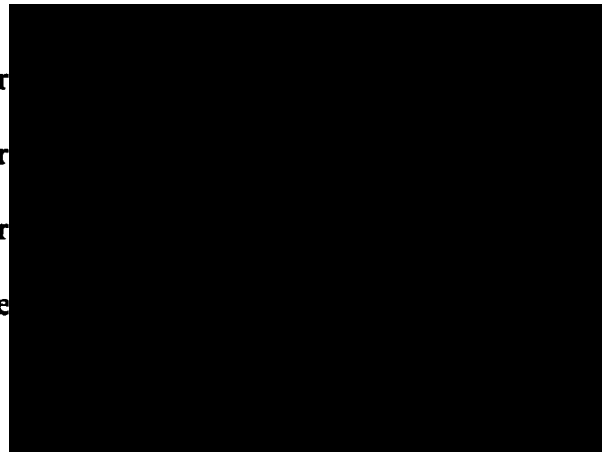
The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Nonlinear Mixed Effects Models for Meta-Analysis"

by Nicholas James Barrowman

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: August 21, 2000

External Examiner
Research Supervisor
Research Supervisor
Examining Committee



DALHOUSIE UNIVERSITY

Date: August 2000

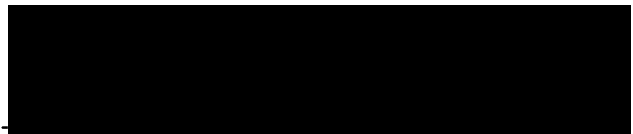
Author: **Nicholas J. Barrowman**

Title: **Nonlinear Mixed Effects Models for Meta-Analysis**

Department: **Mathematics and Statistics**

Degree: **Ph.D.** Convocation: **October** Year: **2000**

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

A large black rectangular box redacting the author's signature.

Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

To my father, Dr. James A. Barrowman

Contents

Acknowledgements	viii
Abstract	ix
List of Symbols	x
1 Introduction	1
1.1 Example 1: Wolves in Québec	3
1.2 Example 2: Ulcer studies	5
1.3 Example 3: Coho salmon spawner-recruitment data	7
1.3.1 Spawner-recruitment models	10
1.4 Conventional approaches to meta-analysis	13
1.4.1 Fixed effects meta-analysis: weighted means	14
1.4.2 Random effects meta-analysis	15
1.4.3 Empirical Bayes estimation	16
1.4.4 Mixed effects meta-analysis	17
1.5 Meta-analysis of raw data	18
1.5.1 Repeated measures data	19
1.5.2 Hierarchical models for raw-data meta-analysis	19
2 Individual estimates	20
2.1 Example: Wolves in Québec	21
2.2 Example: Ulcer studies	24
2.3 Example: Coho salmon	27
2.4 Approaches to robust estimation	35

2.4.1	Weights as diagnostics	39
2.4.2	Example: Wolves in Québec	39
2.5	Elimination of nuisance parameters	42
2.5.1	Conditional likelihood	44
2.5.2	Integrated likelihood	46
2.5.3	Profile likelihood	46
2.6	Raindrop plots	47
2.6.1	Construction of the raindrop plot	51
2.6.2	Example: Ulcer studies	52
2.6.3	Example: Coho salmon Beverton-Holt model	54
2.6.4	Example: Coho salmon hockey-stick models	58
3	Hierarchical models for meta-analysis	63
3.1	Conditionally independent hierarchical models	63
3.2	Parametric empirical Bayes inference	65
3.3	Hierarchical Bayes inference	66
3.4	Comparing empirical Bayes and hierarchical Bayes	68
3.5	Empirical Bayes methods for combining likelihoods	71
3.5.1	Example: Ulcer studies	71
3.5.2	Example: Coho salmon	74
4	Linear mixed effects models	78
4.1	General linear mixed effects model	79
4.2	Estimation	79
4.2.1	Random effects meta-analysis	81
4.2.2	Variance-covariance components estimation	84
4.3	Linear mixed effects model for repeated measures data	88
4.4	Estimation	88
4.4.1	Example: Linear mixed effects models for spawner-recruitment data	89
4.4.2	Example: Linear mixed effects analysis of the wolf data	90
4.5	Asymptotics	97
4.6	Approaches to robust estimation	99
4.6.1	Robust estimation of realized random effects	102

4.6.2	Example: Wolves	102
5	Nonlinear mixed effects models	105
5.1	Nonlinear mixed effects model for repeated measures data	106
5.1.1	Example: Beverton-Holt spawner-recruitment model	107
5.1.2	Related models	108
5.1.3	General nonlinear mixed effects model	109
5.2	Estimation	110
5.2.1	Lindstrom and Bates' algorithm	114
5.2.2	Wolfinger's derivation of an approximate REML likelihood	120
5.2.3	Another Laplacian approximation	121
5.2.4	Asymptotic results	122
5.2.5	Nuñez's approach (SPML)	122
5.2.6	A modification of SPML: Stylized normal samples	130
5.2.7	Uncertainty estimates	134
5.2.8	Example: Coho salmon Beverton-Holt mixed model	140
5.2.9	Example: Coho salmon hockey-stick mixed model	145
5.3	Approaches to robust estimation	146
5.3.1	Modifying Lindstrom and Bates' algorithm	147
5.3.2	Example: Coho salmon Beverton-Holt mixed model	149
5.3.3	Modifying SPML	150
5.3.4	Example: Coho salmon Beverton-Holt mixed model	160
5.3.5	Comparison of the two proposals for robust estimation	163
6	Conclusions	165
6.1	Summary	165
6.1.1	Graphical methods	165
6.1.2	Robust methods	166
6.1.3	Other contributions	167
6.2	Mixed effects models for spawner-recruitment data	167
6.2.1	Application to extinction models	168
6.2.2	Application to management models	168
6.3	Future research	169

Acknowledgements

During my work on this thesis, I have received assistance from many quarters. I wish to thank the Killam Foundation for its financial support. My two supervisors have been tremendous—I wish to thank Chris Field for his patience and clarity and Ram Myers for his penetrating insight, unfailing humour, and for an endless supply of coffee. Rick Routledge kindly agreed to be my external examiner and provided very helpful feedback: the thesis has been substantially improved due to his efforts. I am grateful to David Hamilton for several fruitful conversations concerning nonlinear models and the raindrop plot. Bruce Smith was also kind enough to read my thesis at short notice. Dan Kehler, Shelton Harley, Keith Bowen, and Alice Richardson kindly provided feedback on earlier drafts of the thesis. I would also like to thank Michele Millar, Joanna Mills, and Patricia Moorhead for their assistance and encouragement. My sister-in-law, Adele Megann, provided invaluable editing advice. I wish to thank my families, the Meganns and the Barrowmans, for their financial and moral support. Finally, I wish to thank Gillian, Clare, and Adam for their support, encouragement, and love.

Nonlinear Mixed Effects Models for Meta-Analysis

Nicholas J. Barrowman

Doctor of Philosophy

Department of Mathematics and Statistics

Dalhousie University

2000

Abstract

Meta-analysis is sometimes defined narrowly as the combination of summary statistics from different studies. Here we consider cases where the raw data from each study are available, and explore hierarchical modeling approaches to meta-analysis. Any statistical analysis may present hazards—such as model misspecification and overly influential observations—and meta-analysis is particularly susceptible. One approach to this problem is through the use of graphical methods for diagnosing deviations from our assumptions. In this work we develop several new displays for meta-analysis, including the raindrop plot for displaying information from many studies simultaneously. Approaches to the estimation of linear and nonlinear mixed effects model are reviewed and generalizations of a segmented regression model are developed for use in mixed model analyses. A different way of dealing with the pitfalls of meta-analysis is through the use of robust statistics. Huggins and Richardson previously developed methods for robust estimation of linear mixed effects models. Nonlinear mixed effects models present additional challenges. Based on the approach of Huggins, we propose modifications to robustify two algorithms for estimating nonlinear mixed effects models. Our methods are illustrated using data on ulcer studies, wolf populations, and coho salmon populations.

List of Symbols

Because this thesis involves several different fields of research, the notation can be formidable. So that the symbols don't obscure the meaning, I have tried to keep it simple, for example eschewing the use of boldface notation to distinguish vectors from scalars, as this should be clear from the context. The transpose of a vector (or matrix) x is denoted x^T . I use k to denote a generic index, f to denote a generic function, and I to denote the identity matrix. If f is a vector-valued function of a vector argument x , then $f'(z)$ denotes $\partial f / \partial x^T |_{x=z}$. If f is scalar-valued then $f''(z)$ denotes $\partial^2 f / \partial x \partial x^T |_{x=z}$. Capital letters generally denote matrices. The following is a partial list of the most important symbols used in the thesis, grouped by topic.

Meta-analysis

m number of studies

i index for studies

n_i number of observations in study i

$n = n_1 + \dots + n_m$

j index for observations in studies

y_{ij} j th observation in study i

y_i n_i -dimensional vector of observations in study i

y n -dimensional vector of all observations

Spawner-recruitment models

S_{ij} j th observation of spawner biomass in population i

R_{ij} j th observation of recruitment in population i

α_i slope at the origin for population i

Robustness

r_{ij} j th standardized residual in study i

r_i vector of standardized residuals in study i

$\rho(r_{ij})$ rho-function of a single standardized residual

$\rho(r_i) = \sum_{j=1}^{n_i} \rho(r_{ij})$

$\psi(r_{ij}) = \rho'(r_{ij})$

$\psi(r_i) = (\psi(r_{i1}), \dots, \psi(r_{in_i}))^T$

Mixed effects models for repeated measures data

$$\mu_i = E(y_i)$$

$$\mu = E(y)$$

$$V_i = \text{Var}(y_i)$$

$$V = \text{Var}(y)$$

q number of fixed effects

β q -dimensional vector of fixed effects

c number of random effects (including error)

u_i $(c - 1)$ -dimensional vector of random effects for study i

$$u = (u_1^T, \dots, u_m^T)^T$$

$$\xi = (\beta^T, u^T)^T$$

$$D = \text{Var}(u_i)$$

$$\bar{D} = \text{Var}(u) = \text{diag}(D, \dots, D)$$

ε_{ij} j th error for i th study

$$\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})^T$$

$$\varepsilon = (\varepsilon_1^T, \dots, \varepsilon_m^T)^T$$

$$R_i = \text{Var}(\varepsilon_i)$$

$$R = \text{Var}(\varepsilon) = \text{diag}(R_1, \dots, R_m)$$

ζ parameter vector that determines D and R

$$\lambda = (\beta^T, \zeta^T)^T$$

Hierarchical models for meta-analysis (Chapter 3)

$p(\cdot|\cdot)$ probability density of first argument conditional on second

θ_i parameter of interest for study i (possibly vector-valued)

$$\theta = (\theta_1^T, \dots, \theta_m^T)^T$$

λ hyperparameter common to all studies (possibly vector-valued)

$L_i(\theta_i) = p(y|\theta_i)$: likelihood for study i

$\ell_i(\theta_i) = \log L_i(\theta_i)$: log-likelihood for study i

Linear mixed effects models (Chapter 4)

- X_i $n_i \times q$ design matrix for fixed effects in study i
 $X = (X_1^T, \dots, X_m^T)^T$: $n \times q$ design matrix for fixed effects
 Z_i $n_i \times (c - 1)$ design matrix for random effects in study i
 $Z = (Z_1^T, \dots, Z_m^T)^T$: $n \times (c - 1)$ design matrix for random effects
 $\theta_i = X_i\beta + Z_i u_i$: n_i -dimensional vector for study i
 $\theta = X\beta + Zu = (\theta_1^T, \dots, \theta_m^T)^T$: n -dimensional vector.

Nonlinear mixed effects models (Chapter 5)

- A_i $r \times q$ design matrix for fixed effects in study i
 $A = (A_1^T, \dots, A_m^T)^T$: $mr \times q$ design matrix for fixed effects
 B_i $r \times (c - 1)$ design matrix for random effects in study i
 $B = \text{diag}(B_1, \dots, B_m)$: $mr \times m(c - 1)$ design matrix for random effects
 $\theta_i = A_i\beta + B_i u_i$: r -dimensional vector for study i
 $\theta = A\beta + Bu = (\theta_1^T, \dots, \theta_m^T)^T$: mr -dimensional vector
 $\eta_i(\theta_i)$ n_i -dimensional vector-valued nonlinear function for study i
 $\eta(\theta) = (\eta_1(\theta_1)^T, \dots, \eta_m(\theta_m)^T)^T$: n -dimensional vector-valued nonlinear function
 $\mu_i(\lambda) = E_\lambda(y_i) = E_\lambda[\eta_i(A_i\beta + B_i u_i)]$
 $V_i(\lambda) = \text{Var}_\lambda(y_i) = \text{Var}_\lambda[\eta_i(A_i\beta + B_i u_i)] + R_i$
 $I(f : g)$ Kullback-Leibler information criterion for discrepancy of g from f

Chapter 1

Introduction

A fundamental goal of science is synthesis—diverse natural phenomena are regarded as manifestations of an underlying structure. It is therefore ironic that investigators commonly focus on one data set at a time. For example, in fisheries management, data on a single population, or *stock*, are often analyzed independently of data on related stocks. Of course, any one data set may exhibit tremendous complexity, and much can be learned from carefully studying it in isolation. Yet there seems something perverse in ignoring related data.

Quantitative methods for combining data from different sources have a long history (Stigler 1986), but only in the mid-1970's did they become recognized as a distinct field, under the rubric of *meta-analysis*. Proponents of meta-analysis have pointed to the many benefits of combining information: increased power, understanding of effect modifiers, resolution of apparently contradictory findings, etc.

Yet there remains considerable skepticism and controversy about meta-analysis: psychologist H. J. Eysenck (1978) called it “an exercise in mega-silliness.” Concerns have been raised that meta-analysis “mixes good studies with bad,” that studies may in fact be measuring different things, that unknown factors may make combination of results invalid, that a single study may “drive” the results of a meta-analysis, and that “publication bias” may hide non-significant research findings and thereby bias meta-analytic summaries. A number of graphical methods have been developed to assess the likely import of the various hazards listed above. For example, an early attempt to diagnose the presence of publication bias was the funnel plot—a scatter plot of study estimates versus sample size—and it

remains useful today.

Of course any statistical analysis presents hazards, and graphical diagnosis of deviations from assumptions has become a major theme of recent study. A different approach has been the development of *robust statistics*. The goal has been to develop statistical procedures that are not overly sensitive to slight deviations from their assumptions, without paying a high price in terms of bias or efficiency. Many robust procedures effectively *downweight* outlying observations, and thus provide a diagnostic component. Robustness issues are plainly important in meta-analysis, although relatively little work has directly addressed this.

Another controversy in meta-analysis concerns the use of fixed versus random effects models. Fixed effects models assume that all studies are estimating the same unknown quantity. Random effects models instead assume that the “true effect” being estimated by each study comes from a distribution. In meta-analyses where only summary statistics (typically with standard errors and/or confidence intervals) are combined, the choice is essentially limited to fixed or random effects models. In other cases, however, raw data from each study may be available, and a broader array of modeling choices may be considered. For example, mixed effects models can have both fixed and random effects. Recently *hierarchical models*, which generalize mixed effects models, have become very popular. With improvements in computing power and Markov Chain Monte Carlo (MCMC) methods, Bayesian approaches to hierarchical modeling have been widely applied, in meta-analysis among other areas. Concerns remain, however, about adequacy of assumptions and other dangers.

This thesis concerns the use of nonlinear mixed effects models in meta-analysis, with a particular focus on robustness and graphical methods. I begin by introducing three data sets that will be used throughout the thesis. The practical questions of interest are as follows: (1) Are wolf populations in Québec declining? (2) Are ulcer treatments effective? and (3) What is the maximum reproductive rate of coho salmon? Answers to these questions are of considerable importance, however the broader goal of this thesis is to investigate and develop methodology.

1.1 Example 1: Wolves in Québec

Over the last 100 years, the geographical range of gray wolves (*Canis lupus*) has been greatly reduced (Larivière, Jolicoeur, and Crête 2000). During the 1970's, the province of Québec established a network of wildlife reserves where the harvest of game species, including wolves, became controlled by a system of registered traplines and a quota on hunting licenses. Recently, Larivière, Jolicoeur, and Crête (2000) reported on wolf population trends during the last 15 years in 9 reserves located in southern Québec. They used data from questionnaires distributed to moose hunters and an equation linking the questionnaire data to radio-tracking data. In one of the reserves, Ashuapmushuan, a different management scheme was used, and we omit the data from this reserve. Figure 1.1 displays the data from the remaining 8 reserves.

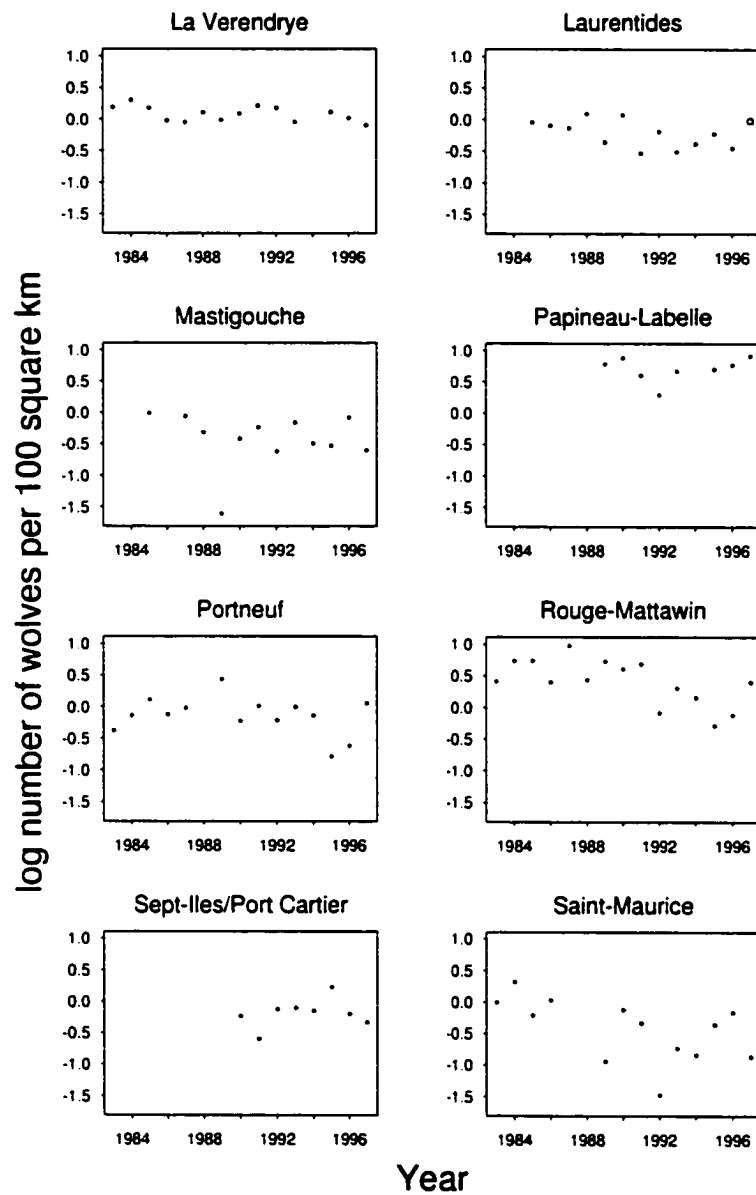


Figure 1.1: Log population numbers of wolves in 8 reserves in southern Québec. The 1997 observation in the Laurentides reserve is plotted as an open circle because the management scheme that year was changed.

Note that the 1989 observation in the Mastigouche reserve seems unusual: it is the lowest observation and deviates from the overall trend. Because it is not at either extreme, it

might not be expected to be very influential on estimates of the slope in a linear regression, however it would be expected to influence the estimate of the intercept, and to inflate the variance estimate.

There is a prior reason for treating the 1997 observation in the Laurentides reserve with caution: the management scheme that year was changed as an experiment.

To estimate long-term trends in wolf densities in each reserve, Larivière, Jolicoeur, and Crête (2000) performed a simple linear regression of density versus year, and concluded that, over the last 15 years, in 7 of the 9 reserves, wolf populations have been relatively stable, with the remaining 2 reserves showing declines. This approach—determining the proportion of studies for which a test exhibits statistical significance in the prescribed direction—is known as “vote counting” in the meta-analysis literature, and is inherently flawed (Hedges and Olkin 1985). Hedges and Olkin (1980) showed that as the number of studies becomes large, the proportion of studies yielding significant results is approximately equal to the average power of the test. Rather than providing inferences about the wolf populations, the procedure may in fact be providing information about the chosen test!

1.2 Example 2: Ulcer studies

Treatments for bleeding peptic ulcers have been investigated in a number of randomized control trials, but questions remain about their effectiveness. To attempt to answer these questions, Sacks et al. (1990) applied conventional meta-analytic methods to data from a number of ulcer trials. Using data from 41 studies obtained from the paper of Sacks et al., Efron (1996) applied a novel meta-analytic approach involving combination of likelihoods in an empirical Bayes framework. Morris (1996) noted several discrepancies between Efron’s data set and that of Sacks et al.; we use that of Efron (1996) for comparability.

Each study was a randomized clinical trial whose outcomes may be written as a 2×2 table (Table 1.1). The i th trial can be denoted (a_i, b_i, c_i, d_i) , where a_i and b_i are the numbers of failures and successes in the treatment group and c_i and d_i are the numbers of failures and successes in the control group. The elements a_i , b_i , c_i , and d_i are sometimes called the *cells* of the table.

	Failure	Success
Treatment	a_i	b_i
Control	c_i	d_i

Table 1.1: Two-by-two table for outcomes of a randomized clinical trial.

The data from all 41 studies are shown in Table 1.2.

i	a_i	b_i	c_i	d_i	i	a_i	b_i	c_i	d_i	i	a_i	b_i	c_i	d_i
1	7	8	11	2	15	3	22	11	21	29	0	22	8	16
2	8	11	8	8	16	4	7	6	4	30	2	16	10	11
3	5	29	4	35	17	2	8	8	2	31	1	14	7	6
4	7	29	4	27	18	1	30	4	23	32	8	16	15	12
5	3	9	0	12	19	4	24	15	16	33	6	6	7	2
6	4	3	4	0	20	7	36	16	27	34	0	20	5	18
7	4	13	13	11	21	6	34	13	8	35	4	13	2	14
8	1	15	13	3	22	4	14	5	34	36	10	30	12	8
9	3	11	7	15	23	14	54	13	61	37	3	13	2	14
10	2	36	12	20	24	6	15	8	13	38	4	30	5	14
11	6	6	8	0	25	0	6	6	0	39	7	31	15	22
12	2	5	7	2	26	1	9	5	10	40	0	34	34	0
13	9	12	7	17	27	5	12	5	10	41	0	9	0	16
14	7	14	5	20	28	0	10	12	2					

Table 1.2: Data from 41 studies on treatment of peptic ulcers listed by Efron (1996) from a meta-analysis by Sacks et al. (1990). For study i , $(a_i, b_i) = (\text{failures}, \text{successes})$ in treatment group, and $(c_i, d_i) = (\text{failures}, \text{successes})$ in control group.

Note that there are some fairly striking differences between the studies. For example, study number 6 had just 11 subjects, whereas study number 23 had 142. Some studies have zero-cells, indicating that subjects in either the treatment or control group had either no

successes or no failures.

It is often desired to estimate the *log odds ratio* of the *i*th table,

$$\theta_i = \log \left(\frac{P_i(\text{Failure}|\text{Treatment})P_i(\text{Success}|\text{Control})}{P_i(\text{Success}|\text{Treatment})P_i(\text{Failure}|\text{Control})} \right).$$

We may also wish to estimate a summary measure—e.g., the mean log odds ratio—over all of the studies.

1.3 Example 3: Coho salmon spawner-recruitment data

In recent years there have been alarming declines in coho salmon (*Oncorhynchus kisutch*) populations on the west coast of North America. A population dynamics approach may help to explain why this is happening and provide guidance for management.

Adult coho spawn in streams and rivers. About 1.5 years later, their offspring—known as smolts at this life stage—migrate to the sea. Another 1.5 years later, the survivors return to spawn (Figure 1.2).

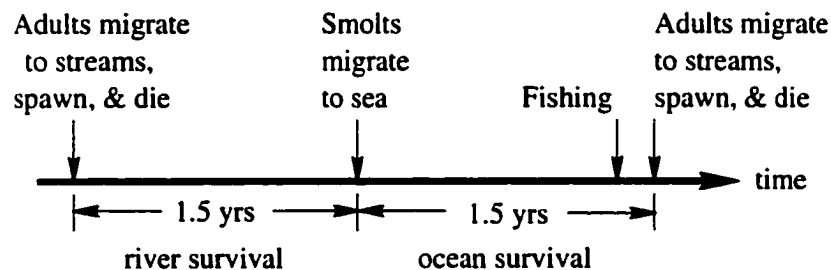


Figure 1.2: Coho salmon life history.

The goal of the analysis that will be developed in this thesis is to model the freshwater proportion of the life-history, so that this information can be combined with independent data on survival at sea to produce improved management models. For example, in Southern British Columbia coho salmon catches and escapements have declined in the last 20 years, and there has been considerable disagreements on the causes of these declines (Walters 1993; Walters and Ward 1998; Beamish, Mcfarlane, and Thomson 1999). However, it is clear that the survival at sea has greatly declined in recent years (Bradford, Myers,

and Irvine 2000). Our analysis of the freshwater survival can produce estimates of the mean and variation among stocks of the freshwater portion of the survival, which can then be combined with the long term data on survival at sea. Furthermore, we will produce population-specific optimal estimates for individual rivers.

Let S_{ij} represent the quantity of spawners from cohort j in population i , measured as the number of spawning females per kilometre of river, and let R_{ij} represent the quantity of “recruits” produced by those spawners, measured as the number of female smolts per kilometre of river. We consider data on 14 coho populations (Figure 1.3). An in-depth analysis of these and related data is in Bradford, Myers, and Irvine (2000).

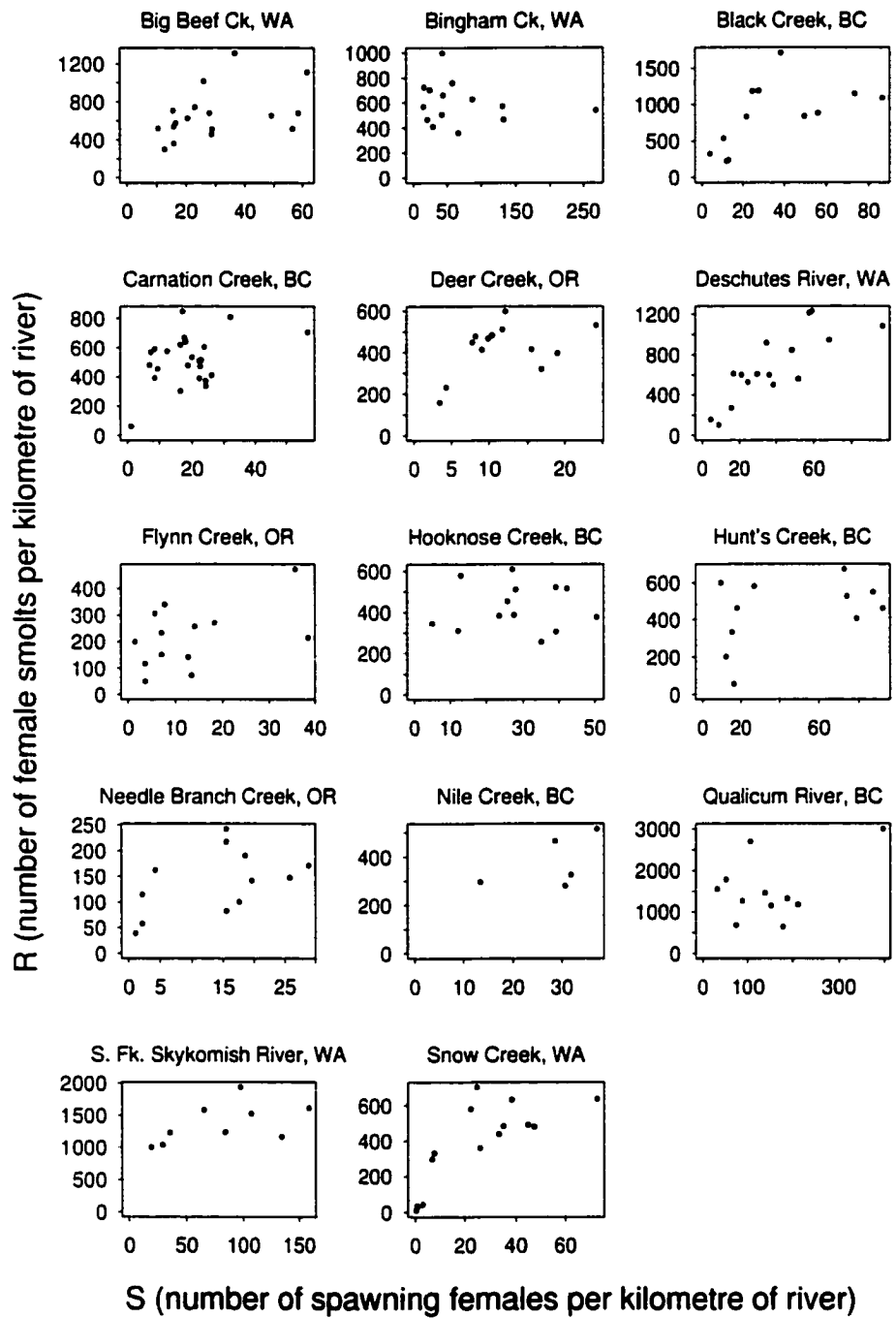


Figure 1.3: Coho salmon spawner-recruitment data.

Borrowing the words of Thomas Hobbes (1651), spawner-recruitment data can be characterized as “nasty, brutish, and short.” The longest spawner-recruitment series in Figure 1.3 has 24 observations (Carnation Creek, BC), and the series show tremendous variability (e.g., Hunt’s Creek, BC). When there are no spawners there can be no recruitment, and the series generally show increasing recruitment with increasing spawner quantity. We will seek spawner-recruitment models that exhibit this behaviour.

Outright errors often find their way into spawner-recruitment data. My initial analyses of the coho salmon data were flawed due to a data-entry error: the length of Hunt’s Creek, BC, had been incorrectly entered as 1.4 km instead of the correct value of 5.4 km. (The source of the error is unclear.) In many applications, this type of “gross error” in data is not uncommon: Hampel et al. (1986, p. 28) suggest that “1–10% gross errors in routine data seem to be more the rule rather than the exception.”

1.3.1 Spawner-recruitment models

Several parametric models for spawner-recruitment data have been proposed; three are shown in Figure 1.4. Simple biological justifications for each model have been proposed.

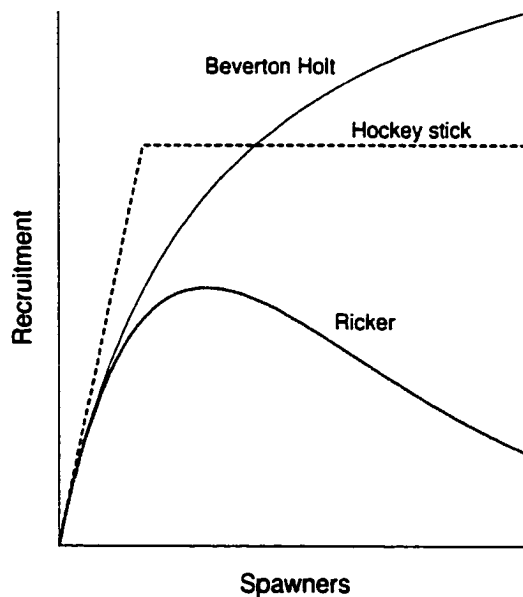


Figure 1.4: Typical shapes of three spawner-recruitment curves.

Spawner-recruitment models are generally nonlinear. The *Beverton-Holt* model (also known as the Michaelis-Menten model in enzyme kinetics) is

$$R_{ij} = \frac{S_{ij}}{1/\alpha_i + S_{ij}/R_i^{\max}}. \quad (1.1)$$

Since Beverton-Holt curves are nondecreasing and approach a horizontal asymptote, they seem to have roughly the behaviour we need to model coho salmon data. The parameters α_i and R_i^{\max} are both positive and have the following interpretations. Geometrically, α_i is the slope of the curve at the origin. It is interpreted biologically as the number of recruits per spawner at low spawner levels, or in the case of the coho salmon data, the number of female smolts per female spawner. Geometrically, R_i^{\max} is the asymptotic level of recruitment. It is interpreted biologically as the *carrying capacity* of a river, or in the case of the coho salmon data, the maximum number of female smolts per kilometre of river.

A brief comment on the units of measurement is in order. Careful consideration of the units of parameters is of fundamental importance in meta-analysis. For combination of results to be meaningful, estimates must be measured on the same scale or be dimensionless. In the wolf example, the rate of decline (or increase) may be measured on the same scale (log number of wolves per 100 square kilometre per year) in each reserve. In the ulcer example, the log odds ratio is dimensionless. Returning to the coho salmon data, recall that S_{ij} and R_{ij} were defined as the quantity of spawners and recruits *per kilometre of river*. This standardization by river length is important because it allows comparison of R_i^{\max} across rivers. Note however, that comparison of α_i across rivers requires no standardization. In general, for other species groups, recruitment should be standardized by the size of the habitat. For coho salmon, river length is a proxy for habitat size since coho reside primarily along the edge of a river.

Another spawner-recruitment model is the *Ricker* model,

$$R_{ij} = \alpha_i S_{ij} e^{-\beta_i S_{ij}}. \quad (1.2)$$

The Ricker model has the convenient property that it can be transformed to obtain a linear model,

$$\log \frac{R_{ij}}{S_{ij}} = \log \alpha_i - \beta_i S_{ij},$$

for which estimation tends to be relatively easy. However, the Ricker is not a good model for the overall dynamics of coho salmon, since it exhibits *overcompensation*, i.e., recruitment is not an increasing function of the spawner quantity. This does not seem to match the behaviour shown in Figure 1.3. Nevertheless, Barrowman and Myers (2000) showed that for a single stock, the Ricker model often gives more reasonable estimates of the slope at the origin than does the Beverton-Holt. Barrowman and Myers (2000) and Bradford, Myers, and Irvine (2000) also proposed the *hockey-stick* model,

$$R_{ij} = \alpha_i \min(S_{ij}, S_i^*) = \begin{cases} \alpha_i S_{ij} & \text{if } S_{ij} < S_i^* \\ \alpha_i S_i^* & \text{if } S_{ij} \geq S_i^* \end{cases} \quad (1.3)$$

and showed that it typically gives reasonable estimates of the slope at the origin as well as matching the population dynamics of coho salmon. Later on we will introduce two families of *generalized hockey stick* models that smooth the abrupt transition of the ordinary hockey stick.

The parameter α_i has dimensions of recruitment per spawner and, in all of the models considered here, gives the slope of the function at $S_{ij} = 0$. It is crucial in setting the limits of overfishing (Mace 1994; Myers and Mertz 1998). It is particularly easy to see why this is the case for species like coho salmon, which die after reproduction. For a given population of coho salmon, suppose that $\alpha_i = 50$ female smolts per female spawner, and assume there is 90% mortality during the ocean stage of the coho life cycle followed by 80% mortality from fishing. This would, on average, result in a stable population (Figure 1.5).

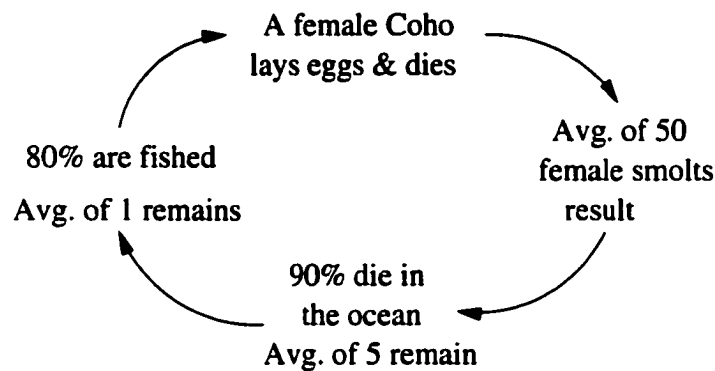


Figure 1.5: Schematic depiction of a stable population of coho salmon, assuming $\alpha_i = 50$ female smolts per female spawner, 90% ocean mortality, and 80% fishing mortality.

We will examine the fits for the above models simultaneously for the 14 rivers; we thus need to consider the patterns of deviations of the observations of recruitment from the mean behaviour model across stocks. Previous work has shown that in the marine environment, recruitment deviations are correlated at separations of roughly 500 km (Myers, Mertz, and Barrowman 1995; Myers, Mertz, and Bridson 1997) compared to less than 50 kilometers in the freshwater environment (Myers, Mertz, and Bridson 1997). These results apply for coho salmon, for which freshwater survival is almost independent among years for stocks greater than 20 kilometers apart, but marine survival is correlated at a much greater spatial scale (Bradford 1999). As described above, the spawner-recruitment data considered here concerns the freshwater part of the life cycle, thus we will assume in what follows that deviations from the spawner-recruitment relationship are independent among stocks.

We will also assume that within a stock there is no autocorrelation in the recruitment residuals among years (Bradford 1999). Myers, Bowen, and Barrowman (1999) considered linear mixed effects models for spawner-recruitment data allowing for autocorrelation.

1.4 Conventional approaches to meta-analysis

For the purposes of this thesis we define meta-analysis as quantitative methods for combining evidence across studies (Hedges and Olkin 1985, p. 13). Note that meta-analysis is often defined more broadly to include non-statistical considerations. These can be of

critical importance, but they are not the subject of the present work. For a broad-ranging survey of statistical and non-statistical aspects of meta-analysis see Cooper and Hedges (1994). For a shorter but very comprehensive tutorial see Normand (1999).

In a different respect, meta-analysis is often defined more narrowly: combination of raw data is excluded, and only the combination of summary statistics, such as standardized mean differences and correlations, is considered. We begin by discussing standard meta-analytic approaches in this narrow context.

1.4.1 Fixed effects meta-analysis: weighted means

Denote the parameter of interest by β and suppose there are m studies. Let t_1, \dots, t_m be independent estimates with common mean β and variances v_1, \dots, v_m . For now, assume that the t_i are normally distributed. Following Cox (1982), let $y_i = t_i / \sqrt{v_i}$ and note that $\text{Var}(y_i) = 1$. Since $E(y_i) = \beta / \sqrt{v_i}$ we can write

$$y_i = \frac{1}{\sqrt{v_i}}\beta + \varepsilon_i, \quad i = 1, \dots, m,$$

where $E(\varepsilon_i) = 0$ and $\varepsilon_1, \dots, \varepsilon_m$ are i.i.d. This is a simple linear regression with no intercept, and from the usual formulas, the least squares estimate of β is

$$\begin{aligned} \hat{\beta} &= \frac{SXY}{SXX} = \frac{\sum_{i=1}^m t_i / v_i}{\sum_{i=1}^m 1 / v_i} \\ &= \sum_{i=1}^m w_i t_i / \sum_{i=1}^m w_i, \end{aligned} \tag{1.4}$$

where $w_i = 1/v_i$. This is just a weighted mean of the estimates, with weights inversely proportional to the variances. To obtain a standard error for $\hat{\beta}$, note that

$$\begin{aligned}\text{Var}(\hat{\beta}) &= \left(\sum_{i=1}^m w_i \right)^{-2} \sum_{i=1}^m w_i^2 \text{Var}(t_i) \\ &= \left(\sum_{i=1}^m w_i \right)^{-2} \sum_{i=1}^m w_i \\ &= \left(\sum_{i=1}^m w_i \right)^{-1}.\end{aligned}\tag{1.5}$$

A check on the validity of our assumptions may be obtained by examining the residual sum of squares

$$\begin{aligned}SSE &= SY^2 - \frac{SXY^2}{SXX} = \sum_{i=1}^m t_i^2/v_i - \frac{(\sum_{i=1}^m t_i/v_i)^2}{\sum_{i=1}^m 1/v_i} \\ &= \sum_{i=1}^m w_i t_i^2 - \frac{(\sum_{i=1}^m w_i t_i)^2}{\sum_{i=1}^m w_i},\end{aligned}\tag{1.6}$$

and since $\text{Var}(y_i) = 1$, it follows that $SSE \sim \chi_{m-1}^2$. We shall call SSE the *chi-squared homogeneity statistic*; too large a value of SSE suggests that the assumptions are in error. For example, the estimates may not in fact have a common mean, or the variances may not be correctly specified.

Of course, we rarely know the variances v_i , and must use estimated weights in (1.4–1.6).

1.4.2 Random effects meta-analysis

Suppose that t_1, \dots, t_m are actually estimating different quantities $\theta_1, \dots, \theta_m$, i.e., conditional on θ_i , t_i has mean θ_i and variance v_i , for $i = 1, \dots, m$. Write $\theta_i = \beta + u_i$. It may be possible to model the u_i in terms of explanatory variables. Alternatively, the u_i may be taken as random. For example, suppose the u_i are i.i.d. normal with zero mean and variance σ_u^2 . Then, marginally, t_i is normally distributed with mean β and variance $\sigma_u^2 + v_i$, for $i = 1, \dots, m$. We call σ_u^2 the *random effects variance* and v_i the *estimation variance*. Provided both σ_u^2 and the v_i are known, the weighted mean (1.4) with weights $w_i = (\sigma_u^2 + v_i)^{-1}$

can be used to estimate β .

We can write this as a *hierarchical model*:

$$t_i | \theta_i, v_i \sim N(\theta_i, v_i) \quad (1.7)$$

$$\theta_i | \beta, \sigma_u^2 \sim N(\beta, \sigma_u^2). \quad (1.8)$$

Alternatively, we can write the model as

$$\begin{aligned} t_i &= \beta + u_i + \varepsilon_i \\ u_i &\stackrel{\text{iid}}{\sim} N(0, \sigma_u^2) \text{ independent of} \\ \varepsilon_i &\stackrel{\text{indep}}{\sim} N(0, v_i) \\ i &= 1, \dots, m. \end{aligned} \quad (1.9)$$

1.4.3 Empirical Bayes estimation

Though the focus of meta-analysis is usually to estimate β , it is sometimes of interest to obtain the “best” estimate of the effect for a particular study. Let us denote the study of interest as study 0 (normally one of the m studies) with estimate t_0 and estimation variance v_0 . In the context of model (1.9), suppose we wish to estimate the realized value of θ_0 , the parameter specific to study 0. A reasonable estimator is the mean of θ_0 given t_0 . If v_0 , β , and σ_u^2 are all known, then the distribution of θ_0 conditional on t_0 is normal with mean

$$B_0\beta + (1 - B_0)t_0 \quad (1.10)$$

and variance

$$v_0(1 - B_0),$$

where $B_0 = v_0 / (v_0 + \sigma_u^2)$. Note that $0 \leq B_0 \leq 1$. When $B_0 = 0$, v_0 is negligible compared to σ_u^2 , and our estimate (1.10) of θ_0 is simply the observed value t_0 . On the other hand, when $B_0 = 1$, σ_u^2 is negligible compared to v_0 , and the estimate of θ_0 is “shrunk” from the observed value t_0 to β . For values of B_0 between zero and one, the result is partial “shrinkage” of the observed value t_0 towards β . For this reason, B_0 is called the *shrinkage factor* for study 0.

If we consider the distribution of (1.8) to be a prior distribution for θ_0 , then the distribution of θ_0 conditional on t_0 is the posterior distribution of θ_0 . The posterior mean (1.10) is a weighted mean of the estimate t_0 and the grand mean β , and the shrinkage of t_0 towards β incorporates our prior knowledge.

Since we do not know β and σ_u^2 , and perhaps not v_0 either, one strategy is to replace them by estimates in (1.10). This is the *parametric empirical Bayes approach* and it incorporates all the data into the estimate for a single study. This is sometimes called “borrowing strength” from other studies: when the information contained in an individual study is relatively weak (so that v_0 is large), the information contained in all of the other studies is “borrowed” to obtain better estimates.

For example, if we knew σ_u^2 and v_0 , but not β , we might estimate the realized value θ_0 by

$$B_0\hat{\beta} + (1 - B_0)t_0, \quad (1.11)$$

where $\hat{\beta}$ is an estimate of β based on t_1, \dots, t_m .

1.4.4 Mixed effects meta-analysis

Model (1.9) is called a random effects model because all terms, except for the grand mean, β , are random effects. We may wish to include fixed effects in the model, however, by defining

$$\theta_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_q x_{iq} + u_i, \quad (1.12)$$

where x_{i1}, \dots, x_{iq} are characteristics of study i . The model (1.12) is known as a *mixed effects model* since it has both random effects and fixed effects (in addition to β_1). In the remainder of this work, when mixed effects models are mentioned, they should be understood to include random effects models as a special case.

If we collect the fixed effects into a vector $\beta = (\beta_1, \dots, \beta_q)^T$, we can write (1.12) as

$$\theta_i = x_i^T \beta + u_i,$$

where $x_i = (1, x_{i2}, \dots, x_{iq})^T$. In Chapter 4, we will generalize this further to give the general

linear mixed effects models for repeated measures data Laird and Ware (1982).

Mixed effects models are a type of *hierarchical model*: in the model given by (1.7) and (1.8), the hierarchy has two levels. As discussed above, we typically do not know β and σ_u^2 and one way to proceed is to treat them as fixed unknown values to be estimated, as in the empirical Bayes approach. Alternatively we can place prior distributions on these parameters, which is to say add another level to the hierarchy, and then proceed using Bayesian inference.

1.5 Meta-analysis of raw data

When only summary statistics are available, analysis is usually limited to the methodologies discussed above, and the estimated effect sizes, t_i , are generally assumed to be approximately normal. In Chapter 2 it will be shown that there are cases where the normality approximation is dubious, and, where possible, it is advisable to use the raw data instead. This thesis focuses on this broader context, and the richer variety of models that may be entertained.

Summary statistics, or *individual estimates* as we will call them, are nevertheless of considerable value. In conducting a meta-analysis it is important to be cautious in several respects. Summary statistics provide an indication of what the individual studies are telling us, without the intermediary of a possibly complex statistical model. Another way to be cautious is by using robust statistical procedures; Chapter 2 includes a discussion of one approach to robust estimation for regression models, namely *M-estimation*.

Modeling may be greatly simplified when we have single-parameter likelihoods, and Chapter 2 also features a review of the elimination of nuisance parameters. Single-parameter likelihoods are also exploited in Chapter 2 to develop a new graphical display for meta-analysis called the “raindrop plot”. The raindrop plot provides a compact and informative way of displaying the information provided by groups of studies, particularly in cases with small sample sizes and nonlinear models where individual likelihoods may exhibit deviations from normality.

1.5.1 Repeated measures data

In the examples considered in this work, the data have a *repeated measures* structure in which several observations are available from each study. Suppose the observation vector for study i is of length n_i . In the wolf data, there are $m = 8$ studies (reserves) with observation vectors of length 14, 13, 12, 8, 14, 15, 8, and 13. The ulcer data consists of $m = 41$ studies whose 2×2 table structure means that the observation vectors for each study are of length 4. In the coho salmon data, there are $m = 14$ studies (rivers) with observation vectors of lengths 17, 14, 12, 24, 13, 16, 13, 13, 11, 12, 5, 11, 9, and 15.

In this thesis we will use the term “studies” generically. In the literature, other terms such as clusters, groups, units, individuals, and subjects are commonly used.

One type of repeated measures data is *longitudinal* data, where the observations in each study are ordered, typically by time. The wolf and coho salmon data are longitudinal, while the ulcer data are not.

1.5.2 Hierarchical models for raw-data meta-analysis

Chapter 3 provides an overview of hierarchical models for meta-analysis, with special emphasis on the connections between frequentist, empirical Bayes, and fully Bayes inference. This provides a broad framework for the mixed effects models discussed in subsequent chapters. Efron (1996) proposed an empirical Bayes approach to combining likelihoods and illustrated it using the ulcer data. We review his approach in Chapter 3 and use raindrop plots to display some of his results. We also consider the application of Efron’s approach to meta-analysis of the coho salmon data.

One of the simplest cases of raw data meta-analysis occurs when we wish to model linear relationships in several studies. The wolf data provide a simple example of this. The Ricker spawner-recruitment model, upon transformation, provides another example. Chapter 4 introduces linear mixed effects models and reviews methods for parameter estimation and inference, as well as approaches to robust estimation.

Except for the Ricker model, the spawner-recruitment models introduced in section 1.3.1 are nonlinear, and in Chapter 5 nonlinear mixed effects models are introduced and methods for parameter estimation and inference are reviewed. Approaches to robust estimation of nonlinear mixed models are discussed, and two methods proposed.

Chapter 2

Individual estimates

A cautious approach to modeling dictates that before we consider combining data from several studies, we should carefully examine the individual data sets, and estimates obtained from each one. We follow Davidian and Giltinan (1995) in referring to these as *individual estimates*. Of course, if the studies in fact concern related phenomena, then this approach will be inefficient. But it protects us from hasty judgements—there may be much to be gained from combining information, but we should not lose sight of the hazards.

We begin, in sections 2.1–2.3, by returning to our examples to illustrate how individual estimates are obtained in different contexts. In section 2.4, we consider robust estimation for individual data sets, and illustrate the diagnostic use of observation-weights obtained from a robust analysis using a new type of graphical display. In multi-parameter problems, the focus of interest may be a single parameter. In section 2.5, we review methods for the elimination of so-called “nuisance” parameters, with illustrations involving the ulcer data and the salmon data. In nonlinear problems, particularly with small sample sizes, conventional methods for displaying individual estimates may be misleading or inadequate. In section 2.6, we introduce the *raindrop plot*, a new graphical display for meta-analysis that circumvents the problems with conventional methods.

We use the following generic notation for our data. There are assumed to be m studies, with n_i observations in study i . In study i , we denote the j th observation by y_{ij} and let $y_i = (y_{i1}, \dots, y_{in_i})^T$ denote the entire vector of observations for that study. Finally the observations for all studies are collected into one vector $y = (y_1^T, \dots, y_m^T)^T$.

2.1 Example: Wolves in Québec

For the i th wolf population, let y_i denote the vector of log numbers of wolves per 100 square km and let x_i denote the associated vector of years. Larivière, Jolicoeur, and Crête (2000) performed linear regressions on each of the data sets individually. For the i th data set, the model is

$$y_i = \theta_{i1} + \theta_{i2}x_i + \varepsilon_i, \quad (2.1)$$

where θ_{i1} is the intercept, θ_{i2} is the slope, and $\varepsilon_i \sim N(0, \sigma_i^2 I)$. Figure 2.1 shows the individual regressions.

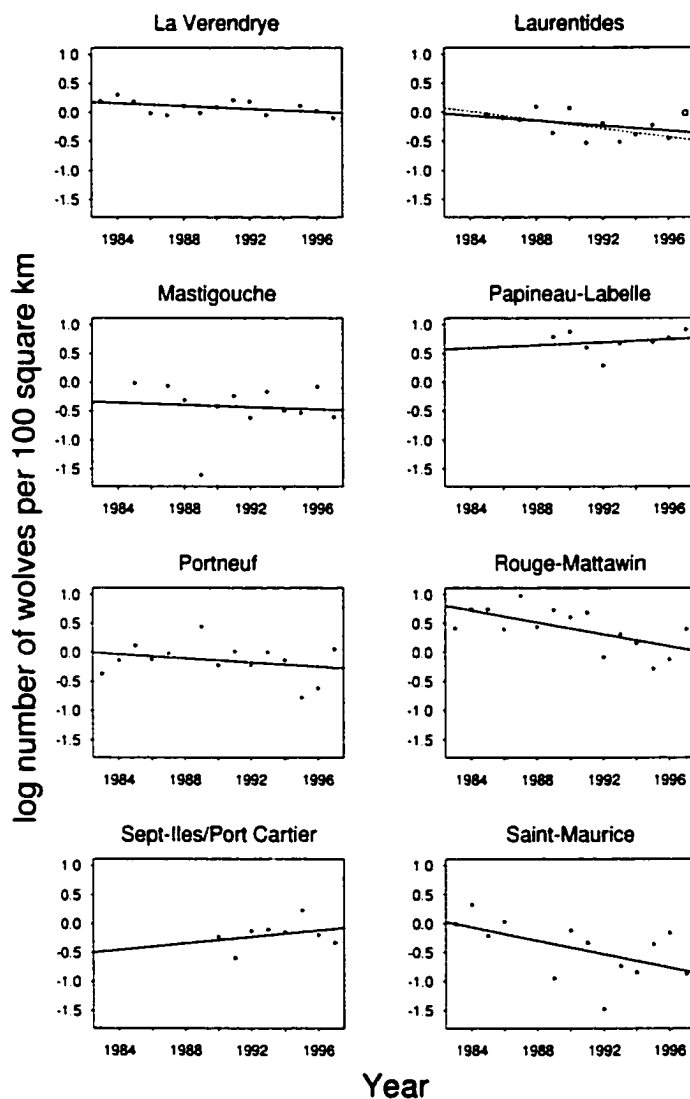


Figure 2.1: Log population numbers of wolves in 8 reserves in southern Québec with fitted least squares regression lines. The 1997 observation in the Laurentides reserve is plotted as an open circle because the management scheme that year was changed. The dashed line shows the regression excluding that point.

Only two of the reserves give slopes significantly different from zero at the 95% confidence level—Rouge Mattawin, and Saint Maurice—and both show declines. Two of the reserves show apparent (but non-significant) increases: Papineau-Labelle and Sept-Iles/Port

Cartier.

Olkin (1999) encourages the meta-analyst to “Plot, plot, plot whenever and whatever you can.” A standard display used in meta-analysis consists of a sequence of point estimates and confidence intervals from individual studies followed by a meta-analytic summary, typically a combined estimate with a confidence interval (Light et al. 1994; Galbraith 1988). This has also been called a forest plot (Bijens et al. 1996). In subsequent chapters of this work, we will consider meta-analytic summaries, but for now, the focus is on individual estimates. Figure 2.2 shows a display of the individual estimates of the slopes of the regression lines in Figure 2.1.

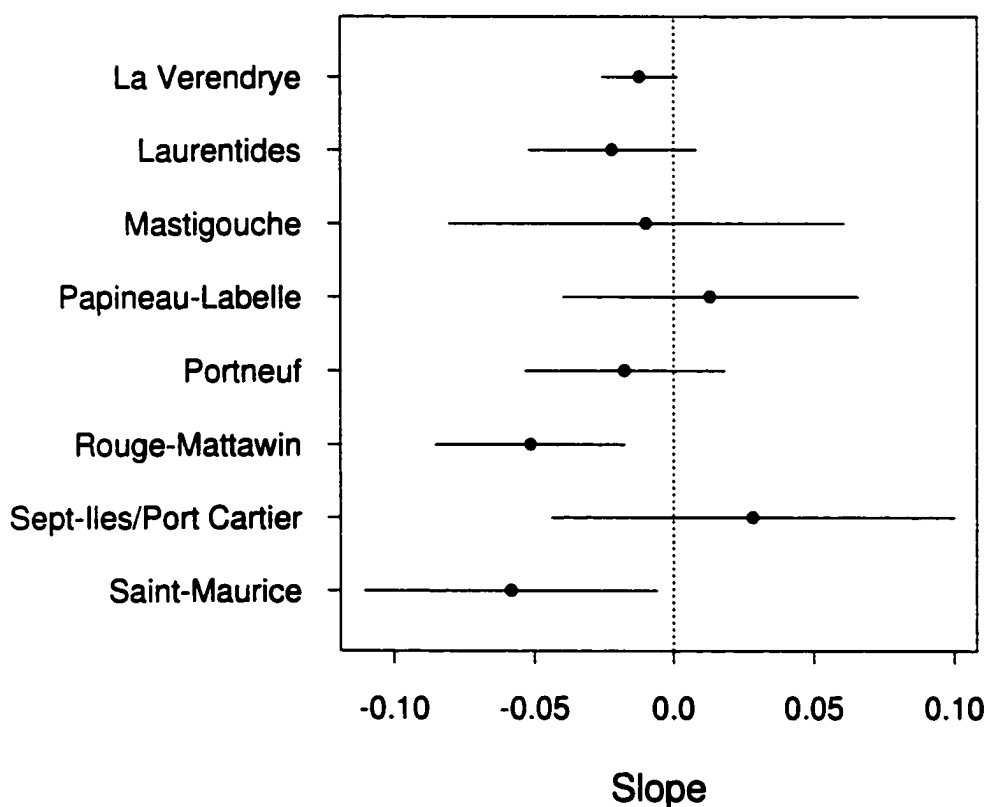


Figure 2.2: Point estimate and 95% confidence interval for each reserve for the slope of the least squares regression of log numbers of wolves versus year.

Figures 2.2 and 2.1 show completely independent reserve-specific estimates. In particular, the error variances, σ_i^2 , are assumed to be unrelated. As Figure 2.1 shows, the error

variances differ between reserves. For example, the error variance for the La Vérendrye reserve is clearly much smaller than that for the Saint-Maurice reserve. It may be that the true error variances differ between studies solely because of differences in the effective sample size underlying each observation from the various reserves. Approximate effective sample sizes, n_i^* , for observations from each reserve were obtained by multiplying the number of moose hunting zones by the number of hunting periods (since the estimates of wolf densities were obtained from moose hunters) and are shown in Table 2.1.

Reserve	Effective sample size, n_i^*
La Vérendrye	378
Laurentides	355
Mastigouche	168
Papineau-Labelle	198
Portneuf	100
Rouge-Mattawin	130
Sept-Îles/Port-Cartier	48
Saint-Maurice	66

Table 2.1: Effective sample sizes for observations from each reserve.

If we assume that the error variance for each reserve is given by σ^2/n_i^* , then we can estimate a single variance parameter, σ^2 , thereby linking the 8 regressions. The resulting slope estimates are no longer “individual estimates” in the sense used above, but the assumption is relatively weak, and is unlikely to distort our conclusions seriously.

2.2 Example: Ulcer studies

Recall that the data for trial i are in the form of a 2×2 table (a_i, b_i, c_i, d_i) , where a_i and b_i are the numbers of failures and successes in the treatment group and c_i and d_i are the numbers of failures and successes in the control group. It is often desired to estimate the

log odds ratio of the i th table,

$$\theta_i = \log \left(\frac{P_i(\text{Failure}|\text{Treatment})P_i(\text{Success}|\text{Control})}{P_i(\text{Success}|\text{Treatment})P_i(\text{Failure}|\text{Control})} \right).$$

The i th sample log odds ratio is given by

$$\hat{\theta}_i = \log \left(\frac{a_i d_i}{b_i c_i} \right), \quad (2.2)$$

with approximate asymptotic standard error

$$\text{SE}(\hat{\theta}_i) = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right)^{\frac{1}{2}} \quad (2.3)$$

(Agresti 1990). However when at least one of the cells is zero, $\hat{\theta}_i$ does not exist, and the alternative estimator

$$\tilde{\theta}_i = \log \left(\frac{(a_i + \frac{1}{2})(d_i + \frac{1}{2})}{(b_i + \frac{1}{2})(c_i + \frac{1}{2})} \right) \quad (2.4)$$

is often used, together with standard error

$$\text{SE}(\tilde{\theta}_i) = \left(\frac{1}{a_i + \frac{1}{2}} + \frac{1}{b_i + \frac{1}{2}} + \frac{1}{c_i + \frac{1}{2}} + \frac{1}{d_i + \frac{1}{2}} \right)^{\frac{1}{2}} \quad (2.5)$$

(Agresti 1990). Both $\hat{\theta}_i$ and $\tilde{\theta}_i$ are asymptotically normal, but for small sample sizes their sampling distributions are highly skewed, and use of the standard display of point estimate plus or minus two standard errors (Figure 2.3) may be misleading. In studies with zero cells this is particularly acute. Study number 5, for example, has $a = 3$, $b = 9$, $c = 0$, and $d = 12$. In other words, there were no failures in the control group. Therefore, $\hat{\theta}_i = \infty$, which suggests that the true log odds ratio is arbitrarily large. The alternative estimator gives $\tilde{\theta}_i = 2.22$ with an approximate 95% confidence interval of $(-0.86, 5.3)$. But this is rather arbitrary, and could be very misleading if the control really were much better than the treatment.

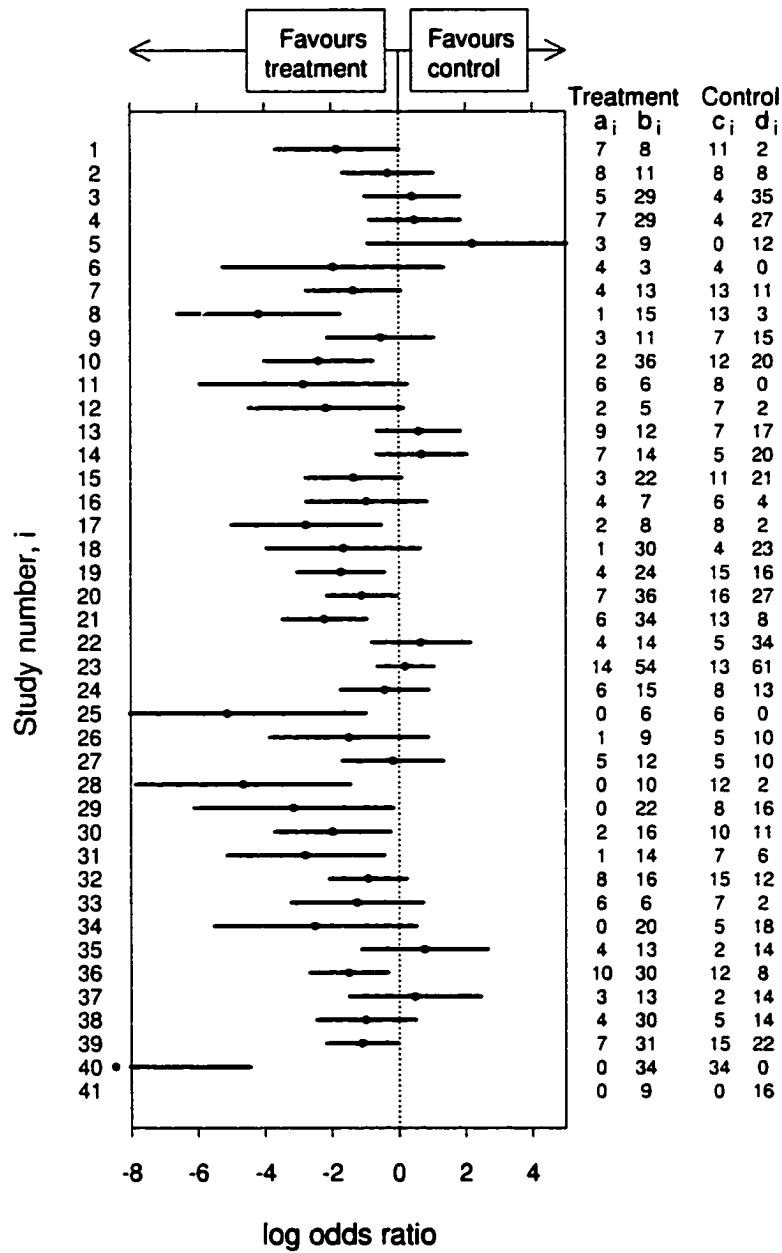


Figure 2.3: Point estimate and 95% confidence interval for the log odds ratio in each ulcer study. When none of the cells of the 2×2 table for a study are zero, expressions (2.2) and (2.3) are used. When at least one of the cells is zero, expressions (2.4) and (2.5) are used.

2.3 Example: Coho salmon

We begin by considering fits to the coho salmon spawner-recruitment data using the Beverton-Holt model. The Beverton-Holt model with multiplicative lognormal error is

$$R_{ij} = \frac{S_{ij}}{1/\alpha_i + S_{ij}/R_i^{\max}} e^{\varepsilon_{ij}},$$

where the ε_{ij} are i.i.d. normal with mean zero and variance σ_ε^2 , and where $\alpha_i > 0$ is the slope at the origin and $R_i^{\max} > 0$ is the asymptotic level of recruitment. Dividing both sides by S_{ij} and taking the logarithm gives

$$y_{ij} = -\log(1/\alpha_i + S_{ij}/R_i^{\max}) + \varepsilon_{ij}, \quad (2.6)$$

where $y_{ij} = \log(R_{ij}/S_{ij})$. To interpret y_{ij} , note that in fish reproduction the quantity of eggs produced is typically proportional to the quantity of spawners S_{ij} . The ratio R_{ij}/S_{ij} is thus an index of survival from the egg stage to the smolt stage, and we typically refer to y_{ij} as *log survival*.

Equation (2.6) specifies a nonlinear regression model. Maximizing the likelihood for this model is equivalent to minimizing the sum of squares

$$\sum_{j=1}^{n_i} [y_{ij} + \log(1/\alpha_i + S_{ij}/R_i^{\max})]^2. \quad (2.7)$$

Maximum likelihood Beverton-Holt fits to the coho salmon data are shown in Figure 2.4.

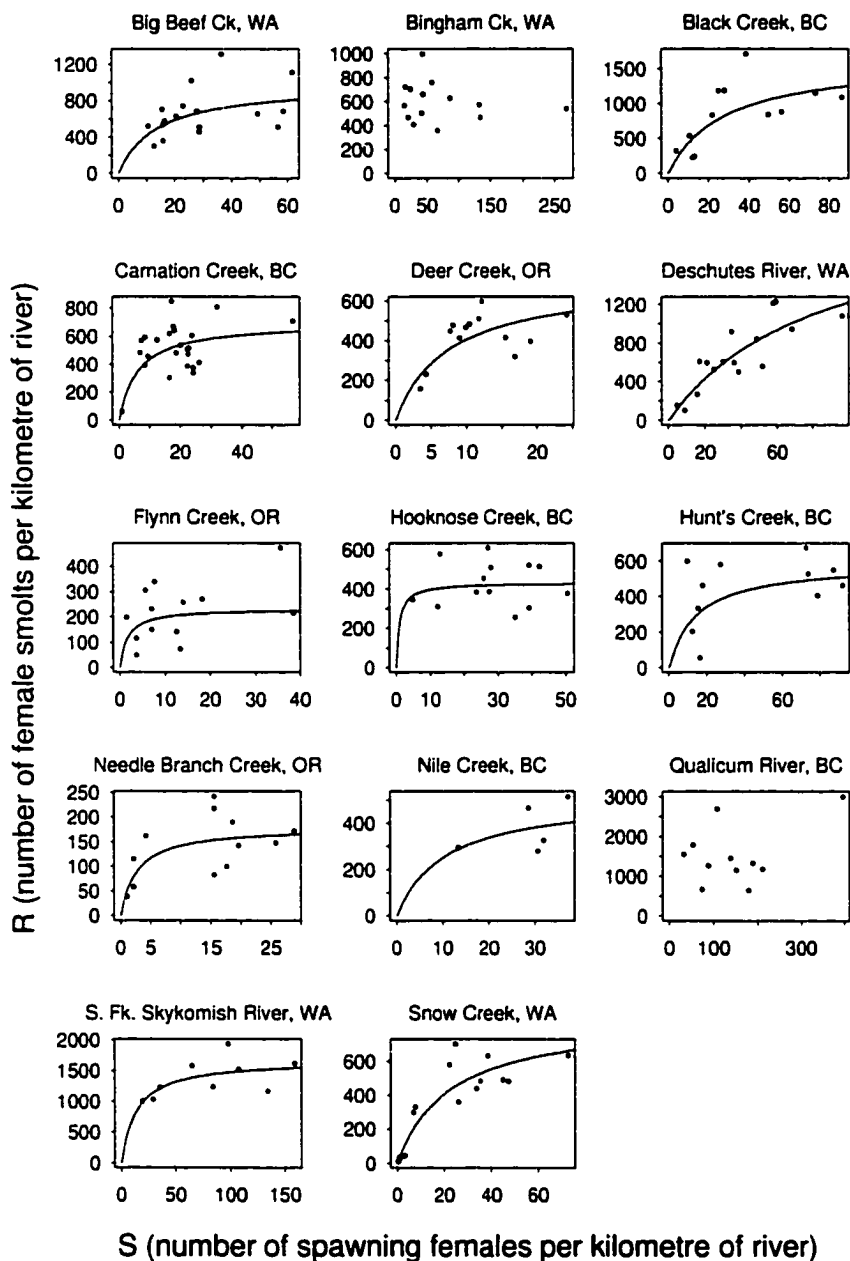


Figure 2.4: Coho salmon data with superimposed median recruitment curves from individual maximum likelihood fits of the Beverton-Holt model assuming a lognormal recruitment distribution. Note that for Bingham Creek, WA, and Qualicum River, BC, no fitted curves are shown because the likelihood is not maximized by a finite slope at the origin.

For two rivers, Bingham Creek and Qualicum River, the data suggest that the slope at the origin is arbitrarily large, i.e., $\alpha_i = \infty$ or equivalently $1/\alpha_i = 0$. Let $\pi_i = 1/\alpha_i$, so that the sum of squares (2.7) is

$$\sum_{j=1}^{n_i} [y_{ij} + \log(\pi_i + S_{ij}/R_i^{\max})]^2. \quad (2.8)$$

Let $\hat{\pi}_i$ and \hat{R}_i^{\max} be the least squares estimates of π_i and R_i^{\max} . Equating the derivative of (2.8) with respect to R_i^{\max} to zero, we have

$$\sum_{j=1}^{n_i} [y_{ij} + \log(\hat{\pi}_i + S_{ij}/\hat{R}_i^{\max})] \frac{S_{ij}}{\hat{R}_i^{\max}(\hat{\pi}_i \hat{R}_i^{\max} + S_{ij})} = 0.$$

When $\hat{\pi}_i = 0$, this reduces to

$$n_i \log \hat{R}_i^{\max} = \sum_{j=1}^{n_i} y_{ij} + \sum_{j=1}^{n_i} \log S_{ij}.$$

But since $y_{ij} = \log(R_{ij}/S_{ij})$, we have

$$\hat{R}_i^{\max} = \sqrt[n_i]{\prod_{j=1}^{n_i} R_{ij}},$$

the geometric mean of the observed recruitments, which we denote \bar{R}_i . Equating the derivative of (2.8) with respect to π_i to zero, we have

$$\sum_{j=1}^{n_i} [y_{ij} + \log(\hat{\pi}_i + S_{ij}/\hat{R}_i^{\max})] / [\hat{\pi}_i + S_{ij}/\hat{R}_i^{\max}] = 0.$$

When $\hat{\pi}_i = 0$, this reduces to

$$\sum_{j=1}^{n_i} [y_{ij} + \log(S_{ij}/\bar{R}_i)] / S_{ij} = 0.$$

Again, substituting $y_{ij} = \log(R_{ij}/S_{ij})$, this becomes

$$\sum_{j=1}^{n_i} \log(R_{ij}/\bar{R}_i)/S_{ij} = 0.$$

But $\hat{\pi}_i = 0$ is a boundary of the region of admissible values for π_i , and so for $\hat{\pi}_i = 0$ to be a least squares estimate, it is necessary that

$$\sum_{j=1}^{n_i} \log(R_{ij}/\bar{R}_i)/S_{ij} \geq 0.$$

Indeed this condition does hold for Bingham Creek and Qualicum River. Furthermore, in both of these cases, when α_i is not constrained to be positive, a numerical optimizer converges to a negative estimate. We conclude that for these cases, the likelihood is maximized by an infinite slope at the origin. But this is not credible, because there cannot be more female smolts than the number of eggs produced by a female spawner, which though large (approximately 3570 (Hutchings and Morris 1985)), is finite.

Referring once again to Figure 2.4, we see that some of the individual estimates of the slope at the origin, α_i , are quite reasonable and well determined. For example, for Black Creek, the slope at the origin is estimated to be roughly 60 female smolts per female spawner. In cases such as Hooknose Creek, however, extremely high estimates are obtained with poor precision. The difficulty is that precise determination of the slope at the origin depends on observations at low levels of S , which tend to be sparse.

Alternative spawner-recruitment models (Figure 1.4, p. 10) can give quite different estimates of the slope at the origin, α_i . To understand why, we express the Ricker model (1.2) and hockey-stick model (1.3) in terms of log survival $y_{ij} = \log(R_{ij}/S_{ij})$. For the Ricker model, we have

$$y_{ij} = \log \alpha_i - \beta_i S_{ij},$$

which is linear. For the hockey-stick model, note that $\alpha_i S_i^*$ is the maximum recruitment, and we therefore write $R_i^{\max} = \alpha_i S_i^*$. We therefore have

$$y_{ij} = \begin{cases} \log \alpha_i & \text{if } S_{ij} < S_i^* \\ \log(R_i^{\max}/S_{ij}) & \text{if } S_{ij} \geq S_i^*. \end{cases}$$

Barrowman and Myers (2000) compared fits of the three spawner-recruitment models using an unstandardized version of the coho salmon dataset. (The units of spawners and recruitment were not divided by river length.) Figure 2.5 shows fits for three of the coho populations.

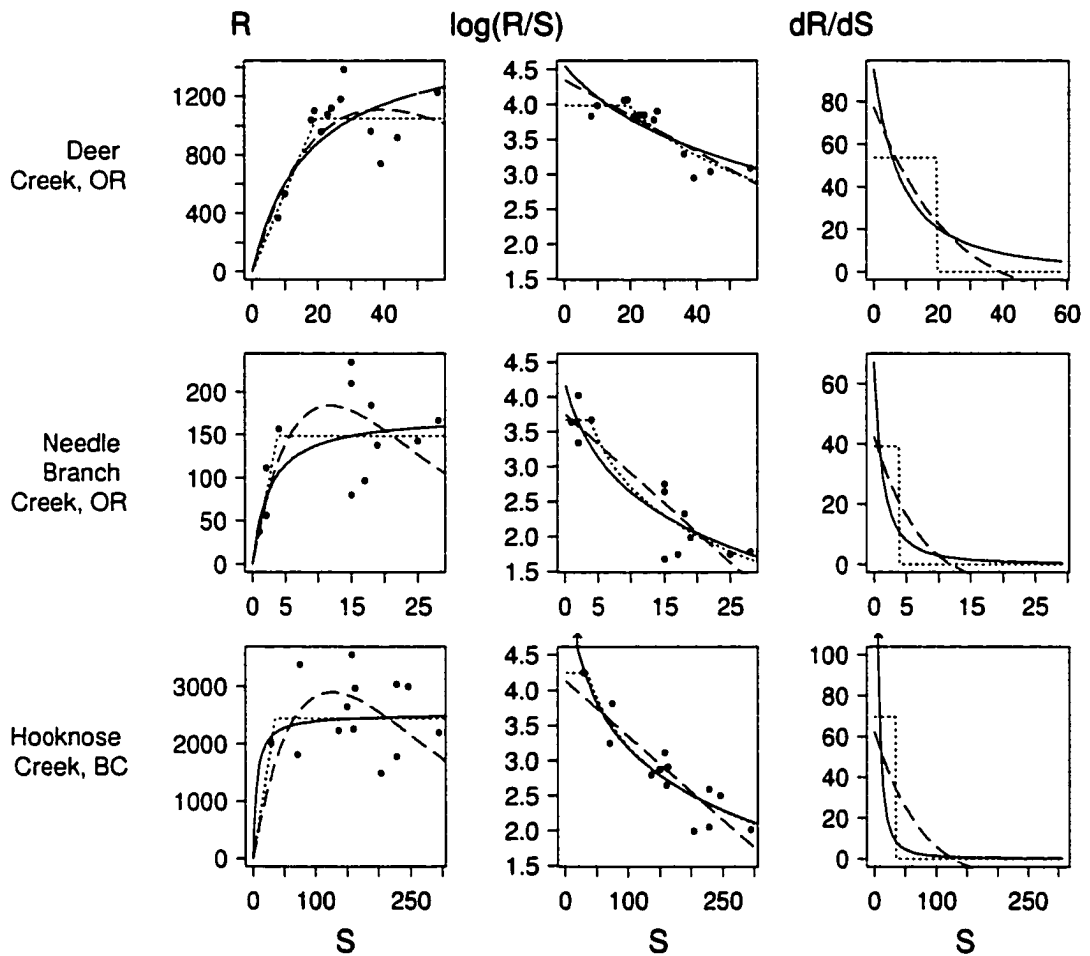


Figure 2.5: Coho salmon spawner-recruitment data for three populations (rows) with superimposed median recruitment curves from individual maximum likelihood fits of the Beverton-Holt model (solid line), the Ricker model (dashed line), and the hockey-stick model (dotted line) assuming a lognormal recruitment distribution. The first column shows R versus S ; the second shows $\log(R/S)$ (log survival) versus S with fitted curves on this scale; the third column shows the derivative of R with respect to S versus S for the fitted curves.

Examination of the $\log(R/S)$ versus S panels in Figure 2.5 shows why the estimates of α_i are so different. Note that when $S_{ij} = 0$, we have $y_{ij} = \log \alpha_i$, so that on this scale, the y -intercept gives $\log \alpha_i$. The different models give very different extrapolations as $S \downarrow 0$.

The hockey-stick model predicts that log survival remains at observed levels (i.e. a horizontal extrapolation), while the Ricker model predicts that the linear trend in log survival continues (i.e. a linear extrapolation). The Beverton-Holt model, however, predicts a sharp increase in log survival as $S \downarrow 0$. We typically have

$$\alpha_{\text{Beverton-Holt}} \gg \alpha_{\text{Ricker}} \approx \alpha_{\text{hockey-stick}}.$$

The choice of spawner-recruitment model for the coho salmon data is therefore not an easy one. The Ricker model seems to give reasonable estimates of α but does not match the spawner-recruitment dynamics of coho salmon for larger values of S . The hockey-stick model also seems to give reasonable (though different) estimates of α , and is predicted by a simple biological model (Barrowman and Myers 2000). However its abrupt bend seems implausible and can lead to estimation difficulties. In particular the likelihood surface may feature flat areas and multiple local maxima. Whereas standard nonlinear optimization procedures can be used to fit Ricker or Beverton-Holt models, a grid search is required for the hockey-stick model (Lerman 1980). The Beverton-Holt model is also predicted by a simple biological model and its smoothness seems plausible, however it makes a very strong extrapolation of log survival at low spawner levels and can result in infinite estimates of α .

Alternatively, a formal model selection criterion can be used to choose the “best” model. The Akaike Information Criterion (AIC) is commonly used for this purpose. Since each of the three models has the same number of parameters (two), choosing the model with the largest log likelihood is equivalent to using the AIC for model selection. The maximized individual log likelihood for each of the coho salmon populations using each model are given in Table 2.2.

	Ricker	Beverton-Holt	Hockey stick
Big Beef Ck, WA	-4.6	-4.2	-4.2
Bingham Ck, WA	-7.2	0.3	0.3
Black Creek, BC	-5.3	-4.9	-4.4
Carnation Creek, BC	-12.4	-8.9	-5.6
Deer Creek, OR	3.5	2.5	7.6
Deschutes River, WA	-4.3	-3.8	-4.7
Flynn Creek, OR	-12.1	-9.7	-10.2
Hooknose Creek, BC	-0.8	1.0	1.1
Hunt's Creek, BC	-8.9	-8.2	-8.5
Needle Branch Creek, OR	-3.4	-2.1	-1.1
Nile Creek, BC	2.7	3.6	3.5
Qualicum River, BC	-8.1	-4.5	-4.5
S. Fk. Skykomish River, WA	5.4	7.3	5.7
Snow Creek, WA	-4.9	-4.5	-3.4
sum	-60.5	-36.3	-28.6

Table 2.2: Maximized log likelihoods for each of the coho salmon populations using the Ricker, Beverton-Holt, or hockey-stick models. The final row of the table shows the sum of the individual maximized log likelihoods for each model.

In only one case (Deer Creek), is the maximized log likelihood for the Ricker model larger than that for the Beverton-Holt model. Summing the individual maximized likelihoods (last line of Table 2.2) shows that the Beverton-Holt model provides much better fitting of the data. The comparison between the Beverton-Holt and the hockey-stick model is more equivocal. The AIC favours the Beverton-Holt for seven of the populations and the hockey stick for the other seven. For several populations, however, the maximized likelihood for the hockey stick is considerably larger than that for the Beverton-Holt, so that the sum of the maximized likelihoods for the hockey-stick model is larger than for the Beverton-Holt. On the basis of these results we focus on the Beverton-Holt and hockey-stick models in this work.

2.4 Approaches to robust estimation

We have seen in the examples that the presence of “gross errors” and deviant observations is of considerable concern. If our estimators are overly sensitive to the presence of such flaws, our inferences may be distorted. The approaches to estimation in this thesis are likelihood-based, and as Hilborn and Mangel (1997) note, quoting David Fournier,

“ ‘The problem with likelihood is that some observations are just too unlikely.’ That is, some outliers will dominate the likelihood, and the fitting procedures often go to great lengths to make predictions closer to the outlier so that the total likelihood will not be too low.”

A number of different approaches to robust estimation have been developed during the past 30 years; for an introduction, see Staudte and Sheather (1990). In our context, we would like to modify likelihood-based estimation procedures to make them more robust. One way to do this is using *M-estimators*, a generalization of maximum likelihood estimators introduced below. We concentrate on *M-estimators* in this thesis, using the wolf and coho salmon data for illustration. In this section, we first introduce *M-estimators* in a general context, and then specialize to the regression case required for the wolf and salmon data sets.

For a single study, i , the *M-estimator* $\hat{\theta}_i$ of the study parameter θ_i solves an equation of the form

$$\sum_{j=1}^{n_i} \psi(y_{ij}, \hat{\theta}_i) = 0, \quad (2.9)$$

for some function ψ . For example, if $y_{ij} \stackrel{\text{iid}}{\sim} f_{\theta_i}$, $j = 1, \dots, n_i$, then the log likelihood for θ_i can be written

$$\ell_i(\theta_i) = \sum_{j=1}^{n_i} \log f_{\theta_i}(y_{ij}).$$

An estimating equation may be obtained by differentiating $\ell_i(\theta_i)$ with respect to θ_i and equating with zero at $\theta_i = \hat{\theta}_i$, i.e.

$$\ell'_i(\hat{\theta}_i) = \sum_{j=1}^{n_i} \psi(y_{ij}, \hat{\theta}_i) = 0,$$

where $\psi(y_{ij}, \theta_i) = \partial \log f_{\theta_i}(y_{ij}) / \partial \theta_i$.

A simple example helps to illustrate these concepts. Suppose f_{θ_i} is the normal distribution with known variance σ^2 and mean $\theta_i = \mu_i$. Then we have $\psi(y_{ij}, \mu_i) = (y_{ij} - \mu_i) / \sigma$. Denoting the j th scaled residual by $r_{ij} = (y_{ij} - \mu_i) / \sigma$, we have $\psi(y_{ij}, \mu_i) = r_{ij}$. In other words, the likelihood depends on the data and parameters only through the residuals, i.e. we have a location/scale invariance. Extreme values of r_{ij} will have a strong influence on the estimate of μ_i . With a small abuse of notation, we can write $\psi(y_{ij}, \theta_i) = \psi(r_{ij})$ where $\psi(r) = r$ is the identity function, which is unbounded in r .

One way to assess the robustness of an estimator is by considering its *influence function*. Suppose we have a functional $T(F_{n_i})$ where F_{n_i} is the empirical distribution and the goal is to obtain an estimate of $T(F)$ where F is the unknown data-generating distribution. The influence function of T measures the infinitesimal behaviour of the functional in response to contamination of the distribution. To formalize this, let Δ_y denote a point mass distribution at y , and consider the mixture distribution $F_{y,\varepsilon} = (1 - \varepsilon)F + \varepsilon\Delta_y$. Sampling from $F_{y,\varepsilon}$ models contamination: with probability $1 - \varepsilon$, we obtain a “good” observation (from F), and with probability ε , we obtain a “bad” observation (at y). The *influence function* of T at F is defined for each y by

$$\text{IF}(y) = \lim_{\varepsilon \downarrow 0} \left[\frac{T(F_{y,\varepsilon}) - T(F)}{\varepsilon} \right].$$

The influence function $\text{IF}(y)$ is a directional derivative of T at F in the direction of $\Delta_y - F$. For robustness, we wish to have the influence function bounded in y . For M -estimates of the general form (2.9), it can be shown that the influence function is bounded if and only if ψ is bounded.

Two common choices of bounded ψ are Huber’s function

$$\psi_b(r) = \max(-b, \min(r, b)),$$

and Tukey’s biweight

$$\psi_b(r) = \begin{cases} r(1 - (r/b)^2)^2 & \text{if } -b \leq r \leq b, \\ 0 & \text{otherwise,} \end{cases}$$

where b is a tuning constant in each of the above functions. The top panel of Figure 2.6 shows these two ψ -functions.

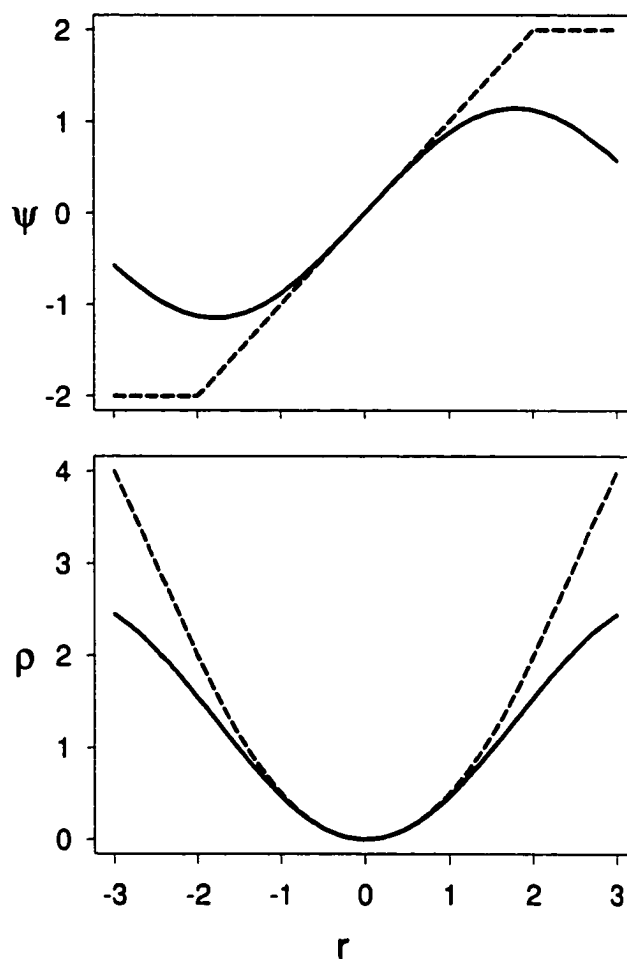


Figure 2.6: Top panel: Tukey's ψ -function with $b = 4$ (solid) and Huber's ψ -function with $b = 2$ (dashed). Bottom panel: Tukey's ρ -function with $b = 4$ (solid) and Huber's ρ -function with $b = 2$ (dashed).

The Tukey ψ -function is said to be *redescending*: near the origin ψ is nondecreasing but far from the origin ψ decreases toward the axis. Later on, we will see why this is an important property.

We will see that it is also useful to consider the antiderivatives of ψ -functions, called ρ -functions. That is, we define $\psi(r) = \rho'(r)$. The ρ functions corresponding to the Huber and Tukey ψ -functions are shown in the bottom panel of Figure 2.6.

We now consider M -estimators for regression models. Consider a linear regression model

$$y_i = X_i \theta_i + \varepsilon_i.$$

Model (2.1) has this form with $X_i = (1_{n_i} \ x_i)$, where 1_{n_i} is an n_i -dimensional vector of 1's, and $\theta_i = (\theta_{i1}, \theta_{i2})^T$. Relaxing the assumption of normal errors, we assume that

$$\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \frac{1}{\sigma_i} f\left(\frac{\varepsilon_{ij}}{\sigma_i}\right),$$

for some pdf f . Given x_i , the y_{ij} 's are independent with density

$$\frac{1}{\sigma_i} f\left(\frac{y_{ij} - (X_i \theta_i)_j}{\sigma_i}\right),$$

where $(X_i \theta_i)_j$ is the j th row of $X_i \theta_i$. The likelihood is thus

$$\frac{1}{\sigma_i^{n_i}} \prod_{j=1}^{n_i} f\left(\frac{y_{ij} - (X_i \theta_i)_j}{\sigma_i}\right),$$

giving log likelihood

$$-n_i \log \sigma_i + \sum_{j=1}^{n_i} \log f\left(\frac{y_{ij} - (X_i \theta_i)_j}{\sigma_i}\right).$$

Differentiating this equation with respect to θ_{ik} and equating to zero gives

$$\sum_{j=1}^{n_i} \frac{1}{\hat{\sigma}_i} \frac{-f'}{f}\left(\frac{y_{ij} - (X_i \hat{\theta}_i)_j}{\hat{\sigma}_i}\right) (X_i)_{jk} = 0,$$

where f' is the derivative of f , $\frac{-f'}{f}(r) = f'(r)/f(r)$, $\hat{\theta}_i$ is an estimate of θ_i , $\hat{\sigma}_i$ is an estimate of σ_i , and $(X_i)_{jk}$ is the (j, k) th element of X_i . Writing $\psi(r) = \frac{-f'}{f}(r)$, we can simplify this

to obtain

$$\sum_{j=1}^{n_i} \psi \left(\frac{y_{ij} - (X_i \hat{\theta}_i)_j}{\hat{\sigma}_i} \right) (X_i)_{jk} = 0. \quad (2.10)$$

When f is the standard normal distribution, we have $\psi(r) = r$. Given an estimate $\hat{\sigma}_i$ of the scale parameter σ_i and for a particular choice of ψ , (2.10) defines an M -estimate of θ_i . Note that these M -estimators are not robust against the effects of leverage, i.e. unusual design points.

2.4.1 Weights as diagnostics

A convenient algorithm for solving (2.10) can be obtained as follows. Write (2.10) as

$$\sum_{j=1}^{n_i} \frac{\psi(r_{ij})}{r_{ij}} r_{ij} (X_i)_{jk} = 0,$$

where $r_{ij} = (y_{ij} - (X_i \hat{\theta}_i)_j) / \hat{\sigma}_i$. Then define $w_{ij} = \psi(r_{ij}) / r_{ij}$, to obtain

$$\sum_{j=1}^{n_i} w_{ij} r_{ij} (X_i)_{jk} = 0, \quad (2.11)$$

which is the k th normal equation for a weighted least squares regression. Because the weights, w_{ij} , depend on the regression parameter estimates θ_i , an iterative algorithm must be used to solve (2.11). When the algorithm has converged, the observation-weights, w_{ij} will be of considerable interest, since they tell us which observations from each study appear to be unusual. The converged parameter estimates fit the bulk of the data and the weights identify unusual observations.

2.4.2 Example: Wolves in Québec

As noted earlier there is at least one suspicious observation: the 1989 observation at Mastigouche. A robust regression procedure applied to these data can be helpful in assessing the sensitivity of the individual regression estimates, and in identifying unusual observations. To display the weights, we propose a new type of display called a *whisker-weight*

scatterplot. The usual scatterplot with fitted regression lines is augmented with “whiskers” in the margins, displaying the robust weights. Full-length whiskers indicate that no down-weighting has occurred; shorter whiskers indicate the presence of down-weighting. An alternative is to use variable-size points in the scatterplot, however this can be distracting. For the wolf data, the observations are equally spaced in time and we display the whiskers on the top margin of the plot (Figure 2.7). More generally, we might wish to use whiskers in both margins; this duplicates information, but can be helpful when the scatter makes visual association of weights and observations difficult.

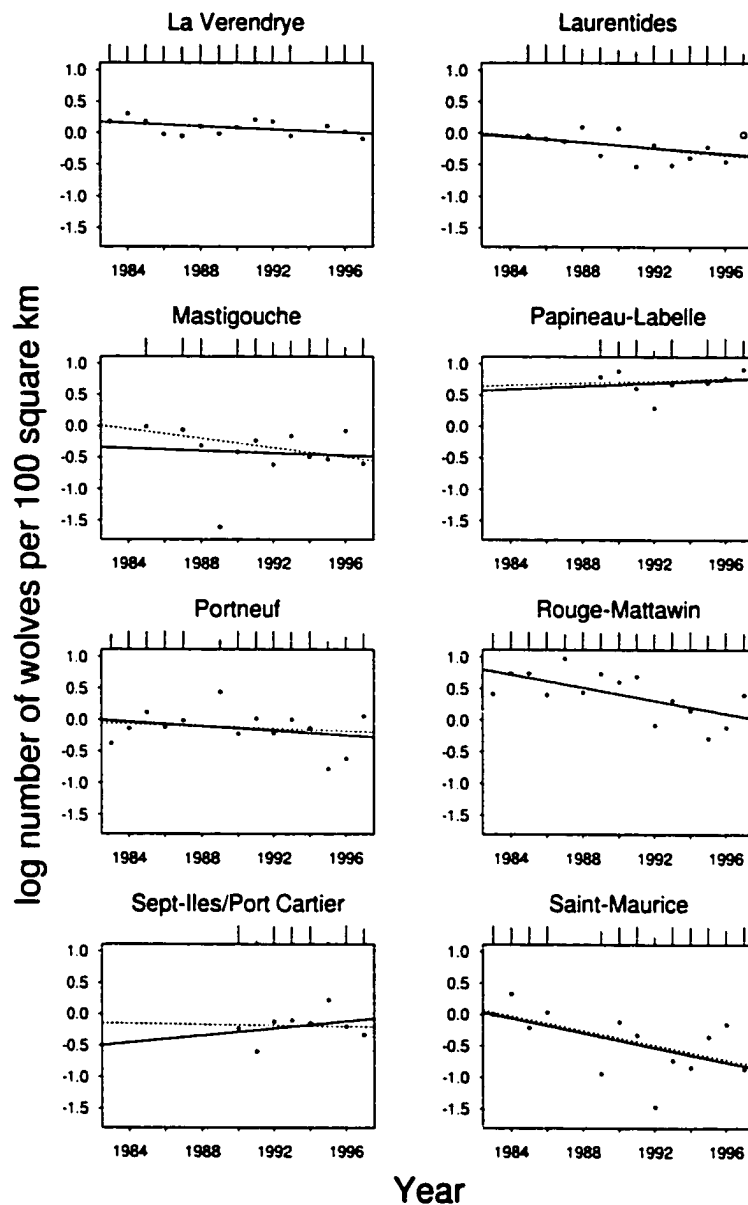


Figure 2.7: Log population numbers of wolves in 8 reserves in southern Québec with fitted least squares regression lines (solid) and robust regression lines (dotted). Along the top of each panel short vertical lines (“whiskers”) indicate the weight for each observation from the robust regression. The 1997 observation in the Laurentides reserve is plotted as an open circle because the management scheme that year was changed.

When an observation is completely downweighted, no vertical line is visible. Note that the

1989 observation at Mastigouche has been entirely eliminated from the analysis. Together with the downweighting of several other points for this reserve, the result is a substantial change in the estimated slope. Several observations at other reserves are also strongly downweighted, e.g. the 1997 observation at Laurentides (which we have a prior reason to reject) and the 1992 observation at Papineau-Labelle. Two points, 1991 and 1995, are downweighted in the Sept-Iles/Port Cartier data, resulting in a substantial change in slope. For such small data sets, it is not clear whether to accept the results from the robust regression. Nevertheless, it helps to highlight unusual observations and sensitivity in our estimates.

It may be argued that simply displaying residual plots conveys similar information to the whiskers, however the whisker-weight scatterplot provides a compact display of data and information about unusual points. What is more, the whiskers show the extent of downweighting by a particular robust procedure, which may vary depending on the choice of tuning constant, ρ -function, etc.

2.5 Elimination of nuisance parameters

For the ulcer data, we saw that for studies with zero cells, point estimates of the log odds ratio are problematic: the sample log odds ratio is not finite while the estimator based on adding $\frac{1}{2}$ to each cell gives rather arbitrary values. Even when there are no zero cells, the standard errors may be misleading. An alternative approach is to try to obtain a “likelihood” which would contain all of the information from a study concerning the log odds ratio.

Denote the full parameter vector for study i by τ_i and the associated likelihood function by $L_i(\tau_i)$. Given data from a single study, the likelihood principle dictates that inferences about τ_i should depend on the data, y_i , only through $L_i(\tau_i)$. In each of our three examples, the full parameter vector τ_i can be decomposed into a univariate parameter of interest θ_i and a (possibly multi-variate) nuisance parameter ν_i , i.e., $\tau_i = (\theta_i, \nu_i)$.

Example: Wolves in Québec

For each wolf reserve, the slopes of the regression line is of primary interest, and the intercept is a nuisance parameter—it must be estimated in order to estimate the slope.

Example: Coho salmon

For each coho salmon population, our primary interest is in the slope at the origin of the spawner-recruitment curve, i.e., the number of female smolts per female spawner. Other parameters of the spawner-recruitment curve, e.g., the asymptotic level of recruitment of the Beverton-Holt, are of secondary importance.

Example: Ulcer studies

For the ulcer studies, our interest is in the log odds ratios, rather than the baseline risks in the control or treatment groups. Denoting

$$p_i = P_i(\text{Success}|\text{Treatment}) \quad \text{and} \quad q_i = P_i(\text{Success}|\text{Control}),$$

the log odds ratio of the i th study is

$$\theta_i = \log \left(\frac{(1-p_i)q_i}{p_i(1-q_i)} \right) = \text{logit}q_i - \text{logit}p_i.$$

Solving for q_i , we have

$$q_i = q_i(\theta_i, p_i) = \text{logit}^{-1}(\theta_i + \text{logit}p_i).$$

In the general notation of this section, the nuisance parameter is $v_i = p_i$, so that the full parameter vector for the i th study is $\tau_i = (\theta_i, p_i)$. For each study, we treat the size of the treatment group, $r_i = a_i + b_i$, and the size of the control group, $s_i = c_i + d_i$, as being fixed and we model the outcomes in the treatment and control groups as being independent. Finally, we assume that

$$b_i \sim \text{Binomial}(p_i, r_i), \quad \text{and} \quad d_i \sim \text{Binomial}(q_i, s_i).$$

The likelihood for the i th study is thus

$$L_i(\theta_i, p_i) = \binom{a_i + b_i}{b_i} p_i^{b_i} [1 - p_i]^{a_i} \binom{c_i + d_i}{d_i} q_i(\theta_i, p_i)^{d_i} [1 - q_i(\theta_i, p_i)]^{c_i}.$$

Study-specific real-valued parameters of interest

In each of the above examples there is a study-specific real-valued parameter of interest. For each study, we would therefore like to “eliminate” the nuisance parameters from the likelihood in order to focus attention on the parameter of interest. Berger, Liseo, and Wolpert (1999) review several methods for eliminating nuisance parameters and give related references. We consider three of them below.

2.5.1 Conditional likelihood

In some situations, the conditional distribution of the data, y_i , given some statistic $z_i = z(y_i)$, does not depend on the nuisance parameter v_i . In this case, we call the conditional pdf (or pmf, for discrete data) the *conditional likelihood* for θ_i . By an abuse of notation, we denote the conditional likelihood by simply $L_i(\theta_i)$.

Example: Ulcer studies

Treating the margins of the i th 2×2 table as fixed, the conditional likelihood for θ_i is given by

$$L_i(\theta_i) = \binom{a_i + b_i}{a_i} \binom{c_i + d_i}{c_i} e^{\theta_i a_i / S_i(\theta_i)}, \quad (2.12)$$

where $S_i(\theta_i)$ is the sum of the numerator above over the allowable choices of a_i subject to the marginal constraints of the table.

Denote the log conditional likelihood by $\ell_i(\theta_i)$, and let θ_i^{MLE} be the conditional MLE of θ_i . A likelihood-ratio based 95% confidence interval for θ_i is given by

$$\{\theta_i : 2[\ell_i(\theta_i^{\text{MLE}}) - \ell_i(\theta_i)] \leq \chi_1^2(0.95)\},$$

or, setting $\ell_i(\theta_i^{\text{MLE}}) = 0$ and since $\chi_1^2(0.95)/2 \approx 1.92$,

$$\{\theta_i : \ell_i(\theta_i) \geq -1.92\}.$$

As an example, consider data $a = 1$, $b = 10$, $c = 15$, $d = 10$. Figure 2.8 shows the conditional likelihood for θ_i based on these data, along with two other approximate likelihoods. In this case the estimated log odds given by (2.2) is very close to the MLE. However the likelihood is asymmetric about its maximum: the 95% confidence interval based on the normal approximation and the standard error given by (2.3) is $(-4.91, -0.50)$, whereas the likelihood-ratio based interval is substantially different: $(-5.65, -0.85)$.

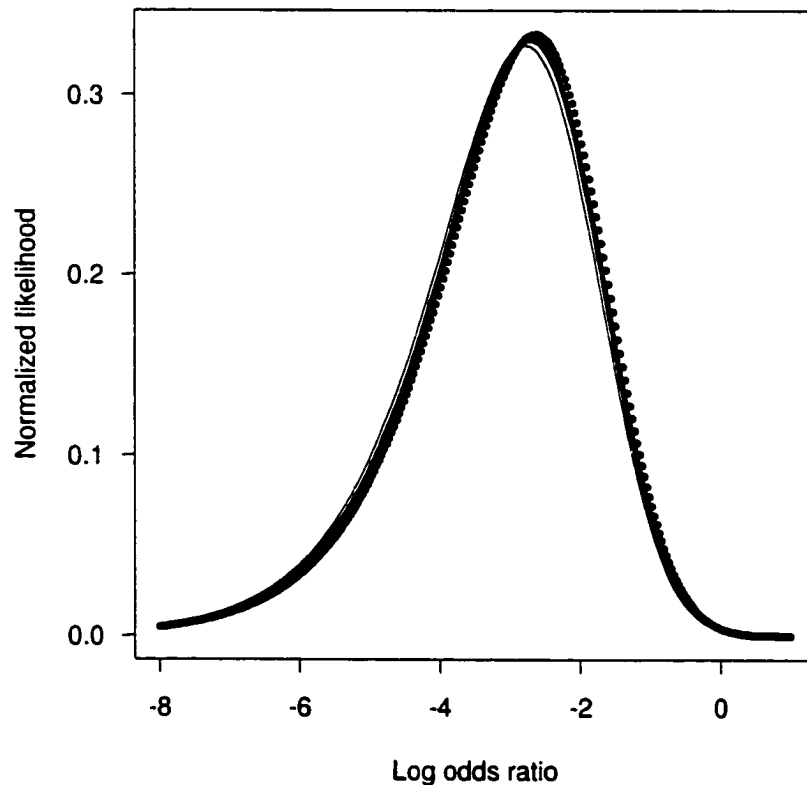


Figure 2.8: Three approximate likelihoods for the log odds ratio, based on data $a = 1$, $b = 10$, $c = 15$, $d = 10$, normalized to integrate to 1 over the range of the graph. The dotted curve is the conditional likelihood; the light curve is the uniform-integrated likelihood; the heavy curve is the profile likelihood. Note that the maximum of the profile likelihood gives the unconditional MLE of the log odds ratio.

2.5.2 Integrated likelihood

Berger, Liseo, and Wolpert (1999) focus on integrated likelihoods for eliminating nuisance parameters. In Bayesian inference, the integrated likelihood is given by

$$L^B(\theta_i) = \int L(\theta_i, \mathbf{v}_i) p(\mathbf{v}_i | \theta_i) d\mathbf{v}_i,$$

where $p(\mathbf{v}_i | \theta_i)$ is the conditional prior of \mathbf{v}_i given θ_i . Berger, Liseo, and Wolpert (1999) state that

“Even if one is not willing to entertain subjective Bayesian analysis, we feel that use of integrated likelihood is to be encouraged. The integration must then be with respect to default or noninformative priors.”

For example, the *uniform-integrated likelihood* is

$$L_i^U(\theta_i) = \int L_i(\theta_i, \mathbf{v}_i) d\mathbf{v}_i,$$

which is the same as the Bayesian integrated likelihood using a uniform (improper “flat”) prior, $p(\mathbf{v}_i | \theta_i) = 1$. However, in some problems where there are ridges in $L_i(\theta_i, \mathbf{v}_i)$, the uniform-integrated likelihood may not exist.

Example: Ulcer studies

The uniform-integrated likelihood is

$$L_i^U(\theta_i) = \int L_i(\theta_i, p_i) dp_i.$$

The uniform-integrated likelihood for the example data set ($a = 1$, $b = 10$, $c = 15$, $d = 10$) is shown in Figure 2.8.

2.5.3 Profile likelihood

The *profile likelihood* replaces integration by maximization:

$$\hat{L}_i(\theta_i) = \sup_{\mathbf{v}_i} L_i(\theta_i, \mathbf{v}_i).$$

Later on in this thesis, we will see several variations on this theme.

Particularly when there are large numbers of nuisance parameters or the likelihood surface has sharp ridges, the profile likelihood can give misleading behaviour (Berger, Liseo, and Wolpert 1999) and several corrections have been proposed, e.g. the modified profile likelihood of Barndorff-Nielsen (1983).

Example: Ulcer studies

For an ulcer study, the profile likelihood is

$$\hat{L}_i(\theta_i) = \sup_{p_i} L(\theta_i, p_i).$$

The profile likelihood for the example data set ($a = 1$, $b = 10$, $c = 15$, $d = 10$) is shown in Figure 2.8. Platt (1998) has evaluated the use of the modified profile likelihood for the log odds ratio of 2×2 tables.

2.6 Raindrop plots

To assist with meta-analysis, in this section we propose the raindrop plot for displaying individual estimates from many studies simultaneously. In later chapters we will repeatedly make use of the raindrop plot, and introduce an extension of the raindrop plot for use in hierarchical modeling.

The conventional plot showing a point estimate of a parameter with confidence limits is often not adequate to display the information contained in studies where the likelihood is not approximately normal, as can occur with small sample sizes or nonlinear models. Furthermore, these plots are sometimes misinterpreted by users as implying that all values within a confidence interval are equally plausible. Since the likelihood ratio provides a gauge of the relative plausibility of different values of θ_i and allows for asymmetry, a graphical display of the likelihood ratio would be desirable. An extra dimension is needed to represent the likelihood ratio. The principles of graph construction and the paradigm of graphical perception developed by Cleveland (1985) provide a guide to designing an appropriate display. To display likelihood ratios, we introduce a new kind of display, the *raindrop plot*, so called because the visual effect is reminiscent of raindrops streaking across a car

window. Figure 2.9 shows how a “raindrop” shape is obtained for the example data set of section 2.5.1, using the conditional log likelihood.

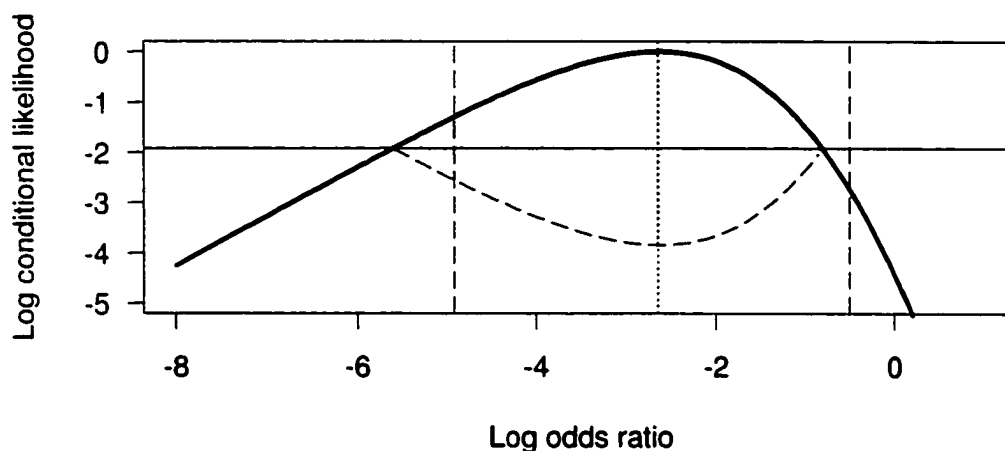


Figure 2.9: Conditional log likelihood for the log odds ratio, based on data $a = 1$, $b = 10$, $c = 15$, $d = 10$, showing how the raindrop shape is obtained. The log conditional likelihood (solid curve) has been graphed with its maximum (indicated by the dotted vertical line) equal to 0. A drop in the log likelihood of 1.92 (indicated by the horizontal line) is significant at an approximate 95% level. The corresponding approximate 95% confidence interval for the log odds ratio is $(-5.65, -0.85)$. By reflecting the part of the curve above -1.92 about the horizontal line, we obtain a symmetric region (shaded in the figure). The height of the region at a particular value of the log odds ratio relative to the maximum height gauges the relative plausibility of that value. The vertical dashed lines on the left and right hand sides are the lower and upper 95% confidence limits based on the normal approximation and the standard error given by (2.5).

The reason for reflecting the curve and shading the resulting region is to produce a visually appealing symmetry and facilitate comparisons among several raindrops. An alternative would be simply to use the top half of the raindrop, however this would result in vertical asymmetry, which is of no interest. Instead, the raindrop plot lets the viewer concentrate on any horizontal asymmetry which may be present. Similar approaches have been used by Lee and Tu (1997) and Hintze and Nelson (1998). In many cases, the likelihood is asymptotically normal. A normal likelihood has a quadratic log likelihood and in this case the raindrop has a reflected parabola shape.

In the example above, it could be argued that a point estimate with non-symmetric error bars would suffice. However, consider the data set $a = 3$, $b = 9$, $c = 0$, $d = 12$ (Figure 2.10b). The zero cell results in a “ramped” likelihood, i.e. as θ_i decreases, the likelihood increases monotonically. There is thus no unique maximum likelihood. The normal approximation (the point estimate and error bars in Fig. 2.10a) is woefully inadequate. For example, the 95% confidence interval based on the normal approximation contains 0, whereas the raindrop does not come close to 0. For 2×2 tables this is not a great surprise, and any investigator would be wary of zero cells. However, in more complex situations, such as the nonlinear models discussed later in this paper, ramped likelihoods are not always apparent in conventional analyses.

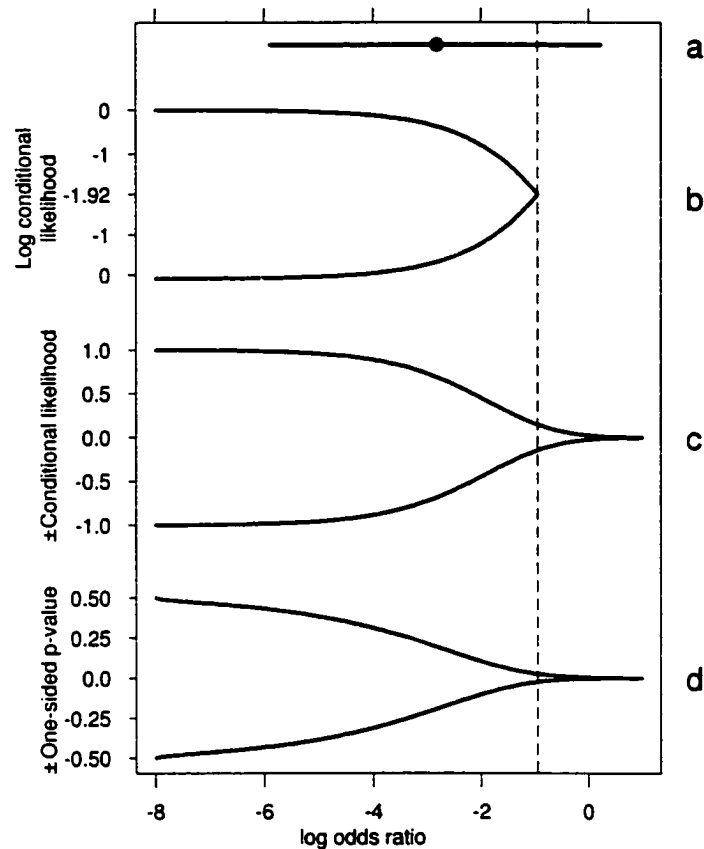


Figure 2.10: Four different ways of displaying information on the log odds ratio based on data $a = 3$, $b = 9$, $c = 0$, $d = 12$. (a) Point estimate with approximate 95% confidence interval based on the normal approximation and the standard error given by (2.5). (b) Raindrop plot. The dashed vertical line shows the upper limit of the associated likelihood-based confidence interval. (c) Raindrop-type plot using the conditional likelihood. (d) Raindrop-type plot using the p -value function. For display purposes it might be best to truncate (c) and (d) at the dashed vertical line.

Figure 2.10 also shows two alternative raindrop-type plots. Figure 2.10c is based on the conditional likelihood, rather than the log conditional likelihood. Figure 2.10d is based on the “confidence curve” (Birnbaum 1961), which depicts the $100(1 - \alpha)\%$ confidence intervals for all $\alpha \in [0, 1)$. In the present context, a $100(1 - \alpha)\%$ confidence interval for θ_i

is given by

$$\{\theta : \ell_i(\theta) \geq -\frac{1}{2}\chi_1^2(1-\alpha)\}.$$

Because of the correspondence between hypothesis testing and interval estimation, the confidence curve is also known as the “ p -value function” (Miettinen 1985): the one-sided p -value for testing the hypothesis $H_0 : \theta_i = \theta$ is given by

$$p = \frac{1}{2}[1 - C_1(-2\ell_i(\theta))],$$

where $C_1(\cdot)$ is the cdf for a chi-square random variable with 1 d.f. The p -value function has infinite extent and so we might consider truncating it at $p = 0.025$ (i.e. at the 95% confidence level). By reflecting the p -value function about $p = 0$, we obtain a figure comparable to the raindrop plot (Fig. 2.10d). While Figure 2.10c and Figure 2.10d show the curves for all parameter values, in practice it is very hard to distinguish small differences in heights, particularly when many such plots are displayed. Also, the interpretation will depend upon the widths of the lines plotted.

There are cases when one might wish to use any one of the above three types of plots. Here we concentrate on the raindrop plot because of its direct confidence-interval interpretation, because its simple reflected-parabola shape under normality facilitates detection of non-normality.

2.6.1 Construction of the raindrop plot

Modern graphical data analysis environments, such as S-PLUS, make producing such figures very easy. For example, if `theta` and `l` are two vectors containing values of θ_i and $\ell_i(\theta_i)$ respectively, the following S-PLUS commands produce a 95% raindrop:

```
cutoff <- -1.92
select <- l > cutoff
THETA <- theta[select]
L <- l[select]
thickness <- 1-L/cutoff
plot(0,xlim=range(theta),ylim=c(-1,1),type="n",xlab="theta",ylab="")
polygon(c(THETA, rev(THETA)),c(-thickness, rev(thickness)))
```

2.6.2 Example: Ulcer studies

Figure 2.11 uses raindrops to display the log conditional likelihoods for the log odds ratio in the ulcer studies.

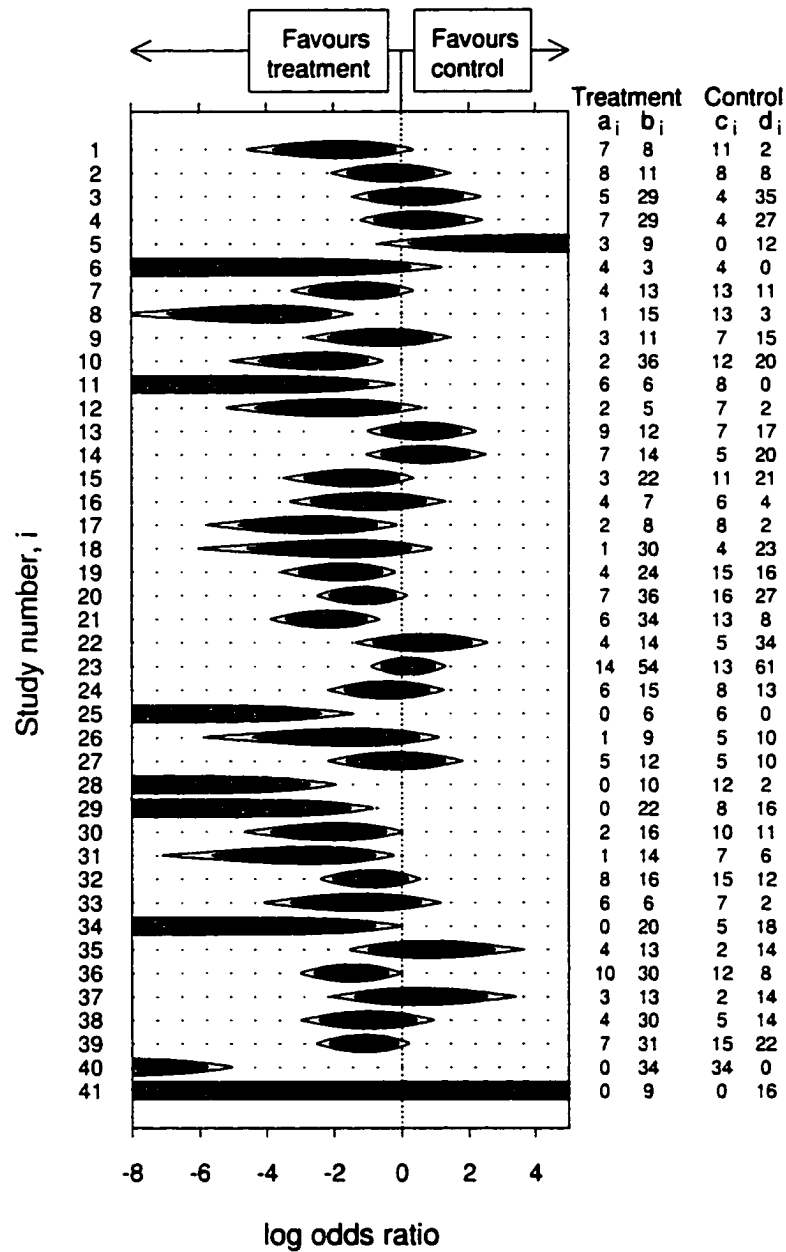


Figure 2.11: Raindrop plot of log odds ratios for studies of ulcer treatments. For each study, 95% raindrops (shaded) are superimposed on 99% raindrops (unshaded).

Note that nine of the studies have zeros in at least one cell. The resulting likelihoods do not have peaks: except for study number 41 (which has a zero marginal total and hence

a completely flat likelihood), the likelihoods are ramped. Such likelihoods are sometimes excluded from analyses, yet they clearly contain information. However, the point estimates and confidence intervals from (2.4) and the normal approximation using (2.5) are of dubious quality. The raindrop plot allows the meta-analyst to display the information provided by these types of data sets, without resorting to questionable approximations.

2.6.3 Example: Coho salmon Beverton-Holt model

We now consider a more complex example. Estimates of the parameters of population dynamics models like the Beverton-Holt are of central importance for fisheries management, estimation of extinction rates, and predictions concerning the recovery of over-exploited populations. Whereas in the 2×2 table case, an explicit form was available for the conditional likelihood for the log odds ratio, in this case no such convenience is available. Instead, we use profile likelihoods in constructing the raindrops.

We typically assume that recruitment is lognormally distributed. Under this assumption, the spawner-recruitment models we will consider can be expressed in the form

$$R_{ij} = \alpha_i f_\varphi(S_{ij}) e^{\varepsilon_{ij}}, \quad j = 1, \dots, n_i, \quad (2.13)$$

where the ε_{ij} ($j = 1, \dots, n_i$) are i.i.d. normal and $f_\varphi(S)$ is a function of S and additional parameters φ . The logarithm of (2.13) is

$$\log R_{ij} = \log \alpha_i + \log f_\varphi(S_{ij}) + \varepsilon_{ij},$$

and maximum likelihood estimation for this model is identical to least squares, i.e., the objective function is

$$\sum_{j=1}^{n_i} [\log R_{ij} - \log \alpha_i - \log f_\varphi(S_{ij})]^2.$$

Because $\log \alpha_i$ enters linearly, it is easily profiled out of the likelihood. For a given trial estimate of φ , the least squares estimate of $\log \alpha_i$ is

$$\widehat{\log \alpha_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} \log \frac{R_{ij}}{f_\varphi(S_{ij})}.$$

The profile log likelihood for φ is given by

$$-\frac{n}{2} \log \sum_{j=1}^{n_i} \left[\log R_{ij} - \widehat{\log \alpha_i} - \log f_{\varphi}(S_{ij}) \right]^2.$$

The profile log likelihood for α_i must be obtained numerically. Figure 2.12 shows side-by-side raindrop plots for α_i and R_i^{\max} based on profile likelihoods.

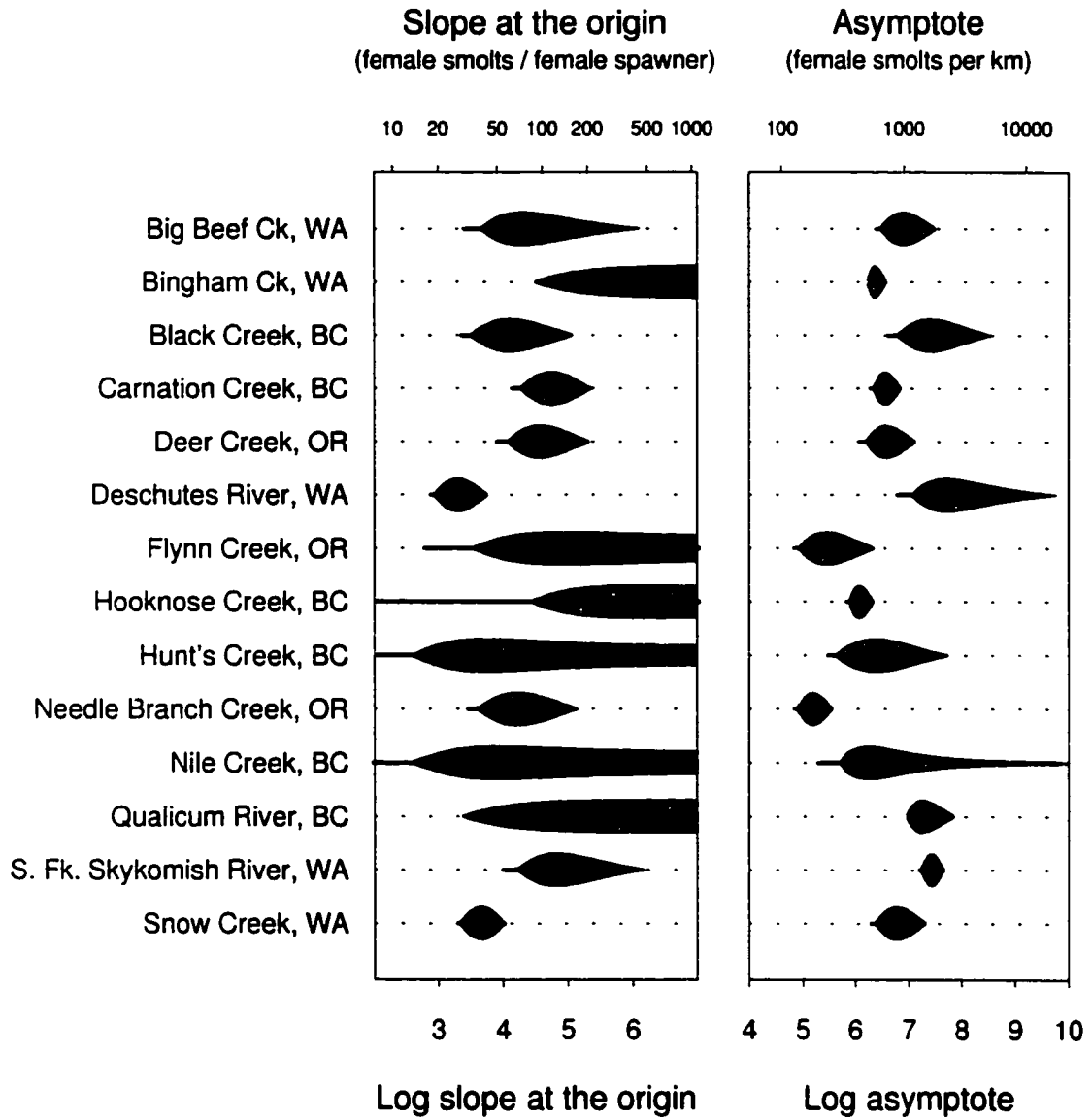


Figure 2.12: Side-by-side 95% raindrop plots for the parameters of the Beverton-Holt model for the 14 coho salmon populations. The left panel shows raindrops for the slope at the origin α_i , while the right panel shows raindrops for the asymptotic level of recruitment, R_i^{\max} . The superimposed dots and error bars on the individual population raindrops are the maximum likelihood estimates obtained by nonlinear regression and approximate asymptotic 95% confidence intervals (based on nonlinear least squares theory).

Note that for two populations convergence of the nonlinear least squares algorithm was not obtained because of “ramping” behavior in the likelihood surface. It appears that the asymptotics on which the least squares estimates and standard errors are based are often poor in that the asymptotic confidence interval often does not match the profile-likelihood-based interval well.

Figure 2.12 clearly shows the difficulties of estimation for these data sets. For example, the Bingham Creek data determine the asymptote well, however the slope at the origin is poorly determined and the nonlinear regression algorithm did not converge in this case. In the case of the Deschutes River data, estimates and standard errors are available from nonlinear regression, but the asymptote is poorly determined, and its standard error is of dubious quality. Both the slope at the origin and the asymptote are poorly determined for the Nile Creek data; the 95% likelihood intervals both have lower limits, but no upper limits. However the raindrops reveal an interesting difference in the shapes of the two profile likelihoods.

Discussion

Raindrop plots are a compact and informative way of displaying the information provided by groups of studies.

An alternative is to superimpose graphs of the individual likelihoods (e.g. (Efron 1996; van Houwelingen and Zwinderman 1993)). However, judgments involving superimposed curves are difficult (Cleveland 1985, p. 271) and if there are more than, say, 5 studies involved, the result is hard to decipher. Furthermore, individual confidence intervals are not easily detected.

Another alternative is simply to display the point estimates with error bars representing a likelihood-based confidence interval. This is adequate if the log likelihood is close to being quadratic, but fails completely in many real examples, e.g. Figures 2.10 and 2.12. In such cases, a second dimension is required: raindrop plots effectively communicate the relative plausibility of different parameter values.

The raindrop plot is based on reflecting the log likelihood, however alternative approaches are possible, using e.g., the likelihood or the p -value function. We argue that the raindrop’s simple shape under normality—a reflected parabola—facilitates detection of non-normality.

With small sample sizes and nonlinear models, likelihoods may exhibit asymmetry and “ramping” behavior that complicate plotting. In such cases raindrop plots may provide a useful tool.

Bates and Watts (1988) suggest several related profiling techniques for parameters of nonlinear regression models, which are useful for individual data sets. It should also be noted that in some problems likelihood surfaces are not smooth and may have multiple local maxima. An example of this is for the hockey stick spawner-recruitment model.

2.6.4 Example: Coho salmon hockey-stick models

Figure 2.13 shows the hockey-stick model raindrop for the slope at the origin, α , for the coho salmon population in Flynn Creek, Oregon.

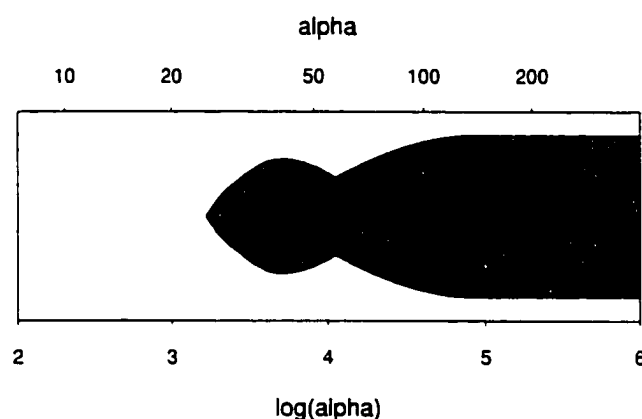


Figure 2.13: Raindrop for the slope at the origin for the coho salmon population in Flynn Creek, Oregon.

There is a local maximum in the profile likelihood near $\log \alpha = 3.7$, but the profile likelihood is globally maximized for larger values of $\log \alpha$ (in fact, the profile likelihood is constant for values of $\log \alpha$ greater than about 5). For likelihoods with multiple maxima, raindrops may consist of more than one piece. In other words the likelihood ratio procedure may give confidence *sets* rather than confidence *intervals*.

The raindrop plot illustrates how badly-behaved segmented regression models like the hockey stick can be: not only is there a local maximum in the profile likelihood, there is also a completely flat region in the likelihood surface.

Because of the irregular likelihood surfaces that the hockey-stick model can give, Barrowman and Myers (2000) investigated ways to generalize the hockey-stick model. The goal was to find a spawner-recruitment function that gives the hockey-stick model as one limiting case and behaviour similar to the Beverton-Holt model as another limiting case. In fact, two *generalized hockey stick* models were developed.

The quadratic hockey stick

A simple approach is to retain the hockey-stick model except in a region close to S^* , and allow a smooth transition between the two parts of the hockey stick. We define this region of transition to be between $S^* - \omega$ and $S^* + \omega$, where $0 \leq \omega \leq S^*$. It would be reasonable to require a curve with continuous first derivatives at all points. The piecewise polynomial approach (Tishler and Zang 1981) can produce a curve with the desired properties. The resulting equation is

$$R = \begin{cases} \alpha S & \text{if } S \leq S^* - \omega \\ \alpha \left(S - \frac{(S - S^* + \omega)^2}{4\omega} \right) & \text{if } S^* - \omega < S < S^* + \omega \\ \alpha S^* & \text{if } S \geq S^* + \omega. \end{cases}$$

We have found it convenient to reparametrize the above equation in terms of a smoothness parameter $\delta = \omega/S^*$, so that $0 \leq \delta \leq 1$. The above equations become

$$R = \begin{cases} \alpha S & \text{if } S \leq S^*(1 - \delta) \\ \alpha \left(S - \frac{(S - S^*(1 - \delta))^2}{4\delta S^*} \right) & \text{if } S^*(1 - \delta) < S < S^*(1 + \delta) \\ \alpha S^* & \text{if } S \geq S^*(1 + \delta), \end{cases} \quad (2.14)$$

which we call the *quadratic hockey stick*. The simplest way to think about the quadratic hockey stick is the two pieces of a hockey stick connected by a parabolic curve, with all the right attributes (continuity and continuity of the first derivative), at the ends of the parabolic curve. Note, however, that even with $\delta = 1$, the quadratic hockey stick does not reproduce the Beverton-Holt.

Figure 2.14 illustrates the quadratic hockey stick and another generalized hockey-stick model, the *logistic hockey stick*.

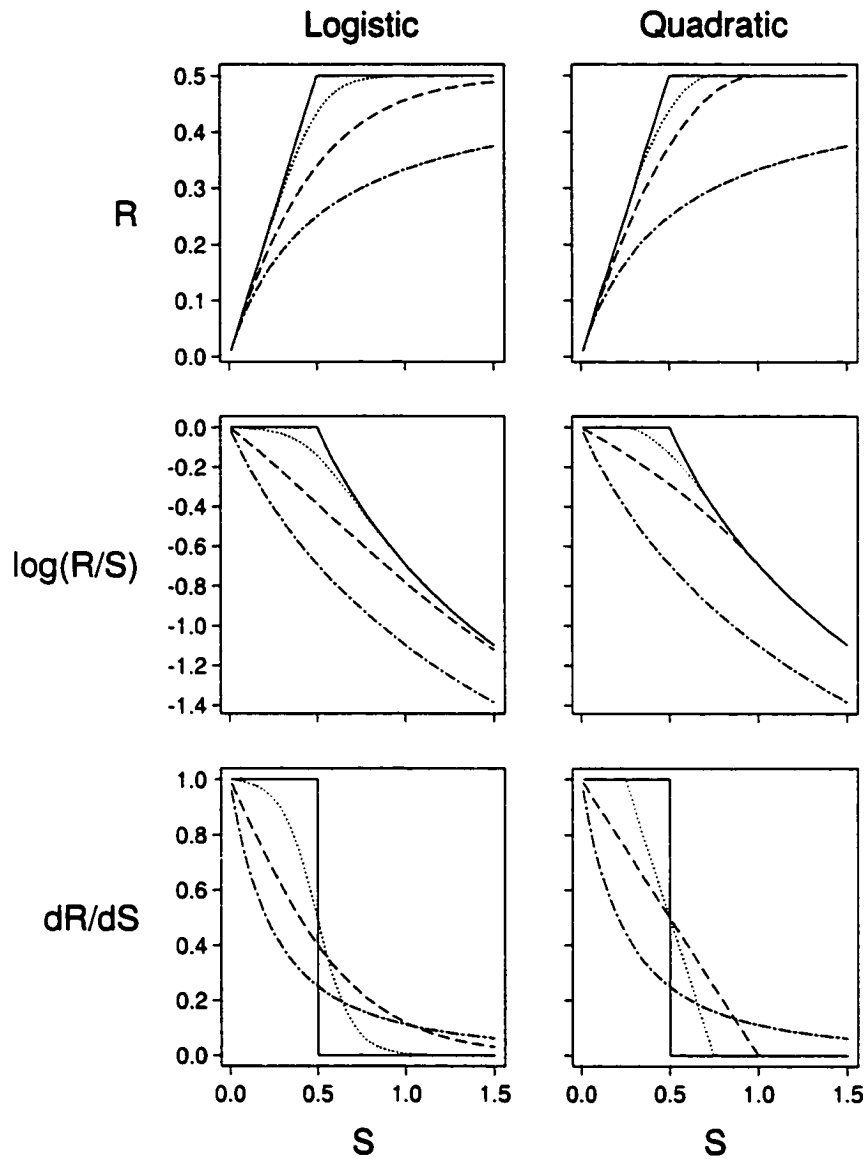


Figure 2.14: Generalized hockey-stick models (solid, dotted, and dashed lines) compared with a Beverton-Holt model (dot-dash line). The left-hand column shows logistic hockey sticks with $\delta = 0$ (solid line), 0.2 (dotted line) and 100 (dashed line). The right-hand column shows quadratic hockey sticks with $\gamma = 0$ (solid line), 0.5 (dotted line), and 1 (dashed line). Each column shows recruits versus spawners (top panel), log survival versus spawners (middle panel), and the derivative of recruitment versus spawners (bottom panel).

The Logistic Hockey Stick

An alternative approach is to define the spawner-recruitment model in terms of the derivative of recruitment with respect to spawner abundance. For the hockey stick we have

$$\frac{dR}{dS} = \begin{cases} \alpha & \text{if } S < S^* \\ 0 & \text{if } S \geq S^*. \end{cases} \quad (2.15)$$

We need a generalization of the above model such that the slope at the origin, α , remains comparable in the generalization, and it includes the above model as a special case.

Equation (2.15) is easily expressed in terms of the limiting case of any of several standard cumulative distribution functions, e.g. the normal or the logistic. We will use the logistic because it has a simple analytic form, and will call the model the “logistic hockey stick.” A preliminary version of our generalization is defined by

$$\frac{dR}{dS} = \alpha \frac{1}{1 + \exp\{(S - \mu)/(\gamma\mu)\}}, \quad (2.16)$$

where μ is the inflection point of spawner abundance and the product $\gamma\mu$ is the scale parameter of the logistic, an analog of the standard deviation parameter in the normal. We parameterize the model in terms of the smoothness parameter γ , an analog of the coefficient of variation of the normal, because it provides a more appropriate tuning parameter. It is easy to see that as $\gamma \rightarrow 0$, the hockey-stick model is recovered.

This model has the right behaviour since dR/dS is monotonically decreasing and as $S \rightarrow \infty$, $dR/dS \rightarrow 0$. However, $\lim_{S \downarrow 0} dR/dS \neq \alpha$ except in the limit as $\mu \downarrow 0$. To ensure that $\lim_{S \downarrow 0} dR/dS = \alpha$ we multiply by $1 + e^{-1/\gamma}$ and (2.16) becomes

$$\frac{dR}{dS} = \alpha \frac{1 + e^{-1/\gamma}}{1 + \exp\{(S - \mu)/(\gamma\mu)\}}.$$

Integrating this expression with respect to S , and setting the integration constant so that $R = 0$ when $S = 0$ gives

$$R = \alpha\gamma\mu(1 + e^{-1/\gamma}) \left(\frac{S}{\gamma\mu} - \log \left(\frac{1 + e^{(S-\mu)/(\gamma\mu)}}{1 + e^{-1/\gamma}} \right) \right). \quad (2.17)$$

Note that an application of l'Hôpital's rule gives

$$\lim_{S \rightarrow \infty} R = \alpha\gamma\mu(1 + e^{-1/\gamma}) \left(\frac{1}{\gamma} + \log(1 + e^{-1/\gamma}) \right).$$

Note, however, that like the quadratic hockey stick, the logistic hockey stick cannot reproduce the behaviour of the Beverton-Holt.

Chapter 3

Hierarchical models for meta-analysis

The models considered in this work fall into the category of *conditionally independent hierarchical models* (CIHMs) (Kass and Steffey 1989). These are two-stage models of the kind used in parametric empirical Bayes methodology. The CIHM framework, introduced in Section 3.1, allows us to see the connections between several different approaches to estimation and modeling. In Sections 3.2 and 3.3, parametric empirical Bayes and hierarchical Bayes approaches are placed in this framework, following Efron (1996). It is well known that in some situations, the parametric empirical Bayes and hierarchical Bayes solutions may give similar results; this is discussed in Section 3.4. Recently, Efron (1996) proposed an empirical Bayes approach for combining likelihoods, which he illustrated with the ulcer data. This is discussed in Section 3.5 and the raindrop plots of Section 2.6 are extended to display results from Efron's methods. Efron's method is also applied to profile likelihoods for spawner-recruitment model parameters and a simulation conducted to examine the performance of the methods in this case.

3.1 Conditionally independent hierarchical models

In what follows, I will use the generic notation $p(\cdot|\cdot)$ to indicate the probability density of the first argument given the second. The resulting loss of precision is more than made up for in simplicity of notation.

In general, a two-stage hierarchical model for a random vector y specifies a density for y conditional on parameters θ and λ , $p(y|\theta, \lambda)$, and a density for θ conditional on λ , $p(\theta|\lambda)$.

Note that θ and λ may each be vector valued.

Here we suppose that y is composed of observation vectors y_1, \dots, y_m collected from m studies, i.e., $y = (y_1^T, \dots, y_m^T)^T$. Because there is usually variability both within and between studies, a two-stage hierarchical model is natural. Associated with study i is an unobservable *study-specific* parameter θ_i , and we define $\theta = (\theta_1^T, \dots, \theta_m^T)^T$. The parameter λ is taken to be associated with all studies, and may be called the *common parameter*, although for reasons to be discussed later, we often call it the *hyperparameter*.

Now suppose that we observe data y_0 from an additional study and we would like to make inference about its study-specific parameter θ_0 . We call y_0 the *direct data* that contains information about θ_0 . However, because of the link provided by λ , the *supplementary data*, y , also contains information about θ_0 .

The complete formulation of a CIHM is given below:

Conditionally Independent Hierarchical Model

Stage 1 Conditional on $(\theta_0, \theta_1, \dots, \theta_m)$ and λ , the vectors y_i are independent with densities $p(y_i|\theta_i, \lambda)$, $i = 1, \dots, m$. Note that the densities may have different forms for different studies.

Stage 2 Conditional on λ , the vectors θ_i are i.i.d. with densities $p(\theta_i|\lambda)$, $i = 1, \dots, m$.

From a Bayesian viewpoint, the second-stage densities $p(\theta_i|\lambda)$ are known as prior distributions, whereas from a mixed model viewpoint they are known as *random effect distributions*.

Results are simplified when the sampling densities do not depend on λ , i.e. when

$$p(y_i|\theta_i, \lambda) = p(y_i|\theta_i). \quad (3.1)$$

For example, for the ulcer data, the conditional likelihood $L_i(\theta_i)$ of (2.12) does not depend on any common parameters. In the notation used here, $L_i(\theta_i) = p(y_i|\theta_i)$ taken as a function of θ_i rather than y_i .

From the assumed conditional independence structure,

$$p(y|\theta, \lambda) = \prod_{i=1}^m p(y_i|\theta_i, \lambda),$$

and

$$p(\theta|\lambda) = \prod_{i=1}^m p(\theta_i|\lambda),$$

so that the joint distribution of the data and the study-specific parameters is given by

$$p(y, \theta|\lambda) = p(y|\theta, \lambda)p(\theta|\lambda). \quad (3.2)$$

Therefore, the marginal density of the data is given by

$$\begin{aligned} p(y|\lambda) &= \int p(y|\theta, \lambda)p(\theta|\lambda)d\theta \\ &= \int \left[\prod_{i=1}^m p(y_i|\theta_i, \lambda) \right] \left[\prod_{i=1}^m p(\theta_i|\lambda) \right] d\theta_1 \cdots d\theta_m \\ &= \int \prod_{i=1}^m p(y_i|\theta_i, \lambda)p(\theta_i|\lambda)d\theta_i \\ &= \prod_{i=1}^m \int p(y_i|\theta_i, \lambda)p(\theta_i|\lambda)d\theta_i \\ &= \prod_{i=1}^m p(y_i|\lambda), \end{aligned} \quad (3.3)$$

which also shows that the y_i 's are marginally independent given λ .

3.2 Parametric empirical Bayes inference

In parametric empirical Bayes inference, we treat $p(\theta|\lambda)$ as a prior distribution for θ with a fixed but unknown parameter λ . The observed marginal distribution of the data provides information that can be used to obtain an estimate of λ . For example, the MLE, $\hat{\lambda}$, of λ is obtained by maximizing the marginal likelihood $p(y|\lambda)$ (treated as a function of λ rather than y):

$$\hat{\lambda} = \operatorname{argmax}_{\lambda} p(y|\lambda).$$

Consider the posterior density for θ_0 based on y_0 conditional on λ , $p(\theta_0|y_0, \lambda)$. By Bayes' rule,

$$p(\theta_0|y_0, \lambda) \propto p(y_0|\theta_0, \lambda)p(\theta_0|\lambda).$$

Substituting $\hat{\lambda}$ for λ gives what Efron (1996) terms the *MLE posterior* for θ_0 :

$$p(\theta_0|y_0, \hat{\lambda}) \propto p(y_0|\theta_0, \hat{\lambda})p(\theta_0|\hat{\lambda}).$$

The mean or mode of the MLE posterior can then be used as an empirical Bayes point estimate of θ_0 . Efron (1996) calls $p(\theta_0|\hat{\lambda})$ the *MLE prior* for θ_0 . In the special case that the sampling distributions are independent of λ (equation (3.1)), we have

$$p(\theta_0|y_0, \hat{\lambda}) \propto p(y_0|\theta_0)p(\theta_0|\hat{\lambda}). \quad (3.4)$$

Changing notation slightly shows that this is Bayes' rule using the MLE prior as if it were a true prior:

$$p_{\hat{\lambda}}(\theta_0|y_0) \propto p(y_0|\theta_0)p_{\hat{\lambda}}(\theta_0).$$

The parametric empirical Bayes strategy is thus to substitute an estimate of the parameters of the prior into the usual Bayesian calculations. The principal disadvantage of this approach is that it ignores our uncertainty about λ . A fully Bayes approach, discussed in the next section, allows for this uncertainty, but with the cost of specifying a prior distribution for λ .

One fine point should be noted. In our exposition, the estimate of the common parameter, λ , was based on the supplementary data $y = (y_1^T, \dots, y_m^T)^T$. As Efron (1996) points out, it would be reasonable also to include the direct data y_0 in the estimation of λ . Formally excluding y_0 makes the connection between empirical Bayes and hierarchical Bayes somewhat more direct, although in practice y_0 would typically be included.

3.3 Hierarchical Bayes inference

Since λ is unknown, the Bayesian approach is to place a prior on λ . Since this is a prior on the parameters of a prior, it is sometimes called a *hyperprior*. Then the posterior for θ_0

based on all of the data is

$$\begin{aligned} p(\theta_0|y_0, y) &= \int p(\theta_0, \lambda|y_0, y) d\lambda \\ &= \int p(\theta_0|\lambda, y_0, y) p(\lambda|y_0, y) d\lambda. \end{aligned}$$

But conditional on λ and y_0 , θ_0 is independent of y (Deely and Lindley 1981)¹. Thus, we have

$$p(\theta_0|y_0, y) = \int p(\theta_0|\lambda, y_0) p(\lambda|y_0, y) d\lambda. \quad (3.5)$$

Note that by Bayes' rule,

$$p(\theta_0|\lambda, y_0) \propto p(y_0|\theta_0, \lambda) p(\theta_0|\lambda).$$

In the special case that the sampling distributions are independent of λ , we have $p(y_0|\theta_0, \lambda) = p(y_0|\theta_0)$, so that (3.5) becomes:

$$p(\theta_0|y_0, y) \propto p(y_0|\theta_0) \int p(\theta_0|\lambda) p(\lambda|y_0, y) d\lambda.$$

Note that $p(\lambda|y_0, y) \propto p(\lambda|y)$, so we can write this as

$$p(\theta_0|y_0, y) \propto p(y_0|\theta_0) \int p(\theta_0|\lambda) p(\lambda|y) d\lambda. \quad (3.6)$$

Also, as noted earlier, given λ , θ_0 is independent of y , so that $p(\theta_0|\lambda) = p(\theta_0|\lambda, y)$, and so we can write

$$\begin{aligned} p(\theta_0|y_0, y) &\propto p(y_0|\theta_0) \int p(\theta_0|\lambda, y) p(\lambda|y) d\lambda \\ &= p(y_0|\theta_0) p(\theta_0|y), \end{aligned} \quad (3.7)$$

¹To see this, note that by Bayes' rule,

$$p(\theta_0|\lambda, y_0, y) \propto p(y_0|\theta_0, \lambda, y) p(\theta_0|\lambda, y).$$

Neither term on the right hand side depends on y since, by the assumed conditional independence structure, given θ_0 and λ , y_0 is independent of y , and given λ , θ_0 is independent of y .

where $p(\theta_0|y)$ is the predictive distribution for θ_0 , termed the *induced prior* by Efron (1996). As Deely and Lindley point out, (3.7) neatly separates the role of the direct and supplementary data in inference about θ_0 . Figure 3.1, adapted from Efron (1996), shows the structure of the hierarchical Bayes model.

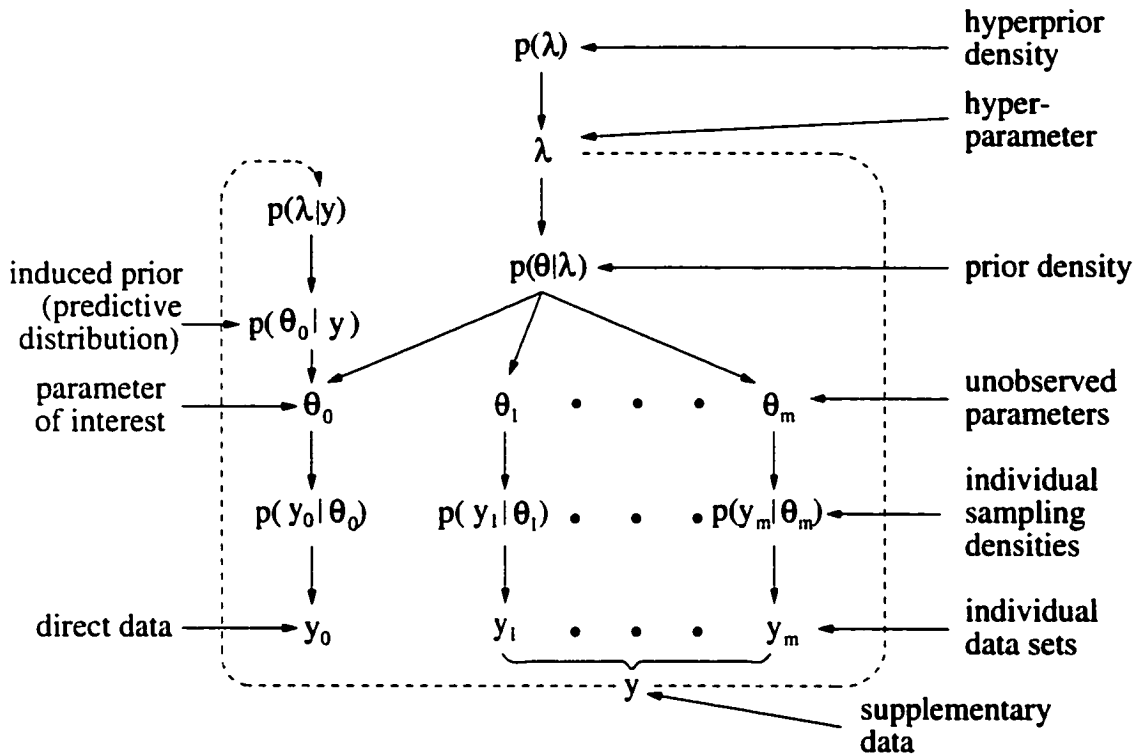


Figure 3.1: The hierarchical Bayes model, following Efron (1996).

3.4 Comparing empirical Bayes and hierarchical Bayes

For simplicity, in this section we will assume that the sampling distributions are independent of λ (equation (3.1)). Comparing (3.4) and (3.7) shows the difference between the empirical Bayes and hierarchical Bayes solutions. In particular, the empirical Bayes solution substitutes the MLE prior $p(\theta_0|\hat{\lambda})$ for the induced prior $p(\theta_0|y)$. From (3.6) and (3.7),

we see that the induced prior is given by

$$p(\theta_0|y) = \int p(\theta_0|\lambda)p(\lambda|y)d\lambda.$$

The two strategies are thus to maximize out λ or to integrate out λ , which is reminiscent of the choice in Chapter 2 between integrated likelihood and profile likelihood for the elimination of nuisance parameters.

Applying Bayes' rule to $p(\lambda|y)$ gives

$$p(\lambda|y) \propto p(y|\lambda)p(\lambda).$$

If $p(\lambda)$ is relatively uninformative, then the posterior for the hyperparameter $p(\lambda|y)$ is approximately equal to the marginal likelihood for the hyperparameter $p(y|\hat{\lambda})$. In addition, if $p(y|\lambda)$ is relatively sharply peaked, then the induced prior $p(\theta_0|y)$ is approximately equal to the MLE prior $p(\theta_0|\hat{\lambda})$ (Smith 1983). These relationships are depicted in Figure 3.2.

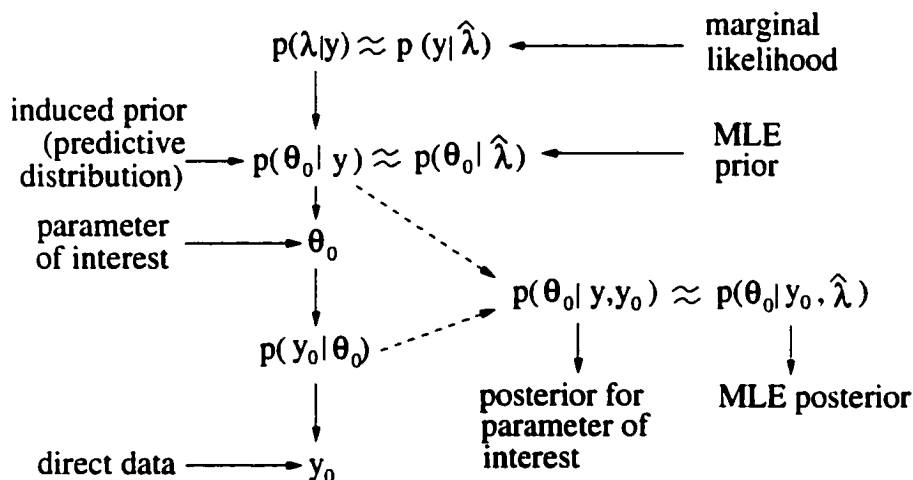


Figure 3.2: Connection between empirical Bayes and hierarchical Bayes. Compare with left hand side of Figure 3.1.

Suppose the mean of the MLE posterior, $E(\theta_0|y_0, \hat{\lambda})$, is used as an empirical Bayes point estimate. Denote the expectation and variance with respect to $p(\lambda|y)$ by $E_{\lambda|y}$ and $\text{Var}_{\lambda|y}$. Kass and Steffey (1989) show that $E_{\lambda|y} [E(\theta_0|y_0, \hat{\lambda})] \approx E(\theta_0|y)$, so that the mean

of the MLE posterior is an approximate fully Bayes posterior mean. However the variance of the MLE posterior generally underestimates the fully Bayes posterior variance. To see this, note that

$$\text{Var}_{\lambda|y}(\theta_0|y) = E_{\lambda|y}[\text{Var}(\theta_0|y_0, \lambda)] + \text{Var}_{\lambda|y}[E(\theta_0|y_0, \lambda)].$$

The variance of the MLE posterior, $\text{Var}(\theta_0|y_0, \hat{\lambda})$, approximates only the first term above and the second term is non-negative. Several methods for inflating estimates of the uncertainty of empirical Bayes estimates have therefore been proposed, e.g., (Laird and Louis 1987; Carlin and Gelfand 1990). This is also described as empirical Bayes bias correction or calibration, since the goal is to match a fully Bayes analysis based on an uninformative hyperprior $p(\lambda)$. Similar corrections are made for profile likelihoods.

Except for the use of improper flat priors, this thesis does not make use of the fully Bayes approach. This is not because of a lack of appreciation of its virtues. The fully Bayes approach allows for complete specification of prior knowledge instead of feigned ignorance, and consequently allows valid posterior probability statements that accurately reflect all remaining uncertainty. Furthermore, modern *Markov Chain Monte Carlo* (MCMC) methods make Bayesian calculations relatively straightforward.

However, Bayesian methods do have their disadvantages. The choice of informative prior distributions may be controversial. Noninformative priors would seem to avoid this difficulty, but in some situations it is not clear what constitutes a noninformative prior. For example, noninformative priors for variances remain a topic of current research; we shall return to this point in Chapter 4. MCMC methods may be computationally intensive and assessment of convergence is a topic of ongoing research. Another point of concern is that for improper priors the posterior may not be proper and this may not be detected when using MCMC (Hobert and Casella 1996).

In any case, as Breslow and Clayton (1993) suggest,

“There is still room for simple, approximate methods both for exploratory analyses and to provide starting values for use with other, more exact procedures.”

3.5 Empirical Bayes methods for combining likelihoods

We now consider an interesting illustration of the hierarchical modeling approach introduced in this chapter. Efron (1996) proposed an empirical Bayes methodology for “the practical solution of hierarchical Bayes problems”. In this section, we describe Efron’s approach and display some of his results using an extension of the raindrop plots of Section 2.6. We then consider the application of his approach to the meta-analysis of the coho salmon data.

Efron began by reducing the individual data sets to approximate single-parameter likelihoods for the study-specific parameters. Section 2.5 reviewed three methods for eliminating nuisance parameters which can be used to obtain single-parameter likelihoods; Efron (1996) gives references for several others. With single-parameter likelihoods, the hierarchical model calculations introduced in this chapter are relatively straightforward. Often, hierarchical models use analytically tractable probability distributions. For example, traditional mixed models use normal random effects and normal errors, giving the so-called *normal-normal* hierarchical model. Another example is the beta-binomial model, often used to describe observed proportions that exhibit overdispersion. In order to flexibly model the prior densities $p(\theta_i|\lambda)$, Efron used specially designed exponential families of distributions and carried out the calculations using numerical integrals. Efron also proposed schemes for preventing overshrinkage and for empirical Bayes bias correction, although we shall not examine them here.

3.5.1 Example: Ulcer studies

The ulcer data were used by Efron to illustrate his approach. The conditional likelihood $L_i(\theta)$ of (2.12) provides a single-parameter likelihood. Efron only considered $\theta \in (-8, 5)$ because most of the 41 likelihood functions were concentrated in that range. The exponential family of prior densities used by Efron has the form

$$p(\theta|\lambda) = e^{\lambda^T t(\theta) - \phi(\lambda)} g_0(\theta),$$

where $t(\theta)$ is the *sufficient vector*, depending on θ —e.g., $t(\theta) = (\theta, \theta^2)$ — $g_0(\theta)$ is the *carrier density*, and $\phi(\lambda)$ is chosen so that the prior density integrates to 1. For the ulcer data,

Efron used the average normalized likelihood for the carrier density, i.e.,

$$g_0(\theta) = \frac{1}{m} \sum_{i=1}^m L_i(\theta) / \int_{-8}^5 L_i(\theta) d\theta.$$

Using the exponential family of priors based on this carrier density, with a numerical estimation algorithm, Efron obtained an MLE prior $p(\theta|\hat{\lambda})$ for the log odds ratio. (He also obtained a bias-corrected version of the MLE prior, which had slightly heavier tails, but we shall not consider it here.)

A modification of the scheme for producing raindrop shapes can be used to display the MLE prior or any other probability density. In place of the log likelihood, we use the log probability density, with a cutoff corresponding to a probability of 0.95. In other words, we use a raindrop based on the log density over the highest density region, or HDR (Hyndman 1996). We therefore call this an HDR-raindrop, and as a visual cue, use darker shading for HDR-raindrops than for ordinary raindrops. Note that the HDR-raindrop provides more information than the interval or region defined by any single highest density region. Figure 3.3 shows a raindrop plot for the ulcer data, together with meta-analytic summaries from Efron's analysis.

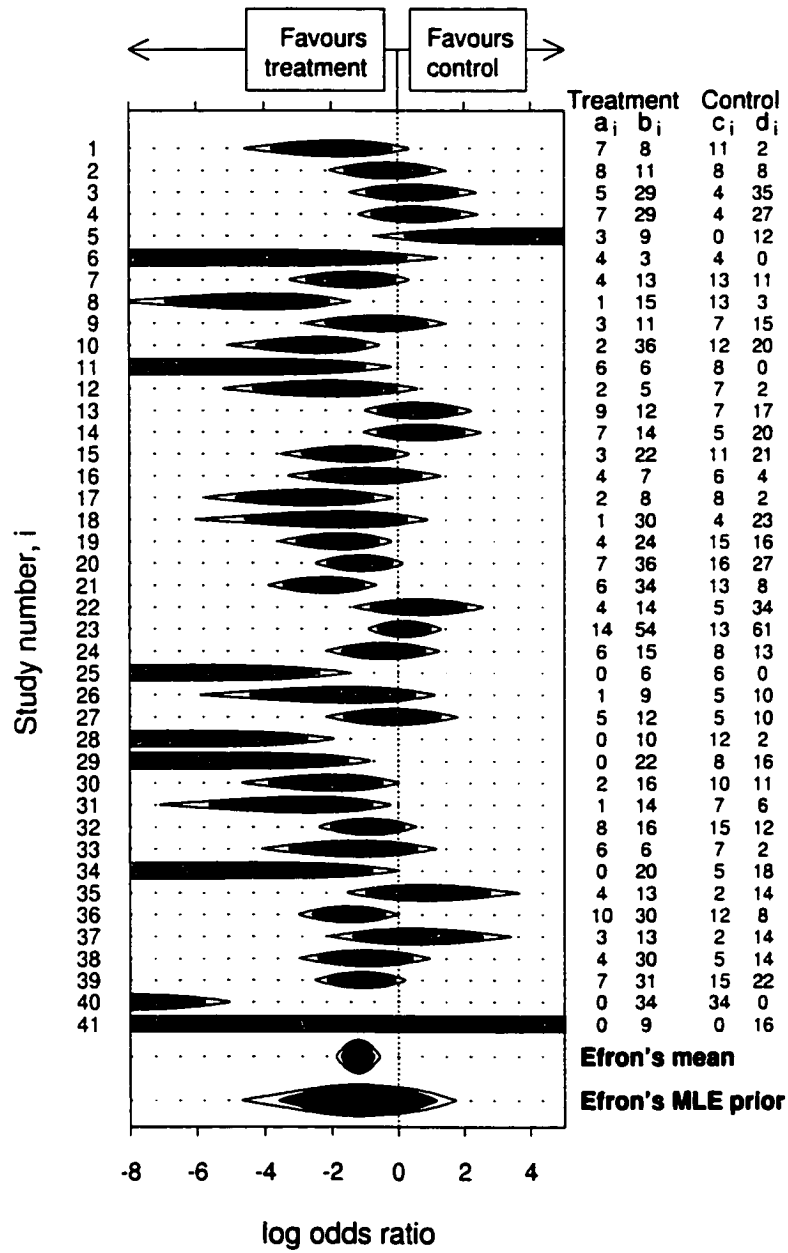


Figure 3.3: Raindrop plot of log odds ratios and meta-analytic summaries for studies of ulcer treatments from the analysis of Efron (1996). For each study, 95% raindrops (shaded) are superimposed on 99% raindrops (unshaded).

Two meta-analytic summaries from Efron (1996) are shown in Figure 3.3: a normal

95% raindrop for the mean (based on a point estimate and jackknife standard error reported by Efron) and a 95% HDR-raindrop for the MLE prior (shaded darker to emphasize that its interpretation is different). The MLE-prior raindrop shows a somewhat subtle deviation from normality. One of the strengths of the raindrop plot is its ability to display such deviations in a compact format. Note that Efron’s analysis excluded study number 40 (which was judged to be an overly influential outlier) and study number 41 (which has a flat likelihood).

From a mixed model viewpoint, the MLE-prior raindrop is our “best” estimate of the random effect distribution for the log odds ratio, i.e., the distribution from which log odds ratios θ_i are generated. This gives very different information from the raindrop labeled “Efron’s mean”, which shows the point estimate of the mean log odds ratio together with a confidence interval, which is conventionally shown in meta-analysis plots. We believe that both pieces of information are important and in the mixed model methods explored in subsequent chapters, we will generally show both.

From an empirical Bayes viewpoint, the MLE-prior raindrop is an approximation to the predictive distribution for the log odds ratio of an as-yet unobserved study. From the discussion in Section 3.4, we expect that, as an approximation to the predictive distribution, the MLE prior does not have enough spread. However, in practice the required correction may not be very large. Corrections to posteriors for individual studies may be substantial, however (Efron 1996).

The summary raindrops at the bottom of Figure 3.3 are reminiscent of the diamond-shaped summary symbols often used in meta-analytic plots. However, unlike the diamonds, the second dimension of the raindrops is informative.

The same approach can be used for Bayesian posterior or predictive distributions. HDR-raindrops are related to the violin plots of Hintze and Nelson (1998), although the violin plot was designed for investigating data samples rather than distributions *per se* and uses the density instead of the log density.

3.5.2 Example: Coho salmon

To study the characteristics of Efron’s approach for this type of data, a simulation was designed. The goal was to simulate data roughly similar to the coho salmon data. In the simulation, groups of 14 spawner-recruitment data sets were generated, with each data set

consisting of 15 spawner-recruitment pairs. To generate the data, a Beverton-Holt model with fixed $K = R^{\max}/\alpha$ ($K = 0.5$) and random $\log\alpha$, sampled from $N(0, 1)$, was used. For each data set a set of fixed spawner values were chosen. Figure 2.12 revealed that 6 of the 14 coho salmon data sets gave rise to “ramped” likelihoods, i.e., they contain relatively little information about α , typically due to a lack of observations at low levels of spawner abundance. To mimic this, our simulated groups of 14 data sets had 6 with spawner observations only at relatively high levels, and the remainder with observations at lower levels. Figure 3.4 shows an example of each type of data set.

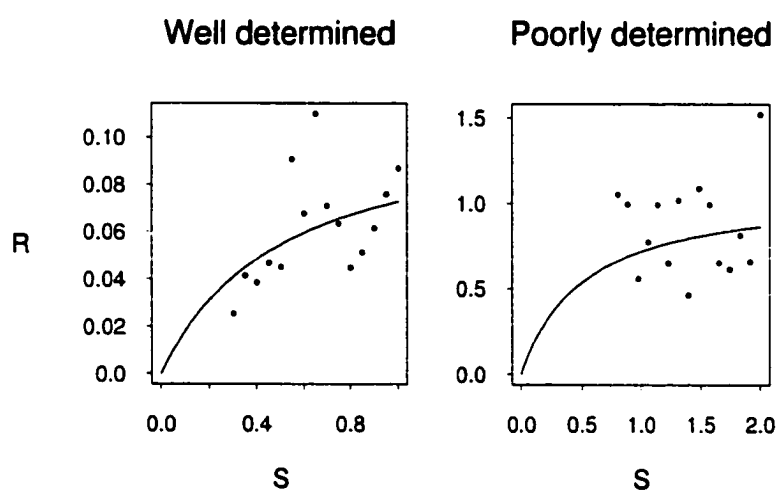


Figure 3.4: Two simulated spawner-recruitment data sets. The superimposed curves are the true median recruitment curves used to generate the data. The left-hand data set determines α fairly well, due to spawner observations at the lower limb of the curve. The right-hand data set determines α quite poorly, due to an absence of spawner observations at the lower limb of the curve.

For each S value, an R value was randomly generated from a normal distribution with mean $\log\alpha - \log S - \log(1 + S/K)$ and standard deviation $\sigma = 0.4$. The value of 0.4 was chosen by trial and error to give data sets whose scatter plots and raindrop plots (Figure 3.5) appeared similar to those of the real data.

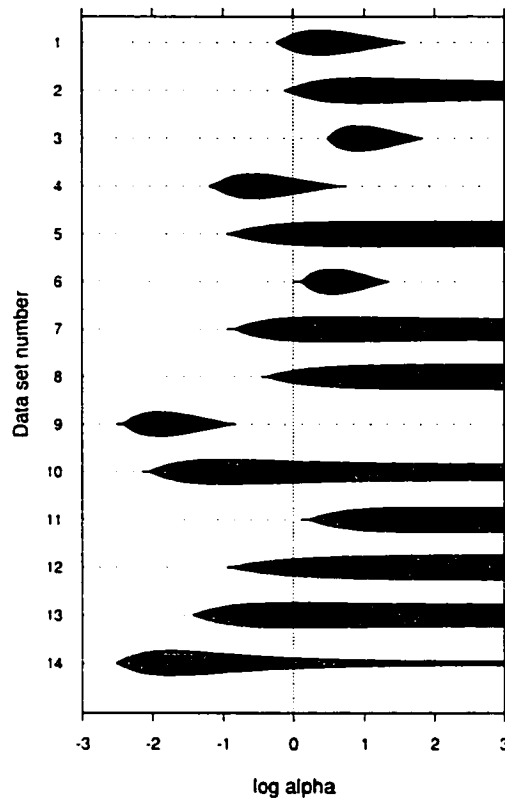


Figure 3.5: Raindrop plots for $\log \alpha$ for one group of simulated data sets. (Compare with raindrop plot based on actual data shown in Figure 2.12.) Note that the true mean of $\log \alpha$ is zero. Data sets 1–8 were assigned spawner observations at relatively low levels so that α would be “well determined”, whereas data sets 9–14 were assigned spawner observations only at higher levels so that α would be “poorly determined”. However the configuration of spawner observations is not the only determinant of whether α is well determined: data sets 2, 5, 7, and 8 have “ramped” profile likelihoods, indicating that α is poorly determined and for data set 9, α is relatively well determined.

The simulation described above was performed 30 times. In one of the cases, Efron’s method failed to converge. The mean estimate of the standard deviation of $\log \alpha$ was 1.04, only slightly different from its true value of 1. The mean estimate of the mean of $\log \alpha$ was 0.5, which differs substantially from its true value of 0.

Based on these results, we conclude that the approximate likelihood for $\log \alpha$ loses too

much information about the data. In Chapters 4 and 5, we consider linear and nonlinear mixed effects models that do not reduce the data to a single-parameter likelihood. Unlike Efron's approach, these models have traditionally been restricted to normal-normal hierarchies.

In the literature on nonlinear mixed effects models, a number of *two-stage* methods have been proposed (Davidian and Giltinan 1995, Chapter 5). These are methods in which individual estimates are obtained first and then combined in some way to estimate the model parameters. Some schemes, such as the *global two-stage* method (Steimer et al. 1984), incorporate the uncertainty in the individual estimates, based on estimated asymptotic variances. In simulations, Sheiner and Beal (1980) found that a two-stage approach produced good fixed effects estimates, but biased and imprecise estimates of variability among individuals.

In any case, conventional two-stage approaches are problematic when individual estimates are not available for all studies, as with the coho salmon data (e.g., see Figure 2.4). The application of Efron's method in this context avoids this difficulty and allows the use of better approximations (e.g. profile likelihoods). However, the simulation described above suggests that bias remains.

Chapter 4

Linear mixed effects models

In the previous chapter we considered the general framework of conditionally independent hierarchical models. Here we consider one of the first such models to be widely used: the linear mixed effects model. Much of the chapter reviews well-known methods and results, however some novel graphical displays are introduced in the context of the example datasets. Furthermore, the material covered here lays the groundwork for the nonlinear mixed effects models introduced in Chapter 5.

In Section 4.1 we introduce the general version of the linear mixed effects model. We show how the mixed model equations are obtained and how they lead to estimators of the fixed effects and predictors of the random effects. In Section 4.2.1 we return to the random effects model for meta-analysis of Section 1.4.2, and show that it fits into this framework and, as for the fixed effects model, leads to an estimator that is a weighted mean of individual estimates. We consider the estimation of variance-covariance components in Section 4.2.2. For the meta-analytic data considered in this work, linear mixed effects models have a special structure; in Section 4.3 we introduce linear mixed effects models for repeated measured data. Asymptotic properties for linear mixed effects models are a subject of continuing research and in Section 4.5 we review the available results. Approaches to robust estimation of linear mixed effects models are the subject of Section 4.6. The methods in this chapter are illustrated using the three example datasets with a particular focus on the wolf data.

4.1 General linear mixed effects model

The general linear mixed effects model can be written

$$y = X\beta + Zu + \varepsilon, \quad (4.1)$$

where X is an $n \times q$ design matrix, β is a q -dimensional fixed effects vector, Z is an $n \times m(c-1)$ design matrix, u is an $m(c-1)$ -dimensional random effects vector with mean 0 and variance-covariance matrix \bar{D} , and ε is an n -dimensional error vector with mean 0 and variance-covariance matrix R . The errors ε and random effects u are assumed independent.

A good introduction to linear mixed effects models is given by Searle, Casella, and McCulloch (1992).

4.2 Estimation

Estimation of linear mixed effects models has long been a topic of research interest. A comprehensive review of maximum likelihood and restricted maximum likelihood approaches was given by Harville (1977).

The observation vector y has mean $X\beta$ and variance-covariance matrix

$$V \equiv \text{Var}(y) = Z\bar{D}Z^T + R. \quad (4.2)$$

If V is known up to a multiplicative constant, β may be estimated using generalized least squares. The *generalized least squares* (GLS) estimator of β is given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \quad (4.3)$$

Derivation of the GLS estimator requires only assumptions on the first two moments of y . Traditionally, however, the errors and random effects are both assumed to be normally distributed. For now let us assume that \bar{D} and R are both known up to a multiplicative constant. Let us consider the distribution of u to be a prior distribution, and place an improper flat prior on β , i.e., $p(\beta) = 1$, independent of u . Then the posterior distribution of β and u based on y is proportional to the joint distribution of y and u , i.e., from (3.2) on

page 65,

$$(2\pi)^{-(n+c-1)/2} |R|^{-1/2} |\bar{D}|^{-1/2} \exp \left\{ -\frac{1}{2} [(y - X\beta - Zu)^T R^{-1} (y - X\beta - Zu) + u^T \bar{D}^{-1} u] \right\}. \quad (4.4)$$

To obtain the mode of the posterior, we differentiate the log of (4.4) with respect to β and u and equate to zero. This leads to the *mixed-model equations* of Henderson et al. (1959):

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & \bar{D}^{-1} + Z^T R^{-1} Z \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}. \quad (4.5)$$

Alternatively, the terms in (4.4) that depend on β and u can be written

$$\begin{aligned} & (y - X\beta - Zu)^T R^{-1} (y - X\beta - Zu) + u^T \bar{D}^{-1} u \\ &= \begin{bmatrix} R^{-1/2} (y - X\beta - Zu) \end{bmatrix}^T \begin{bmatrix} R^{-1/2} (y - X\beta - Zu) \end{bmatrix} + \begin{bmatrix} \bar{D}^{-1/2} u \end{bmatrix}^T \begin{bmatrix} \bar{D}^{-1/2} u \end{bmatrix} \\ &= \tilde{e}^T \tilde{e}, \end{aligned}$$

where

$$\begin{aligned} \tilde{e} &= \begin{bmatrix} R^{-1/2} (y - X\beta - Zu) \\ -\bar{D}^{-1/2} u \end{bmatrix} \\ &= \begin{bmatrix} R^{-1/2} y \\ 0 \end{bmatrix} - \begin{bmatrix} R^{-1/2} X\beta + R^{-1/2} Zu \\ \bar{D}^{-1/2} u \end{bmatrix} \\ &= \tilde{y} - \tilde{X}\xi, \end{aligned} \quad (4.6)$$

with

$$\tilde{y} = \begin{bmatrix} R^{-1/2} y \\ 0 \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} R^{-1/2} X & R^{-1/2} Z \\ 0 & \bar{D}^{-1/2} \end{bmatrix}, \quad \xi = \begin{bmatrix} \beta \\ u \end{bmatrix}, \quad \text{and } \tilde{e} \sim N(0, I).$$

Rewriting (4.6), we have

$$\tilde{y} = \tilde{X}\xi + \tilde{e},$$

which is known as a “pseudo-data” regression model (Lindstrom and Bates 1990; Davidian and Giltinan 1995). Defining

$$\hat{\xi} = \begin{bmatrix} \bar{\beta} \\ \bar{u} \end{bmatrix},$$

the normal equations for this regression model are

$$(\check{X}^T \check{X}) \hat{\xi} = \check{X}^T \check{y},$$

which are simply the mixed model equations (4.5). Solving them for $\bar{\beta}$ gives (4.3) (i.e., $\bar{\beta} = \hat{\beta}$, the generalized least squares estimator), and solving for \bar{u} gives

$$\bar{u} = \bar{D}Z^T V^{-1}(y - X\hat{\beta}), \quad (4.7)$$

which is the *best linear unbiased predictor* of u , denoted $BLUP(u)$. The BLUP is *linear* in the sense that it is a linear function of y , *unbiased* in the sense that $E(\bar{u}) = E(u) = 0$, and *best* in the sense that it has minimum mean squared error within the class of linear unbiased predictors (Robinson 1991). In practice we may not know \bar{D} and R (and hence V) and may substitute estimates of them in (4.7) giving the *empirical BLUP*.

4.2.1 Random effects meta-analysis

As a special case of (4.1), consider model (1.9) for random effects meta-analysis (page 16). Here the data vector is $y = (t_1, \dots, t_m)^T$. In this case there is just one fixed effect, β (i.e. $q = 1$) and the X matrix is a vector of 1's and the Z matrix is an identity matrix. The model has $c = 2$ random components and the variance-covariance matrices are $R = \text{diag}(v_1, \dots, v_m)$ and $D = \text{diag}(\sigma_u^2, \dots, \sigma_u^2)$. Therefore $V = \text{diag}(v_1 + \sigma_u^2, \dots, v_m + \sigma_u^2)$. The generalized least squares estimator of β is

$$\begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ &= \left(\sum_{i=1}^m (v_i + \sigma_u^2)^{-1} \right)^{-1} \sum_{i=1}^m (v_i + \sigma_u^2)^{-1} t_i \\ &= \sum_{i=1}^m w_i t_i / \sum_{i=1}^m w_i, \end{aligned}$$

where $w_i = (v_i + \sigma_u^2)^{-1}$. This is the same as the estimate of β (1.4) in the fixed effects model for meta-analysis except that the weights have a contribution from σ_u^2 . In particular, if $\sigma_u^2 = 0$, we recover (1.4). From (4.7), the BLUP of $\theta_i = \beta + u_i$ is

$$\begin{aligned}\hat{\beta} + \text{BLUP}(u_i) &= \hat{\beta} + \sigma_u^2(v_i + \sigma_u^2)^{-1}(t_i - \hat{\beta}) \\ &= B_i \hat{\beta} + (1 - B_i)t_i,\end{aligned}$$

where $B_i = v_i/(v_i + \sigma_u^2)$, which is the same as the empirical Bayes estimate of (1.11).

Example: Analysis of ulcer data using conventional fixed effects meta-analysis

As noted above, when $\sigma_u^2 = 0$, the estimate of β used in conventional fixed effects meta-analysis is recovered. To apply this to the ulcer data, we use the point estimates and standard errors from Section 2.2, treating the standard errors as if they were known without uncertainty. The result is shown in Figure 4.1.

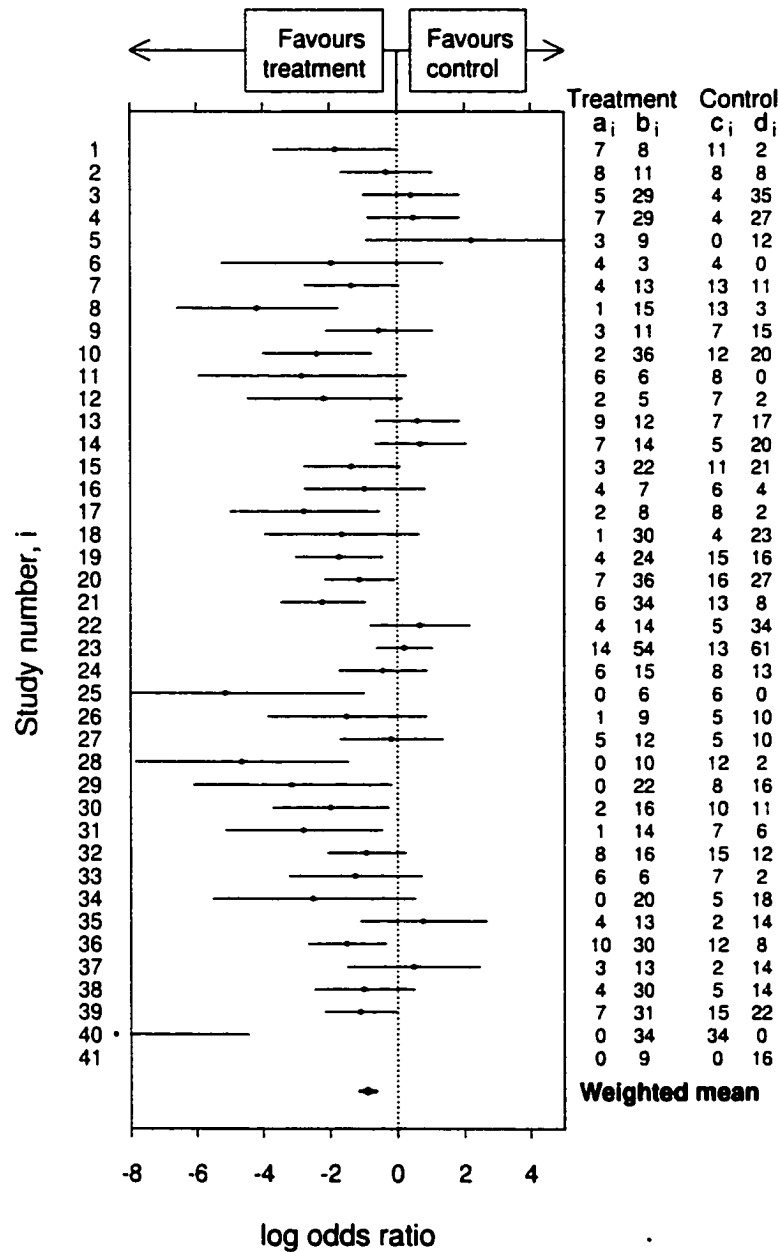


Figure 4.1: Point estimate and 95% confidence interval for the log odds ratio in each ulcer study, together with a meta-analytic summary. The summary, shown at the bottom, is the weighted mean with 95% confidence interval from a fixed effect analysis.

The chi-squared homogeneity statistic of Section 1.4.1 for this fixed effects analysis is 99.0

on 39 degrees of freedom, which is highly significant. This suggests that the fixed effects analysis is not adequate, i.e., there is substantially more variation than we would expect. A random effects model provides one way to account for this extra variation.

4.2.2 Variance-covariance components estimation

As noted above, we typically do not know the variance-covariance matrices \bar{D} and R . The elements of \bar{D} and R are known as *variance-covariance components*. Various structures for \bar{D} and R may be considered. For example, in *variance components models*, the covariance components are assumed to be zero (i.e. \bar{D} is diagonal) and only the variance components need to be estimated. We denote the parameters that determine \bar{D} and R by ζ .

Various methods for estimating variance-covariance components have been proposed (Searle, Casella, and McCulloch 1992). We concentrate here on likelihood-based approaches together with robust modifications.

We begin by assuming that the errors and random effects are both normally distributed. In the general notation of Chapter 3, our model is

$$\begin{aligned} y|\theta, \lambda &\sim N(\theta, R) \\ \theta|\lambda &\sim N(X\beta, Z\bar{D}Z^T), \end{aligned}$$

with $\lambda = (\beta^T, \zeta^T)^T$ and $\theta = X\beta + Zu$. From (3.3), the marginal density of the data is

$$p(y|\lambda) = \int p(y|\theta, \lambda)p(\theta|\lambda)d\theta,$$

where

$$p(y|\theta, \lambda) = (2\pi)^{-n/2}|R|^{-1/2} \exp\left\{-\frac{1}{2}[(y-\theta)^T R^{-1}(y-\theta)]\right\}$$

and

$$p(\theta|\lambda) = (2\pi)^{-(c-1)/2}|Z\bar{D}Z^T|^{-1/2} \exp\left\{-\frac{1}{2}[(\theta-X\beta)^T (Z\bar{D}Z^T)^{-1}(\theta-X\beta)]\right\}.$$

Therefore, integrating out θ ,

$$p(y|\lambda) = (2\pi)^{-n/2} |V|^{-1/2} \exp \left\{ -\frac{1}{2} [(y - X\beta)^T V^{-1} (y - X\beta)] \right\}, \quad (4.8)$$

i.e., y is marginally normally distributed with mean $X\beta$ and variance-covariance matrix V . The log likelihood is thus given by

$$\ell(\beta, \zeta) = -\frac{1}{2} [(y - X\beta)^T V^{-1} (y - X\beta) + \log |V|]. \quad (4.9)$$

Maximum likelihood (ML) estimators of β and ζ are obtained by maximizing (4.9). For known ζ , the ML estimate of β is the generalized least squares estimator (4.3).

Even without assuming normality, we may consider estimators that maximize (4.9), or equivalently minimize

$$(y - X\beta)^T V^{-1} (y - X\beta) + \log |V|. \quad (4.10)$$

This approach was called *extended least squares* (ELS) by Beal (1984), and we will return to it in Chapter 5.

For finite sample sizes, ML estimators for the variance components are biased. Intuitively, the bias occurs because the ML estimators do not take into account the loss in degrees of freedom from the estimation of β . The restricted maximum likelihood (REML) approach corrects for this bias by estimating variance components based on residuals after estimating the fixed effects by least squares. It may be shown (see, e.g., Verbyla (1990), Searle, Casella, and McCulloch (1992)) that this is equivalent to maximizing the likelihood based on $n - q$ linearly independent error contrasts of y . The vector $K^T y$ represents $n - q$ linearly independent contrasts if K is a known, full rank $n \times (n - q)$ matrix such that $K^T X = 0$. For any such matrix K , the log likelihood based on $K^T y$ may be written

$$\begin{aligned} \ell_R(\hat{\beta}, \zeta) &= -\frac{1}{2} [(y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) + \log |V| + \log |X^T V^{-1} X|]. \\ &= \ell(\hat{\beta}, \zeta) - \frac{1}{2} \log |X^T V^{-1} X|. \end{aligned} \quad (4.11)$$

This is the reduced or restricted log-likelihood introduced by Patterson and Thompson

(1971) and Patterson and Thompson (1974). We shall refer to it as the REML log likelihood. Note that (4.11) is written in terms of $\hat{\beta}$ whereas (4.9) is written in terms of β , which highlights the previously mentioned point that the REML likelihood is based on *residuals* after estimating the fixed effects, β . Note that the REML log likelihood can also be written

$$\ell_R(\hat{\beta}, \zeta) = -\frac{1}{2} [y^T P y + \log |V| + \log |X^T V^{-1} X|], \quad (4.12)$$

where $P = V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}$. Harville (1974) showed that the REML likelihood may also be derived by considering the marginal distribution of y when an improper flat prior is placed on β , i.e.,

$$p(y|\zeta) = \int p(y|\theta, \lambda) p(\theta|\lambda) p(\beta) d\theta d\beta, \quad (4.13)$$

with $p(\beta) \equiv 1$.

Example: Analysis of ulcer data using conventional random effects meta-analysis

To illustrate the estimation of variance components, we perform a conventional random effects meta-analysis of the ulcer data, using REML. As for the fixed effects analysis, we use the point estimates and standard errors from Section 2.2, treating the standard errors as if they were known with no uncertainty. The result is shown in Figure 4.2.

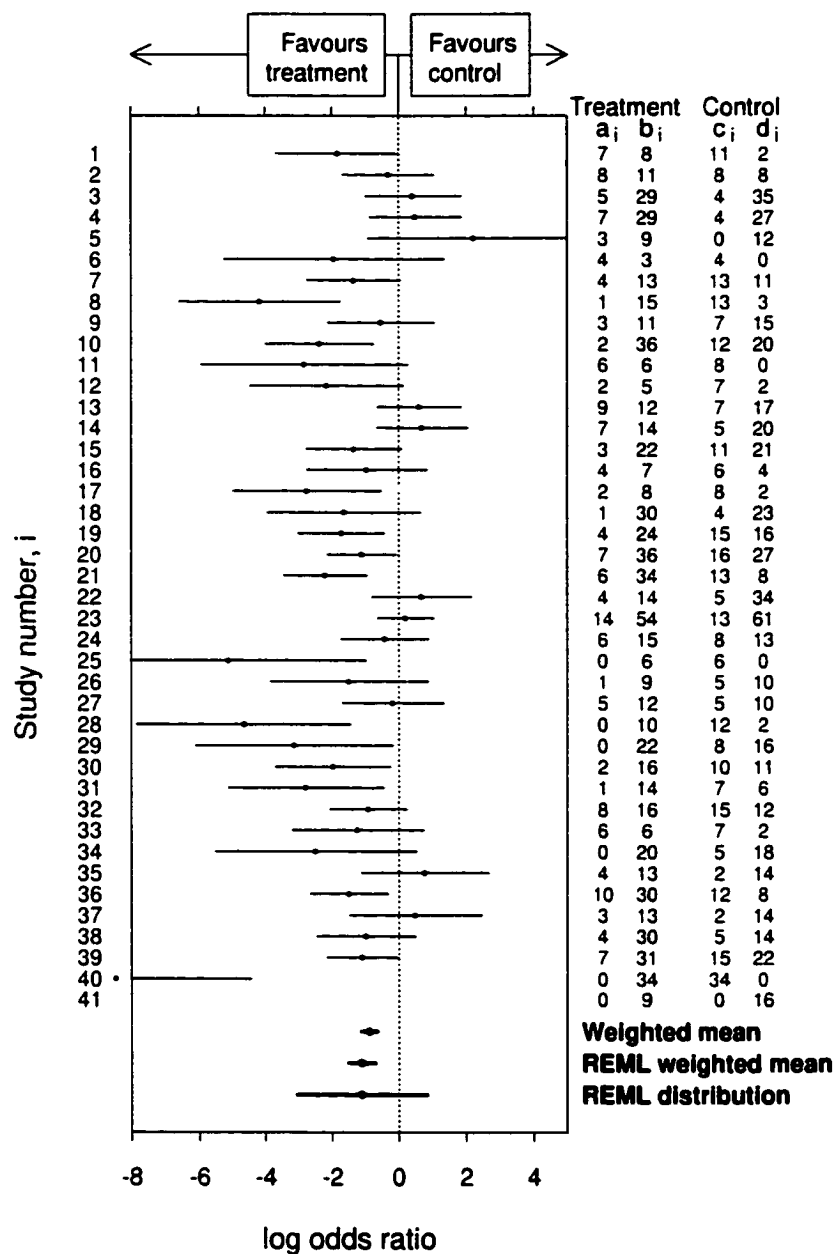


Figure 4.2: Point estimate and 95% confidence interval (CI) for the log odds ratio in each ulcer study, together with three meta-analytic summaries. The summaries, shown at the bottom, include the weighted mean from the fixed effect analysis (with 95% CI), the weighted mean from the REML random effect analysis (with 95% CI), and the 95% highest-density interval for the estimated distribution of a log odds ratio.

The estimated mean log odds ratio from the random effect analysis is slightly more in favour of the treatment than in the fixed effect analysis. As expected, the 95% confidence interval is wider than in the fixed effect analysis since it incorporates inter-study variation. The “REML distribution” interval in Figure 4.2 is a 95% highest-density interval. It indicates where most of the mass of the estimated random effect distribution is located, and is important because it conveys to the user of the graph the typical inter-study variability.

4.3 Linear mixed effects model for repeated measures data

The data considered in this work have a repeated measures structure, i.e. several observations are available from each of the m studies. So we partition the data vector y into individual study data vectors y_i of length n_i , i.e., $y = (y_1^T, \dots, y_m^T)^T$ and $n = \sum_{i=1}^m n_i$, and we correspondingly partition X , Z , u , and ϵ as

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix}, \quad Z = \begin{pmatrix} Z_1 & & \\ & \ddots & \\ & & Z_m \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}, \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix}, \quad (4.14)$$

where X_i is an $n_i \times q$ matrix, Z_i is an $n_i \times (c-1)$ matrix, u_i is a $(c-1)$ -dimensional vector, and ϵ_i is a n_i -dimensional vector. This gives

$$y_i = X_i \beta + Z_i u_i + \epsilon_i. \quad (4.15)$$

Laird and Ware (1982) used this model.

4.4 Estimation

We assume that the error vectors are mutually independent with $\text{Var}(\epsilon_i) = R_i$, so that $R = \text{diag}(R_1, \dots, R_m)$. Similarly, we assume that the random effects vectors are mutually independent with $\text{Var}(u_i) = D$, so that $\bar{D} = \text{diag}(D, \dots, D)$. The i th observation vector

y_i has mean $X_i\beta$ and variance-covariance matrix

$$V_i \equiv \text{Var}(y_i) = Z_i D Z_i^T + R_i, \quad (4.16)$$

so that $V = \text{diag}(V_1, \dots, V_m)$, i.e. V is block diagonal. The y_i 's are marginally independent, as shown by (3.3). The log likelihood can thus be written

$$\ell(\beta, \zeta) = -\frac{1}{2} \sum_{i=1}^m \left[(y_i - X_i\beta)^T V_i^{-1} (y_i - X_i\beta) + \log |V_i| \right]. \quad (4.17)$$

The REML log likelihood cannot be written as a sum because the matrix P in (4.12) is not block diagonal. This complicates the asymptotic theory for REML estimation, however Richardson and Welsh (1995) have shown that without assuming normality of the data and under fairly general conditions, REML estimates are consistent and asymptotically normal.

4.4.1 Example: Linear mixed effects models for spawner-recruitment data

Most of the spawner-recruitment models introduced so far are nonlinear. However, as we have noted, the Ricker model

$$R_{ij} = \alpha_i S_{ij} e^{-\beta_i S_{ij}},$$

can be transformed to obtain

$$\log \frac{R_{ij}}{S_{ij}} = \log \alpha_i - \beta_i S_{ij}$$

which is linear. As discussed in Chapter 1, the Ricker model is not suitable for modeling coho salmon because it “bends over” at high spawner levels, i.e., recruitment is not an increasing function of spawner quantity. For many other species, however, the Ricker model seems quite reasonable.

Recall from Chapter 1 that the α_i 's are positive and all have the same units, i.e., recruits per spawner. It may therefore be reasonable to treat the $\log \alpha_i$'s as normally distributed random effects. If we assume additive normal error on the transformed response, i.e.,

$$\log \frac{R_{ij}}{S_{ij}} = \log \alpha_i - \beta_i S_{ij} + \epsilon_{ij},$$

with $\varepsilon_{ij} \sim N(0, \sigma^2)$, then back-transforming shows that this is equivalent to assuming multiplicative lognormal error in the recruitment R_{ij} , a fairly reasonable model.

With these assumptions, the model can be written in the form of a linear mixed effects model $y_i = X_i\beta + Z_iu_i + \varepsilon_i$. Using a database of spawner-recruitment series for over 500 fish populations, Myers, Bowen, and Barrowman (1999) applied linear mixed effects models to obtain estimates of $\log\alpha_i$ and its variance for a variety of species and families (groups of species). They treated the β_i 's as population-specific nuisance parameters. This was necessary because Myers, Bowen, and Barrowman (1999) did not attempt to standardize the β_i 's: the units of the β_i 's vary from population to population.

4.4.2 Example: Linear mixed effects analysis of the wolf data

Recall model (2.1) for the wolf populations:

$$y_i = \theta_{i1} + \theta_{i2}x_i + \varepsilon_i,$$

where θ_{i1} is the intercept, θ_{i2} is the slope, and $\varepsilon_i \sim N(0, \sigma_i^2)$. The rate of change of the population in each reserve may reflect large-scale phenomena such as climate or management policies. The population trends in the 8 reserves are viewed as realizations from a larger conceptual "superpopulation" of possible population trends for wolf populations given the existing large-scale conditions. Therefore we model the slope for each reserve as a random effect, i.e.,

$$\theta_{i2} = \beta_1 + u_i,$$

where β_1 is the mean slope, and $u_i \sim N(0, \sigma_u^2)$. Treating the reserve-specific slopes as random effects is a reflection of a prior belief that the population trends in the 8 reserves are not completely unrelated. The intercept for each reserve determines the initial population density. Note that the reserves vary in size and density of wildlife, which can be thought of as intrinsic to each reserve, and unrelated to any other reserves. We therefore treat the intercepts as reserve-specific fixed effects, i.e.,

$$\theta_{i1} = \beta_{i+1},$$

where β_{i+1} is the fixed intercept for the i th reserve. We collect the β_k 's together into a vector β of length $q = m + 1$. The model can then be written in the form $y_i = X_i\beta + Z_iu_i + \varepsilon_i$ where $Z_i = x_i$ and X_i is an $n_i \times (m + 1)$ matrix whose first column is the x_i vector, and whose remaining columns are zero's except for the $(i + 1)$ th column which is all ones. Figure 4.3 shows fits from a REML analysis of the wolf data using this model. The dashed lines in each panel are the BLUPs, i.e.,

$$\hat{y}_i = X_i\hat{\beta} + Z_i\hat{u}.$$

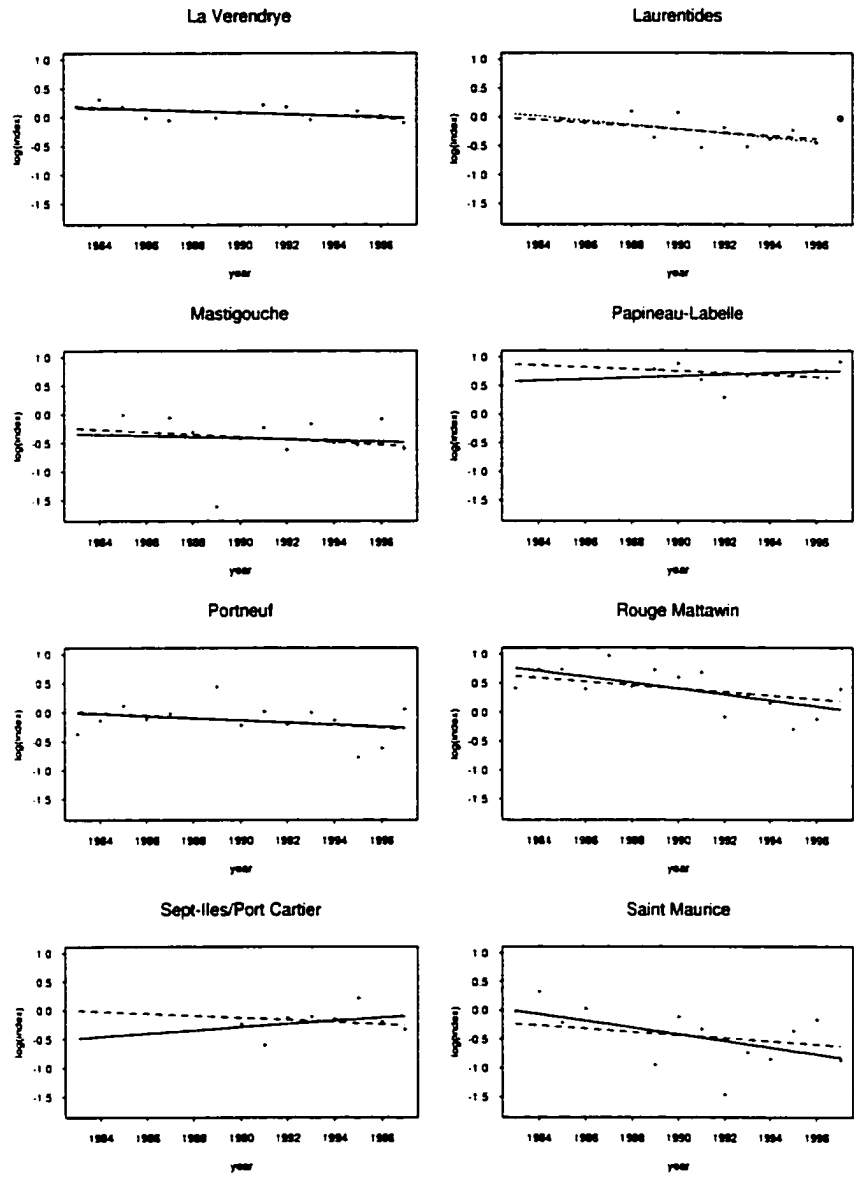


Figure 4.3: Log population numbers of wolves in 8 reserves in southern Québec with fitted least squares regression lines (solid) and BLUPs (dashed).

A conventional meta-analytic plot is shown in Figure 4.4.

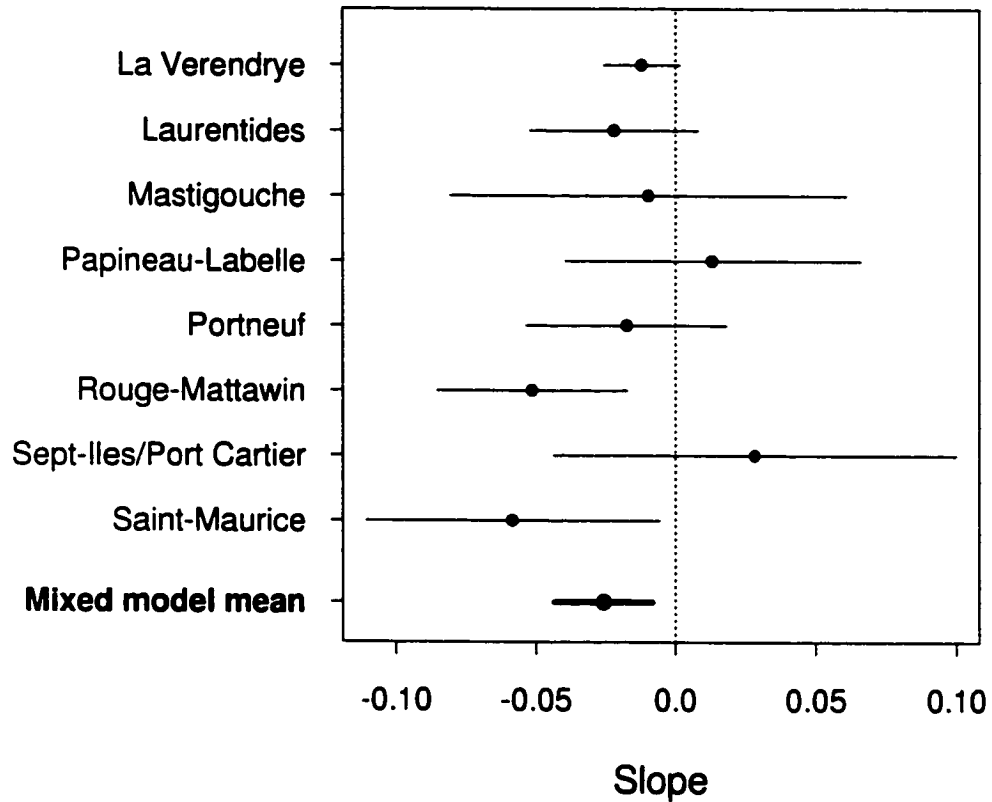


Figure 4.4: Point estimate and 95% confidence interval for each reserve for the slope of the least squares regression of log numbers of wolves versus year together with a meta-analytic summary.

The estimate of the inter-reserve variance of the slope, $\hat{\sigma}_u^2$, is tiny, approximately 10^{-4} . In other words, the model-estimates of the reserve-specific slopes are essentially identical. This suggests that a model with reserve-specific fixed effects would suffer from an excess of parameters with an associated loss of power to discriminate population declines.

Note that the model specifies a separate variance for each reserve. However, recall from Section 2.1 that we actually have some additional information, namely estimates of the effective sample sizes for observations from each reserve. Table 4.1 shows these sample sizes together with estimates of the within-reserve variances $\hat{\sigma}_i^2$ from our model:

Reserve	Effective sample size, n_i^*	Model estimate of σ_i^2
La Vérendrye	378	.012
Laurentides	355	.032
Mastigouche	168	.187
Papineau-Labelle	198	.045
Portneuf	100	.088
Rouge-Mattawin	130	.089
Sept-Îles/Port-Cartier	48	.062
Saint-Maurice	66	.197

Table 4.1: Effective sample sizes for observations from each reserve together with estimate of within-reserve variances from REML fit of linear mixed effects model.

To reduce the number of parameters to estimate, we now change our model so that $\sigma_i^2 = \sigma^2/n_i$. The revised model results in an estimate of zero for the inter-reserve variance of the slope. The mean slope is estimated to be -0.026 with a standard error of 0.009 . Using those point estimates suggests that none of the slopes could in fact be positive. However, this is a misleading conclusion, since it ignores the uncertainty in the estimates. Figure 4.5 is a graphical display of the range of plausible values of the mean and the standard deviation of the slope, with the associated probabilities of a positive slope. The “mushroom”-shaped contour is the joint 95% likelihood-based confidence interval for the mean and standard deviation of the slope. In the upper right portion of the mushroom, the mean and standard deviation are both relatively large. The probability of a positive slope is then up to about 0.4.

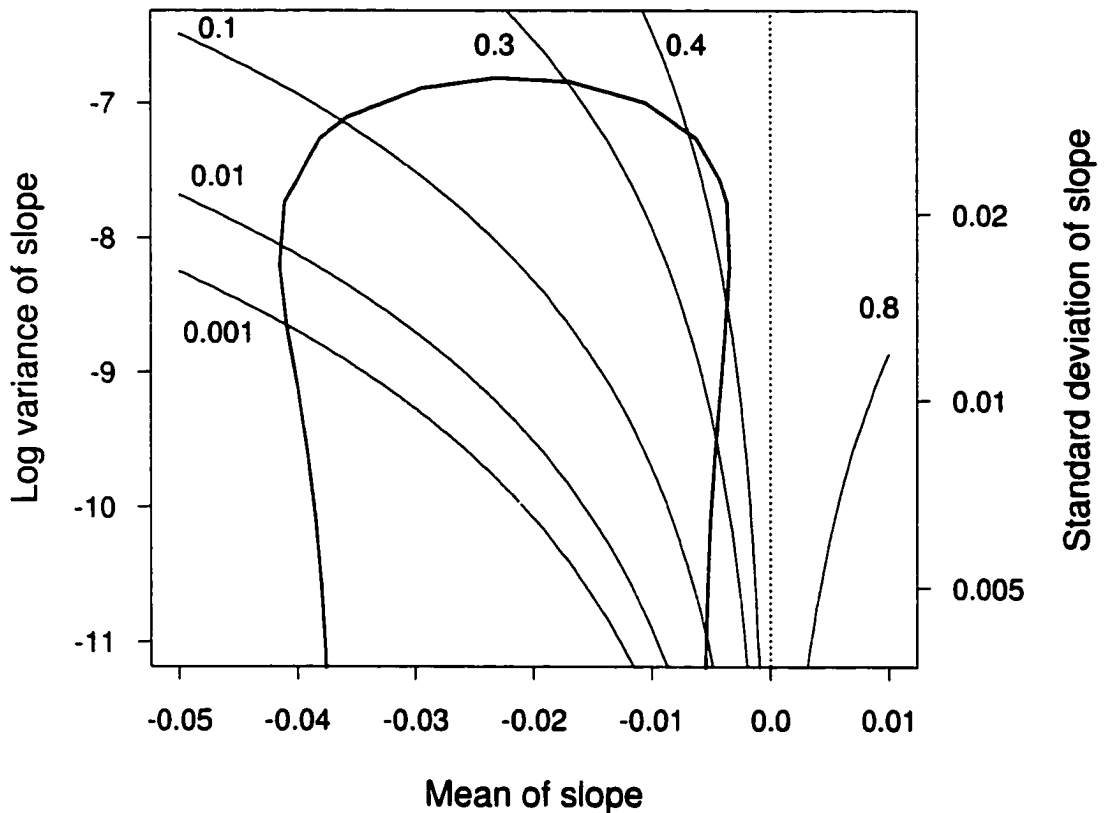


Figure 4.5: Approximate joint 95% confidence interval for $\log \sigma_u^2$ and $\log \beta_1$ (heavy line) with superimposed contours of constant probability (lighter lines), showing the probability of a positive slope given a normal distribution with mean β_1 and variance σ_u^2 . The dotted vertical line indicates a mean slope of zero.

A hierarchical Bayesian approach to this problem would make use of prior distributions for the mean and variance of the slope. But the choice of prior distributions for variances is quite difficult and in small samples “noninformative” priors may in fact be informative (Daniels 1999). Note first that the boundary value $\sigma_u^2 = 0$ is supported by non-negligible likelihood since it is possible that there is no between-study variance (cite DuMouchel). The standard noninformative prior is $p(\sigma_u^2) \propto 1/\sigma_u^2$ (Box and Tiao 1973), however this has an asymptote at zero, which can lead to an improper posterior. Smith, Spiegelhalter,

and Thomas (1995) argue that a reasonable approach is to search for a proper prior for the inter-study variance and provide an example in the context of a meta-analysis of 2×2 tables. However their reasoning is not entirely convincing since it relies on questionable assumptions about expected ranges of effect sizes.

Figure 4.6 shows how little information the likelihood contains about the standard deviation of the slope, σ_u , in the present case. A likelihood ratio test suggests that values of σ_u up to about 0.025 are plausible. For values in this range, up to about 20% of the mass of the distribution of the slope is positive. In practical terms this means that 20% of the wolf populations may in fact be increasing.

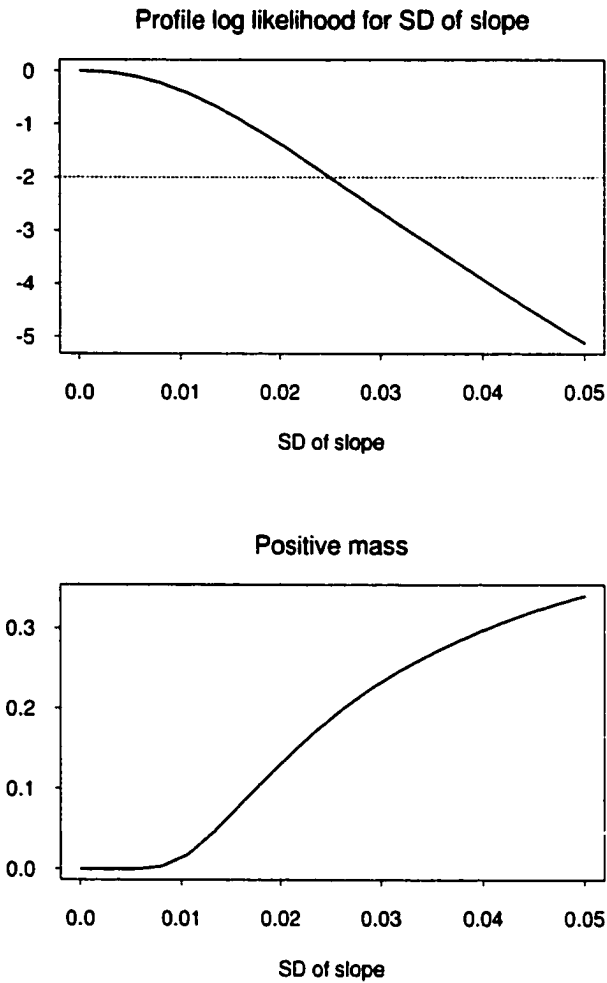


Figure 4.6: Top panel: Profile log likelihood for σ_u , the standard deviation of the slope random effect. The horizontal line depicts a decrease in the log likelihood of 2, which corresponds to a likelihood ratio test at an approximate 95% level. Bottom panel: Positive mass of the distribution of slope as a function of σ_u .

4.5 Asymptotics

Standard asymptotic theory for ML estimators cannot be applied to linear mixed effects models because the observations are not independent. Under the assumption of normality,

Hartley and Rao (1967) and Miller (1977) obtained the asymptotic distribution of the ML estimator for mixed effects ANOVA models. Pinheiro (1994) extended their results to the more general linear mixed effects model (4.1). Recently Cressie and Lahiri (1993) obtained the asymptotic distribution of the REML estimator, again under the assumption of normality.

In practice we can never be sure that data are normally distributed, much less the unobservable random effects. Following work by Westfall (1986), Richardson and Welsh (1994) derived the asymptotic distribution of the ML and REML estimators for a class of linear mixed effects models in which y can be partitioned into independent sub-vectors without the assumption that the data are either normally or spherically symmetrically distributed.

For the models considered by Richardson and Welsh (1994), the Z matrix is assumed to have a special structure, and they make assumptions so that as the number of observations increases, the structure is preserved. For these models, the observations y can be partitioned into independent sub-vectors.

For such models, Welsh and Richardson (1997) summarize asymptotic results for a general class of estimators that are solutions to estimating equations of the form

$$\sum_{i=1}^m \Psi_i(\lambda) = 0.$$

For example, the maximum likelihood estimator is of this form, and the REML estimator is asymptotically of this form. An estimator $\hat{\lambda}$, that satisfies the above equation is estimating a root, λ_ψ , of

$$\sum_{i=1}^m E\Psi_i(\lambda) = 0.$$

The information matrix is given by

$$G_n = \frac{1}{n} \sum_{i=1}^m E \left(\frac{d\Psi_i}{d\lambda} \Big|_{\lambda_\psi} \right).$$

The variance of the normalized score function is

$$F_n = \frac{1}{n} \sum_{i=1}^m E(\Psi_i(\lambda_\psi)\Psi_i(\lambda_\psi)^T).$$

Welsh and Richardson (1997) assume that $m \rightarrow \infty$ as $n \rightarrow \infty$ as that $G_n \rightarrow G$ and $F_n \rightarrow F$. With additional assumptions about bounded moments, bounded norms of the X_i matrices, and nonsingularity of the variance-covariance matrices, they show that

$$n^{-1/2}(\hat{\lambda} - \lambda) \xrightarrow{D} N(0, G^{-1}FG^{-1}), \quad \text{as } n \rightarrow \infty.$$

4.6 Approaches to robust estimation

The methods in this thesis were all developed starting from normality assumptions. One approach is to assume that the errors and random effects are normally distributed. For the linear mixed effects model, it then follows that marginally the observations y are normally distributed; see (4.8). (In the next chapter, we will see that this does not hold for the nonlinear mixed effects model, making estimation more difficult.) Alternatively, the assumption of marginal normality can be made without assuming normality of the errors and random effects. Even without assuming normality, estimators developed for the normal marginal distribution may be used; see (4.10). We will call the distributional assumptions under which we are operating the *model distribution*. Whatever our assumptions, deviations from the model distribution can be detrimental to our inferences, and we seek distributionally robust methods.

A very useful account of approaches to robust estimation was provided by Stahel and Welsh (1997) in the simple case of a balanced one-way analysis of variance model. Stahel and Welsh found that nonrobust methods can be extremely inefficient when the assumption of normality is violated.

Welsh and Richardson (1997) provide a comprehensive review of approaches to robust estimation of linear mixed effects models. They note that deviations from the model distribution can arise from contamination of the errors or any of the random effects. This is an important point from a meta-analytic perspective because we can conceive of outlying observations within studies as well as outlying studies. A common criticism of a meta-analysis is that it “combines apples and oranges”, i.e. that the studies are measuring different things. The notion of “outlying studies” is one way to formalize this, and robust methods may provide some protection or at least diagnosis of the problem.

We adopt the approach of Huggins (1993b); see also (Huggins 1993a). Richardson and

Welsh (1995) extended the approach of Huggins (1993b), which they called Robust ML I. They worked within a class of linear mixed effects models in which y can be partitioned into independent sub-vectors. To obtain the Robust ML I estimator we start with the log likelihood (4.17) for a normal marginal distribution:

$$-\frac{1}{2} \sum_{i=1}^m \left[(y_i - X_i \beta)^T V_i^{-1} (y_i - X_i \beta) + \log |V_i| \right].$$

Now let $V_i^{-1/2}$ be a square root of V_i^{-1} (i.e., setting $V_i^{-1/2} = A$, we have $AA^T = V_i^{-1}$), so we can write the log likelihood as

$$-\frac{1}{2} \sum_{i=1}^m \left\{ \left[V_i^{-1/2} (y_i - X_i \beta) \right]^T \left[V_i^{-1/2} (y_i - X_i \beta) \right] + \log |V_i| \right\}.$$

Let $r_i = V_i^{-1/2} (y_i - X_i \beta)$ be the vector of standardized residuals for study i and let r_{ij} be the j th element of r_i . Then we can write the log likelihood as

$$-\frac{1}{2} \sum_{i=1}^m \left(\sum_{j=1}^{n_i} r_{ij}^2 + \log |V_i| \right).$$

If any of the standardized residuals r_{ij} is large, it will have considerable influence on the likelihood. A natural strategy is to replace r_{ij}^2 above by a less rapidly growing function $\rho(r_{ij})$. This modification will result in a systematic bias when the data actually have a normal marginal distribution, and a *consistency correction* is required. With these modifications, the objective is to minimize

$$\sum_{i=1}^m \left(\sum_{j=1}^{n_i} \rho(r_{ij}) + \frac{\kappa}{2} \log |V_i| \right),$$

where κ is the consistency correction determined by our choice of ρ . When $\rho(r) = r^2/2$, we recover the negative log likelihood by setting $\kappa = 1$. To streamline the presentation, we make use of a slight abuse of notation and let $\rho(r_i) = \sum_{j=1}^{n_i} \rho(r_{ij})$, so that the objective

function can be written

$$\sum_{i=1}^m \left(\rho \left[V_i^{-1/2} (y_i - X_i \beta) \right] + \frac{\kappa}{2} \log |V_i| \right). \quad (4.18)$$

Huggins (1993b) proposed using Tukey's bisquare ρ -function (see Figure 2.6, p. 37). Welsh and Richardson (1997) note that in the linear mixed effects model context, we require both $\psi(r)$ and $r\psi(r)$ to be bounded for robustness. To see why this is so, note that the estimating equations obtained by equating the derivatives of (4.18) to zero are

$$\sum_{i=1}^m X_i^T V_i^{-1/2} \psi \left[V_i^{-1/2} (y_i - X_i \beta) \right] = 0$$

for the fixed effects, and

$$\frac{1}{2} \sum_{i=1}^m \left\{ (y_i - X_i \beta)^T V_i^{-1/2} \left(\frac{\partial V_i^{1/2}}{\partial \sigma_k^2} \right) V_i^{-1/2} \psi \left[V_i^{-1/2} (y_i - X_i \beta) \right] - \kappa \text{tr} \left[V_i^{-1} \left(\frac{\partial V_i}{\partial \sigma_k^2} \right) \right] \right\} = 0$$

for the k th variance component σ_k^2 . We see that the estimating equation for the fixed effects controls the effects of extreme observations provided $\psi(r)$ is bounded. However, the estimating equation for the variance components does not control the effects of extreme observations unless $r\psi(r)$ is bounded. Huber's ψ -function does not have this property, however redescending ψ -functions such as Tukey's biweight do. Unfortunately, redescending ψ -functions may lead to multiple solutions (Staudte and Sheather 1990, p. 118).

To avoid the use of redescending ψ -functions, Richardson and Welsh (1995) proposed the Robust ML II estimator, which is based on modifying the maximum likelihood estimating equations, rather than the likelihood. The Robust ML II estimating equation for the variance components features two (possibly different) ψ -functions. I will not pursue the Robust ML II approach here because my approach involves directly optimizing a robustified likelihood, rather than using estimating equations. In the models used by Welsh and Richardson (1997), the necessary derivatives are obtained analytically, but this is not the case for linear mixed effects models having regression structure, much less for the nonlinear mixed effects models discussed later on.

4.6.1 Robust estimation of realized random effects

Fellner (1986) suggested using analogues of the BLUPs to robustly estimate realized random effects. These were a natural by-product of his algorithm for robust estimation of linear mixed effects models. Fellner's algorithm is based on the Henderson-Harville algorithm (Henderson 1963; Henderson 1973; Harville 1977) for variance components estimation, which involves iterative solution of the mixed model equations (4.5). In the present context, we simply use the empirical BLUP formula (4.7), $\bar{u} = \bar{D}Z^T V^{-1}(y - X\hat{\beta})$, using the robust estimates from Robust ML I. Let $\bar{\theta}_i = X_i\hat{\beta} + Z_i\bar{u}_i$ be the BLUP of θ_i .

When we refer to *empirical Bayes fits* for the wolf data we mean lines defined by

$$y_i = \hat{\beta}_{i+1} + \bar{\theta}_{i2}x_i.$$

When we refer to *marginal mean fits* for the wolf data we mean lines defined by

$$y_i = \hat{\beta}_{i+1} + \hat{\beta}_1x_i.$$

4.6.2 Example: Wolves

We implemented Robust ML I for the wolf data using the Huber ψ -function with a tuning constant of 2. This follows the approach of Welsh and Richardson (1997). As discussed above, this does not ensure robustness of the estimates, however it avoids numerical difficulties associated with redescending ψ -functions.

The estimation algorithm was based on the estimating equation approach of Richardson and Welsh (1995). Richardson's algorithm had to be modified because in the ANOVA models she used, the variance matrix V_i has a special form for which a symmetric square root $V_i^{-1/2}$ and its derivatives with respect to variance components may be obtained explicitly. Because of the regression structure of the wolf data, there is no such convenience. Instead we used a Choleski decomposition for the square root, and numerical derivatives. Richardson's algorithm also required modification to incorporate the sample size weighting.

The parameter estimates were very similar to the nonrobust fits. The variance of the slope random effect is estimated to be very close to zero. Figure 4.7 shows the fits with observation weights.

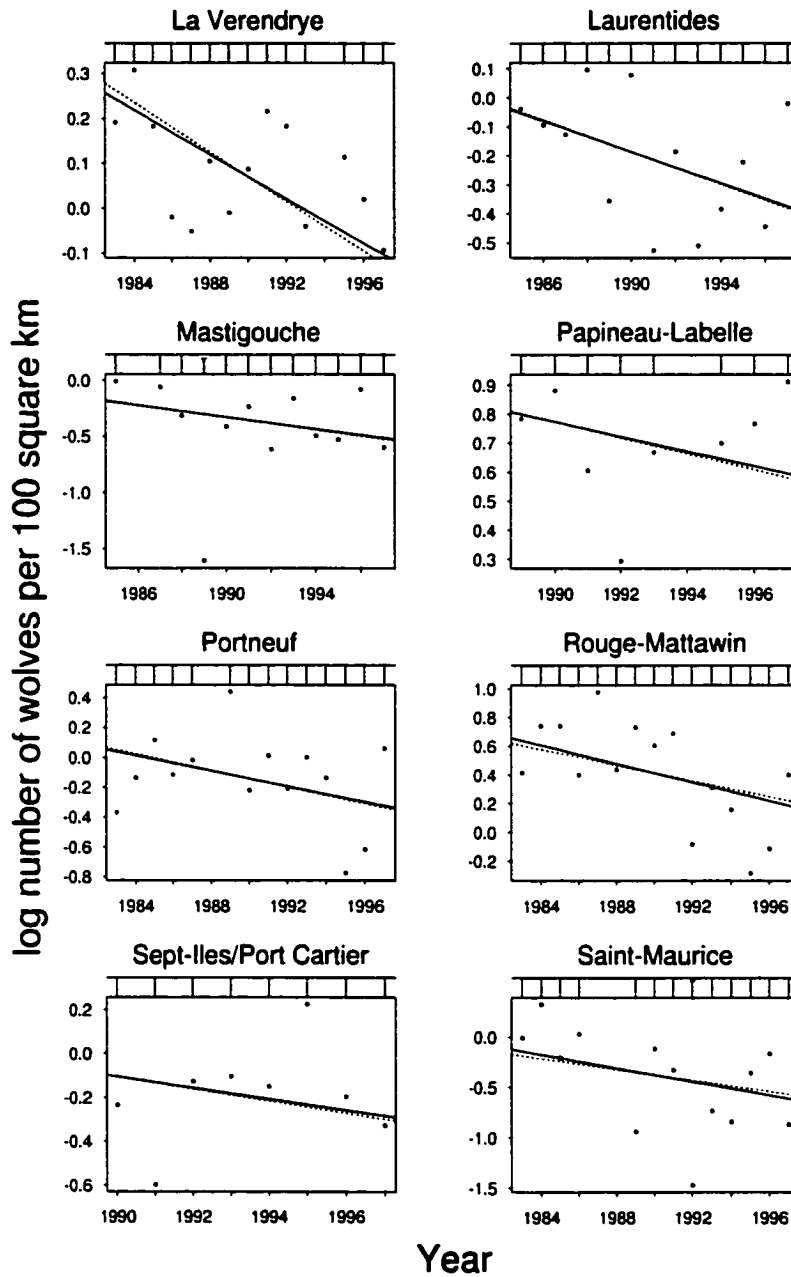


Figure 4.7: Robust mixed model empirical Bayes fits (solid) and marginal mean fit (dotted) to log numbers of wolves. The whiskers along the top margin show the weights from the robust fit, with a horizontal line to help discriminate downweighted observations.

Unlike the individual robust fits shown in Figure 2.7 (p. 41), there is little downweighting in the mixed model fits. Only the highly unusual point from Mastigouche is slightly

downweighted. In a mixed model analysis, the downweighting of individual points seems to be less critical. For example, the 1992 observation in the Papineau-Labelle wolf reserve appears to be quite unusual in the analysis of the individual data set. The whisker for this observation in the mixed model analysis suggests that it is not at all unusual.

Chapter 5

Nonlinear mixed effects models

The linear mixed effects models of the previous chapter are useful and have been widely applied. We have seen that these models may be applied to the wolf data and that Ricker dynamics for multiple fish populations may be estimated using this framework upon applying a simple 0. Similarly, the Beverton-Holt spawner-recruitment model (1.1) is *transformably linear*: its reciprocal is

$$\begin{aligned}\frac{1}{R_{ij}} &= \frac{1}{R_i^{\max}} + \frac{1}{\alpha_i S_{ij}} \\ &= \theta_{i1} + \theta_{i2} x_{ij}\end{aligned}\tag{5.1}$$

where $\theta_{i1} = 1/R_i^{\max}$, $\theta_{i2} = 1/\alpha_i$, and $x_{ij} = 1/S_{ij}$. However as Bates and Watts (1988, p. 34) point out, transformation of a model affects the distributional assumptions we are making. In the linear mixed effects model (4.15), errors are assumed to be additive normal. While (5.1) is linear, we may not wish to assume additive normal errors on this scale, since this would allow negative values of $1/R_{ij}$, which is not meaningful. Some models are not even transformably linear, for example the hockey-stick spawner-recruitment model. For these reasons we must move beyond linear mixed effects models, and the subject of this chapter is a class of nonlinear mixed effects models. A comprehensive account of methods of inference for nonlinear mixed effects models up to 1995 is provided by Davidian and Giltinan (1995).

It is noteworthy that one of the earliest methods for estimation of nonlinear mixed effects models (Sheiner and Beal 1980) was illustrated using the Michaelis-Menten model,

as the Beverton-Holt is known in pharmacokinetics.

In Section 5.1, we introduce a nonlinear mixed effects model for repeated measures data. This is followed, in Section 5.2, by an account of some methods of estimation, illustrated by application to the salmon data. The two methods of primary interest are those of Lindstrom and Bates (1990) and Nuñez (1998). To reduce variability, we propose a minor modification of the method of Nuñez (1998). Little has been published on robust estimation of nonlinear mixed effects models. In Section 5.3 we propose two methods for robust estimation, and discuss their properties.

5.1 Nonlinear mixed effects model for repeated measures data

The linear mixed effects models previously considered generalize in a straightforward way to the nonlinear case. For the i th study, let

$$y_i = \eta_i(\theta_i) + \varepsilon_i, \quad (5.2)$$

and

$$\theta_i = A_i\beta + B_iu_i, \quad (5.3)$$

where $\eta_i(\cdot)$ is an n_i -dimensional vector-valued nonlinear function, θ_i is an r -dimensional vector, A_i is an $r \times q$ design matrix, and B_i is an $r \times (c - 1)$ design matrix. As before, y_i is an n_i -dimensional observation vector, ε_i is an n_i -dimensional vector of errors with mean 0 and variance-covariance matrix R_i , β is a q -dimensional vector of fixed effects, and u_i is a $(c - 1)$ -dimensional vector of random effects with mean 0 and variance-covariance matrix D . The errors ε_i are assumed mutually independent and the random effects u_i are assumed mutually independent. Finally the errors and the random effects are assumed independent of each other. As for linear mixed effects models, the errors and random effects are typically assumed to be normally distributed.

This model has been used by Lindstrom and Bates (1990) and others. A more general version of this model was used by Nuñez (1998). Because this model is fairly new, there is

some variation in terminology in the literature. Some authors reserve the term “nonlinear mixed effects models” for models in which $\eta_i(\theta_i)$ is linear in the random effects u_i , which affords some simplifications. Ramos and Pantula (1995) prefer the term *nonlinear random coefficients model* for models in which $\eta_i(\theta_i)$ has a more general form.

5.1.1 Example: Beverton-Holt spawner-recruitment model

We will now show that the Beverton-Holt model for multiple fish populations can be formulated as a nonlinear mixed model. From Section 2.3 (p. 27), the Beverton-Holt model can be written

$$y_{ij} = -\log(1/\alpha_i + S_{ij}/R_i^{\max}),$$

where $y_{ij} = \log(R_{ij}/S_{ij})$. As discussed in Section 4.4.1 (p. 89), an assumption of additive normal error on this scale seems fairly reasonable. Recruitment is a non-negative quantity, and this error assumption implies that recruitment must be positive. Positivity constraints for the parameters α_i and R_i^{\max} must also be incorporated into the model. One way to do this is by writing

$$\alpha_i = e^{\theta_{i1}} \quad \text{and} \quad R_i^{\max} = e^{\theta_{i2}},$$

for unconstrained parameters θ_{i1} and θ_{i2} . To put this into a mixed model framework, let us write $\theta_{i1} = \beta_1 + b_i$ and $\theta_{i2} = \beta_2 + f_i$ with β_1 and β_2 fixed and

$$\begin{pmatrix} b_i \\ f_i \end{pmatrix} \stackrel{\text{iid}}{\sim} \text{N} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_b^2 & \sigma_{bf} \\ \sigma_{bf} & \sigma_f^2 \end{pmatrix} \right].$$

As usual, we group the elements θ_{i1} and θ_{i2} into a vector $\theta_i = (\theta_{i1}, \theta_{i2})^T$. Thus we can write our model as

$$y_i = \eta_i(\theta_i) + \varepsilon_i,$$

where

$$\eta_i(\theta_i) = \begin{bmatrix} -\log(e^{-\theta_{i1}} + S_{i1}e^{-\theta_{i2}}) \\ \vdots \\ -\log(e^{-\theta_{i1}} + S_{in_i}e^{-\theta_{i2}}) \end{bmatrix},$$

and

$$\theta_i = \beta + u_i,$$

with $\beta = (\beta_1, \beta_2)^T$ and $u_i = (b_i, f_i)^T$. Thus $\theta_i = A_i\beta + B_iu_i$ with $A_i = I$ and $B_i = I$.

5.1.2 Related models

A number of different models are closely related to the nonlinear mixed effects model. An increasingly popular class of models is *generalized linear mixed models* (GLMMs) for repeated measures data. In these models the random effects u_i are again normally distributed, but the distribution of the observations conditional on the random effects belongs to an exponential family with mean

$$E(y_i|u_i) = h^{-1}(X_i\beta + Z_iu_i),$$

for a specified monotone link function h . Compared to the nonlinear mixed effects model, the GLMM allows a more general conditional distribution but restricts the conditional mean function.

The ulcer data, analyzed by Efron (1996) using empirical Bayes methods for combining likelihoods (Chapter 3), were re-analyzed by Platt et al. (1999) using a non-central hypergeometric GLMM.

At first sight, the GLMM framework seems suitable for the Beverton-Holt spawner-recruitment model. In the introduction to this chapter we saw that taking the reciprocal of the Beverton-Holt gives a linear model. A natural approach would appear to be to use a GLMM with an inverse link and $\theta_{i1} = 1/R_i^{\max}$ and $\theta_{i2} = 1/\alpha_i$ as in (5.1). The canonical distribution for an inverse link is the gamma, which would constrain the recruitment to be positive, as it must be. The parameters R_i^{\max} and α_i must also be positive, however the GLMM requires that $\theta_{i1} = 1/R_i^{\max}$ and $\theta_{i2} = 1/\alpha_i$ be normally distributed, which allows negative values. With sufficient data, this might not be a problem, however the assumption that $1/R_i^{\max}$, for instance, is normally distributed seems unacceptable. In a nonlinear mixed effects model, it is always possible to choose a parameterization that constrains the parameters of interest in a suitable way. The restrictive structure of the conditional mean function in the GLMM is thus apparent.

A related approach for the analysis of longitudinal data was developed by Liang and Zeger (1986). In their approach, the marginal distribution of the data is modeled using

generalized linear models and estimates are obtained using generalized estimating equations which take the correlation of repeated measurements into account. The correlation is regarded essentially as a nuisance parameter and estimates are interpreted in terms of the average over the population of studies. In contrast, the nonlinear mixed effects model considered here gives study-specific estimates and the variance-covariance components are themselves of interest.

5.1.3 General nonlinear mixed effects model

More generally, nonlinear mixed effects models may not assume a repeated measures design. As in the previous chapter, a general notation may be used to write our model in terms of the entire observation vector $y = (y_1^T, \dots, y_m^T)^T$. If we define

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_m \end{pmatrix}, \quad \eta(\theta) = \begin{pmatrix} \eta_1(\theta_1) \\ \vdots \\ \eta_m(\theta_m) \end{pmatrix}, \quad \text{and} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix}, \quad (5.4)$$

then we can write (5.2) as

$$y = \eta(\theta) + \varepsilon.$$

If we also define

$$A = \begin{pmatrix} A_1 \\ \vdots \\ A_m \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & & \\ & \ddots & \\ & & B_m \end{pmatrix}, \quad \text{and} \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}, \quad (5.5)$$

then we can write (5.3) as

$$\theta = A\beta + Bu.$$

As in the previous chapter, we define $R \equiv \text{Var}(\varepsilon) = \text{diag}(R_1, \dots, R_m)$ and $\bar{D} \equiv \text{Var}(u) = \text{diag}(D, \dots, D)$, and we define ζ to be the parameters that determine R and \bar{D} .

5.2 Estimation

The fact that the random effects μ enter our model nonlinearly makes estimation considerably more difficult than in the linear case. This is because there is, in general, no closed form expression for the likelihood. Furthermore, it is not clear how to define a restricted likelihood for nonlinear mixed effects models. Estimation algorithms have been proposed by, among others, Sheiner and Beal (1980), Lindstrom and Bates (1990), Pinheiro and Bates (1995), and Nuñez (1998). In this section we consider several different approaches to estimation for the nonlinear mixed effects model.

We begin by considering the marginal distribution of y . Because the normal random effects enter the model through an arbitrarily complex nonlinear function, the marginal distribution of y will in general be non-standard. However, Vonesh and Carter (1992) considered models where the random effects enter linearly, i.e.,

$$\eta_i(\theta_i) = f_i(\beta) + Z_i\mu_i,$$

and in this case, $y_i \sim N(\mu_i, V_i)$, where

$$\mu_i \equiv E(y_i) = f_i(\beta) \quad \text{and} \quad V_i \equiv \text{Var}(y_i) = Z_i D Z_i^T + R_i.$$

If y is normally distributed, then the log likelihood is

$$\begin{aligned} & -\frac{1}{2} [(y - \mu)^T V^{-1} (y - \mu) + \log |V|] \\ = & -\frac{1}{2} \sum_{i=1}^m (y_i - \mu_i)^T V_i^{-1} (y_i - \mu_i) + \log |V_i|, \end{aligned} \quad (5.6)$$

where $\mu = E(y)$ and $V = \text{Var}(y)$. Note that μ and V are functions of β and ξ , and as long as they can be computed, the likelihood can be maximized. The results of Hoadley (1971) show that under regularity conditions, these estimators are consistent and asymptotically normal. As in the previous chapter, even without assuming normality, we may consider estimators that maximize (5.6), or equivalently, minimize

$$(y - \mu)^T V^{-1} (y - \mu) + \log |V|. \quad (5.7)$$

Beal (1984) called these *extended least squares* (ELS) estimators and showed that under regularity conditions, such estimators are consistent and asymptotically normal. As noted above, however, for nonlinear mixed effects models there is in general no explicit form for the mean

$$\mu = E(y) = E[\eta(A\beta + Bu)]$$

and the variance-covariance matrix

$$V = \text{Var}(y) = \text{Var}[\eta(A\beta + Bu)] + R.$$

We will return to this later, but for now turn to an approach similar to the one used for linear mixed effects models.

Suppose instead that u and ε are normally distributed, and suppose that their variance covariance matrices, \tilde{D} and R , are both known up to a multiplicative constant. As in the previous chapter, let us consider the distribution of u to be a prior distribution, and place an improper flat prior on β , i.e., $p(\beta) = 1$, independent of u . Then the posterior distribution of β and u based on y is proportional to the joint distribution of y and u , i.e., from (3.2) on page 65:

$$(2\pi)^{-\frac{1}{2}(n+c-1)} |R|^{-\frac{1}{2}} |\tilde{D}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(y - \eta(A\beta + Bu))^T R^{-1} (y - \eta(A\beta + Bu)) + u^T \tilde{D}^{-1} u] \right\}. \quad (5.8)$$

Following the pseudo-data approach of the previous chapter, the terms in (5.8) that depend on β and u can be written

$$\begin{aligned} & (y - \eta(A\beta + Bu))^T R^{-1} (y - \eta(A\beta + Bu)) + u^T \tilde{D}^{-1} u \\ &= \left[R^{-1/2} (y - \eta(A\beta + Bu)) \right]^T \left[R^{-1/2} (y - \eta(A\beta + Bu)) \right] + \left[\tilde{D}^{-1/2} u \right]^T \left[\tilde{D}^{-1/2} u \right] \\ &= \tilde{e}^T \tilde{e}, \end{aligned}$$

where $\tilde{e} = \tilde{y} - \tilde{\eta}(A\beta + Bu)$, with

$$\tilde{y} = \begin{bmatrix} R^{-1/2} y \\ 0 \end{bmatrix}, \quad \tilde{\eta}(A\beta + Bu) = \begin{bmatrix} R^{-1/2} \eta(A\beta + Bu) \\ \tilde{D}^{-1/2} u \end{bmatrix}, \quad \text{and } \tilde{e} \sim N(0, I).$$

As in the previous chapter, let $\xi = (\beta^T, u^T)^T$. The posterior mode $\tilde{\xi} = (\tilde{\beta}^T, \tilde{u}^T)^T$ may be obtained by solving the nonlinear least squares problem

$$\bar{y} = \bar{\eta}(A\beta + Bu) + \bar{e}, \quad (5.9)$$

i.e., minimizing the sum of squares $\bar{e}^T \bar{e}$ in β and u . Note that

$$\frac{\partial \bar{e}^T \bar{e}}{\partial \xi} = -2 \left(\frac{\partial \bar{\eta}^T}{\partial \xi} \right) [\bar{y} - \bar{\eta}(A\beta + Bu)].$$

Denote

$$\hat{X} = \left. \frac{\partial \eta}{\partial \beta^T} \right|_{\hat{\beta}, \hat{u}} = \eta'(A\hat{\beta} + B\hat{u})A$$

and

$$\hat{Z} = \left. \frac{\partial \eta}{\partial u^T} \right|_{\hat{\beta}, \hat{u}} = \eta'(A\hat{\beta} + B\hat{u})B.$$

Then

$$\frac{\partial \bar{\eta}}{\partial \xi^T} = \begin{bmatrix} R^{-1/2} \hat{X} & R^{-1/2} \hat{Z} \\ \mathbf{0} & \bar{D}^{-1/2} \end{bmatrix}.$$

At the posterior mode, $\partial \bar{e}^T \bar{e} / \partial \xi = 0$, so that

$$\begin{bmatrix} \hat{X}^T R^{-1/2} & \hat{Z}^T R^{-1/2} \\ \bar{D}^{-1/2} & \mathbf{0} \end{bmatrix} \begin{bmatrix} R^{-1/2}(y - \eta(A\hat{\beta} + B\hat{u})) \\ -\bar{D}^{-1/2} \hat{u} \end{bmatrix} = 0.$$

Defining

$$w = y - \eta(A\hat{\beta} + B\hat{u}) + \hat{X}\hat{\beta} + \hat{Z}\hat{u},$$

we have

$$\begin{bmatrix} \hat{X}^T R^{-1} \hat{X} & \hat{X}^T R^{-1} \hat{Z} \\ \hat{Z}^T R^{-1} \hat{X} & \hat{Z}^T R^{-1} \hat{Z} + \bar{D}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \hat{X}^T R^{-1} w \\ \hat{Z}^T R^{-1} w \end{bmatrix}, \quad (5.10)$$

which are the mixed model equations for a linear mixed effects model with response w and design matrices \hat{X} and \hat{Z} . Solving for $\hat{\beta}$ and \hat{u} gives

$$\hat{\beta} = (\hat{X}^T \hat{V}^{-1} \hat{X})^{-1} \hat{X}^T \hat{V}^{-1} w, \quad \text{and} \quad \hat{u} = \bar{D} \hat{Z}^T \hat{V}^{-1} (w - \hat{X} \hat{\beta}), \quad (5.11)$$

where

$$\hat{V} = \hat{Z}\hat{D}\hat{Z}^T + R.$$

Since w , \hat{X} , and \hat{Z} depend on $\hat{\beta}$ and \hat{u} , these must be solved iteratively.

A result that will prove useful later on can be obtained by considering the second derivative of the sum of squares:

$$\frac{\partial^2 \bar{e}^T \bar{e}}{\partial \xi \partial \xi^T} = -2 \left[\frac{\partial^2 \bar{\eta}^T}{\partial \xi \partial \xi^T} \right] \left[\bar{y} - \bar{\eta}(A\beta + Bu) \right] + 2 \left(\frac{\partial \bar{\eta}^T}{\partial \xi} \right) \left(\frac{\partial \bar{\eta}}{\partial \xi^T} \right). \quad (5.12)$$

The square brackets in the first term on the right hand side serve to highlight the fact that this product is not an ordinary matrix multiplication. The term $\frac{\partial^2 \bar{\eta}^T}{\partial \xi \partial \xi^T}$ is not a matrix, but an array of dimension $q \times m(c-1) \times m(c-1)$. The product is a matrix of dimension $m(c-1) \times m(c-1)$.

The second derivative (5.12) is required for the quadratic approximation used by the Newton-Raphson algorithm for nonlinear least squares estimation. The Gauss-Newton algorithm replaces the term $-2 \left[\frac{\partial^2 \bar{\eta}^T}{\partial \xi \partial \xi^T} \right] [\bar{y} - \bar{\eta}(A\beta + Bu)]$ by zero, which is its expectation. This is also known as *Fisher scoring*. Evaluating this approximation at $\hat{\xi} = (\hat{\beta}^T, \hat{u}^T)^T$ gives

$$\left. \frac{\partial^2 \bar{e}^T \bar{e}}{\partial \xi \partial \xi^T} \right|_{\hat{\beta}, \hat{u}} \approx 2 \begin{bmatrix} \hat{X}^T R^{-1} \hat{X} & \hat{X}^T R^{-1} \hat{Z} \\ \hat{Z}^T R^{-1} \hat{X} & \hat{Z}^T R^{-1} \hat{Z} + \bar{D}^{-1} \end{bmatrix}, \quad (5.13)$$

which is twice the matrix on the left hand side of the mixed model equations (5.10).

We have seen that the estimation of the fixed and random effects is relatively straightforward given the parameters ζ that determine \bar{D} and R . The more difficult part of the problem is the estimation of ζ . In the general notation of Chapter 3, our model is

$$\begin{aligned} y|\theta, \lambda &\sim N[\eta(\theta), R] \\ \theta|\lambda &\sim N(A\beta, B\bar{D}B^T), \end{aligned}$$

with $\lambda = (\beta^T, \zeta^T)^T$ and $\theta = A\beta + Bu$. From (3.3), the marginal density of the data is

$$p(y|\lambda) = \int p(y|\theta, \lambda) p(\theta|\lambda) d\theta, \quad (5.14)$$

where

$$p(y|\theta, \lambda) = (2\pi)^{-n/2} |R|^{-1/2} \exp \left\{ -\frac{1}{2} [(y - \eta(\theta))^T R^{-1} (y - \eta(\theta))] \right\} \quad (5.15)$$

and

$$p(\theta|\lambda) = (2\pi)^{-(c-1)/2} |B\bar{D}B^T|^{-1/2} \exp \left\{ -\frac{1}{2} [(\theta - A\beta)^T (B\bar{D}B^T)^{-1} (\theta - A\beta)] \right\}.$$

Unfortunately, because of the presence of the nonlinear function η in (5.15), there is no closed-form expression for this integral. Consequently a number of analytical and numerical approximations to it have been proposed. Pinheiro and Bates (1995) examined several of these approximations and compared their computational and statistical properties. They concluded that the approximation used by Lindstrom and Bates (1990) is reasonably accurate and computationally efficient. We now consider this approach.

5.2.1 Lindstrom and Bates' algorithm

Lindstrom and Bates (1990) proposed linearizing $\eta(A\beta + Bu)$ in (5.15) around estimates $\hat{\beta}$ and \hat{u} of β and u , i.e.,

$$\eta(A\beta + Bu) \approx \eta(A\hat{\beta} + B\hat{u}) + \hat{X}(\beta - \hat{\beta}) + \hat{Z}(u - \hat{u}).$$

Note that although $\hat{\beta}$ and \hat{u} are statistics (i.e., functions of y), we now treat them as if they were constants. Substituting the above approximation into (5.15), we have

$$y|u \sim N \left(\eta(A\hat{\beta} + B\hat{u}) + \hat{X}(\beta - \hat{\beta}) + \hat{Z}(u - \hat{u}), R \right), \quad (5.16)$$

so that the integral (5.14) can be evaluated to give an approximate marginal distribution of

$$y \sim N \left(\eta(A\hat{\beta} + B\hat{u}) + \hat{X}(\beta - \hat{\beta}) - \hat{Z}\hat{u}, \hat{V} \right). \quad (5.17)$$

As before, letting $w = y - \eta(A\hat{\beta} + B\hat{u}) + \hat{X}\hat{\beta} + \hat{Z}\hat{u}$, we can rewrite (5.16) and (5.17) as

$$w|u \sim N (\hat{X}\beta + \hat{Z}u, R),$$

and

$$w \sim N(\hat{X}\beta, \hat{V}).$$

These are the conditional and marginal distributions of a linear mixed effects model. The associated approximate marginal log likelihood is

$$\ell_{LB}(\beta, \zeta) = -\frac{1}{2} [(w - \hat{X}\beta)^T \hat{V}^{-1} (w - \hat{X}\beta) + \log |\hat{V}|]. \quad (5.18)$$

By analogy with the REML likelihood for the linear mixed effects model, (4.11), Lindstrom and Bates (1990) defined an approximate REML log likelihood

$$\ell_{RLB}(\beta, \zeta) = -\frac{1}{2} [(w - \hat{X}\hat{\beta})^T \hat{V}^{-1} (w - \hat{X}\hat{\beta}) + \log |\hat{V}| + \log |\hat{X}^T \hat{V}^{-1} \hat{X}|]. \quad (5.19)$$

Lindstrom and Bates (1990) proposed the following alternating two-step algorithm for approximate maximum-likelihood estimation. Given the k th estimate of ζ , denoted $\zeta^{(k)}$, perform the following steps:

Step 1. Pseudo-data step. With $\zeta = \zeta^{(k)}$, solve the nonlinear least squares problem (5.9) to obtain $\beta^{(k)}$ and $u^{(k)}$. Byproducts of this will be $\hat{X}^{(k)}$ and $\hat{Z}^{(k)}$.

Step 2. Linear mixed effects step. With $\hat{\beta} = \beta^{(k)}$, $\hat{u} = u^{(k)}$, $\hat{X} = \hat{X}^{(k)}$, and $\hat{Z} = \hat{Z}^{(k)}$, maximize (5.18) to obtain $\beta^{(k+1)}$ and $\zeta^{(k+1)}$.

The approximate REML version of their algorithm maximizes (5.19) instead of (5.18).

Example: Coho salmon Beverton-Holt mixed model

Using the model of Section 5.1.1, we have $\alpha_i = e^{\theta_{i1}}$ and $R_i^{\max} = e^{\theta_{i2}}$ with $\theta_{i1} = \beta_1 + b_i$ and $\theta_{i2} = \beta_2 + f_i$. The parameters β_1 and β_2 are treated as fixed effects and b_i and f_i as normally distributed random effects. Since there is no biological reason for suspecting that α_i and R_i^{\max} would be correlated, we set $\sigma_{bf} = 0$, so that

$$b_i \sim N(0, \sigma_b^2) \text{ independent of } f_i \sim N(0, \sigma_f^2).$$

The errors ϵ_{ij} are treated as i.i.d. $N(0, \sigma^2)$

The Lindstrom and Bates algorithm is available in S-PLUS in the `nlme` function. Estimates for the Beverton-Holt model using the approximate maximum likelihood version of the algorithm are shown at the bottom of Figure 5.1.

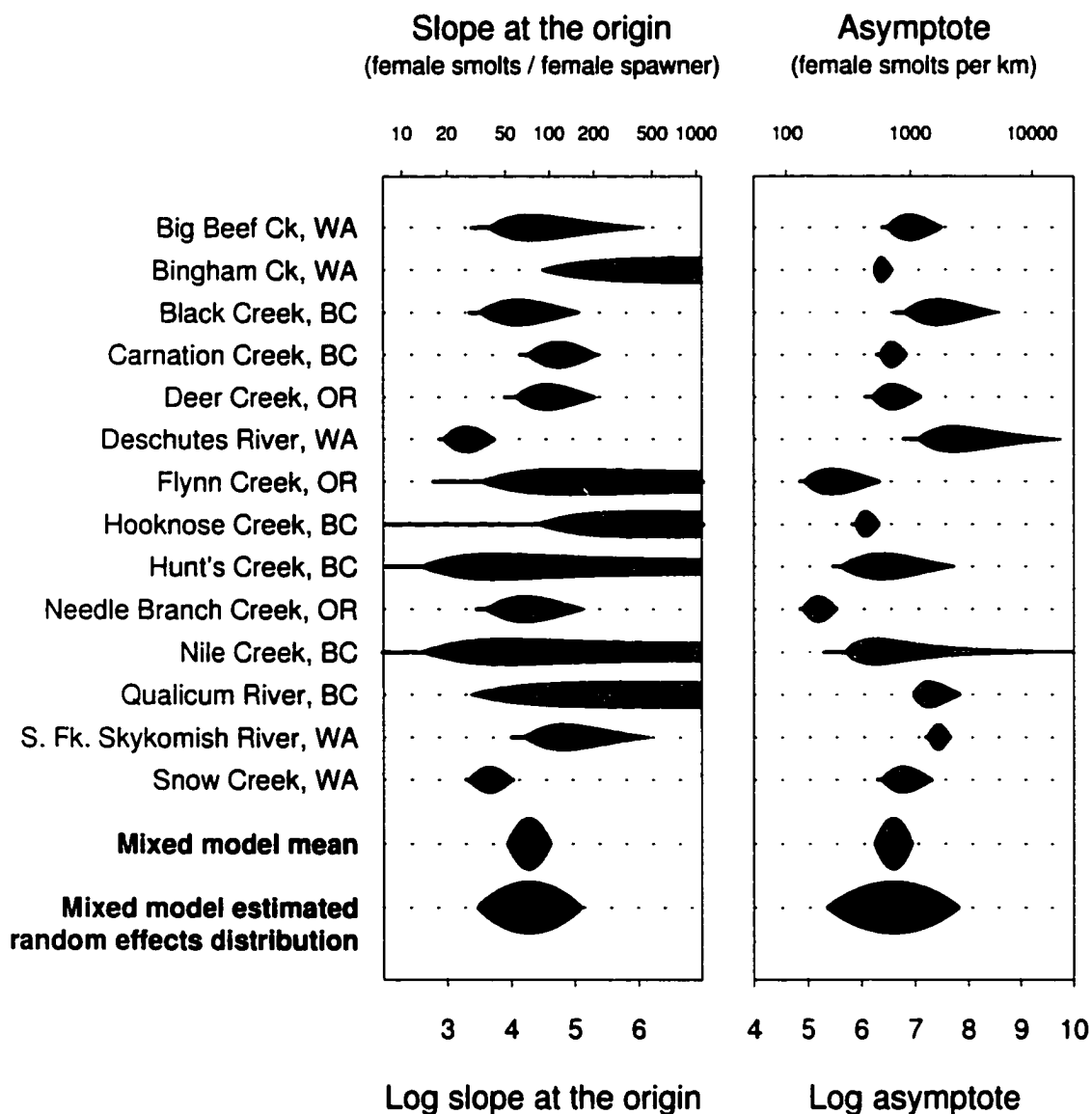


Figure 5.1: Side-by-side 95% raindrop plots for the parameters of the Beverton-Holt model for the 14 coho salmon populations with meta-analytic summaries at the bottom. The superimposed dots and error bars on the individual population raindrops are the maximum likelihood estimates obtained by individual nonlinear regression and approximate asymptotic 95% confidence intervals (based on nonlinear least squares theory). The meta-analytic summaries are from the approximate maximum likelihood version of the Lindstrom and Bates algorithm. The summary raindrops are shown doubly tall for emphasis.

In Figure 5.1, the “Mixed model mean” raindrops summarize the information from the mixed effects model concerning the two parameters. The “Mixed model estimated random effects distribution” raindrops are shaded darker to emphasize that they are HDR-raindrops (Section 3.5.1, p. 71) representing distributions rather than log likelihoods.

As discussed in the previous chapters, estimation of mixed effects models provides not only estimates of the fixed effects and variance components, but also estimates of the realized values of the random effects (empirical Bayes estimates). As shown in Figure 2.4 (p. 28), for two of the stocks there is no individual maximum likelihood fit: the slope at the origin is apparently arbitrarily large, a biological impossibility.

Let $\tilde{\theta}_{i1}$ and $\tilde{\theta}_{i2}$ be the empirical Bayes estimates of θ_{i1} and θ_{i2} . Then empirical Bayes spawner-recruitment curves are given by

$$R = \frac{S}{\exp(-\tilde{\theta}_{i1}) + S \exp(-\tilde{\theta}_{i2})}.$$

Figure 5.2 shows individual maximum likelihood curves and empirical Bayes curves.

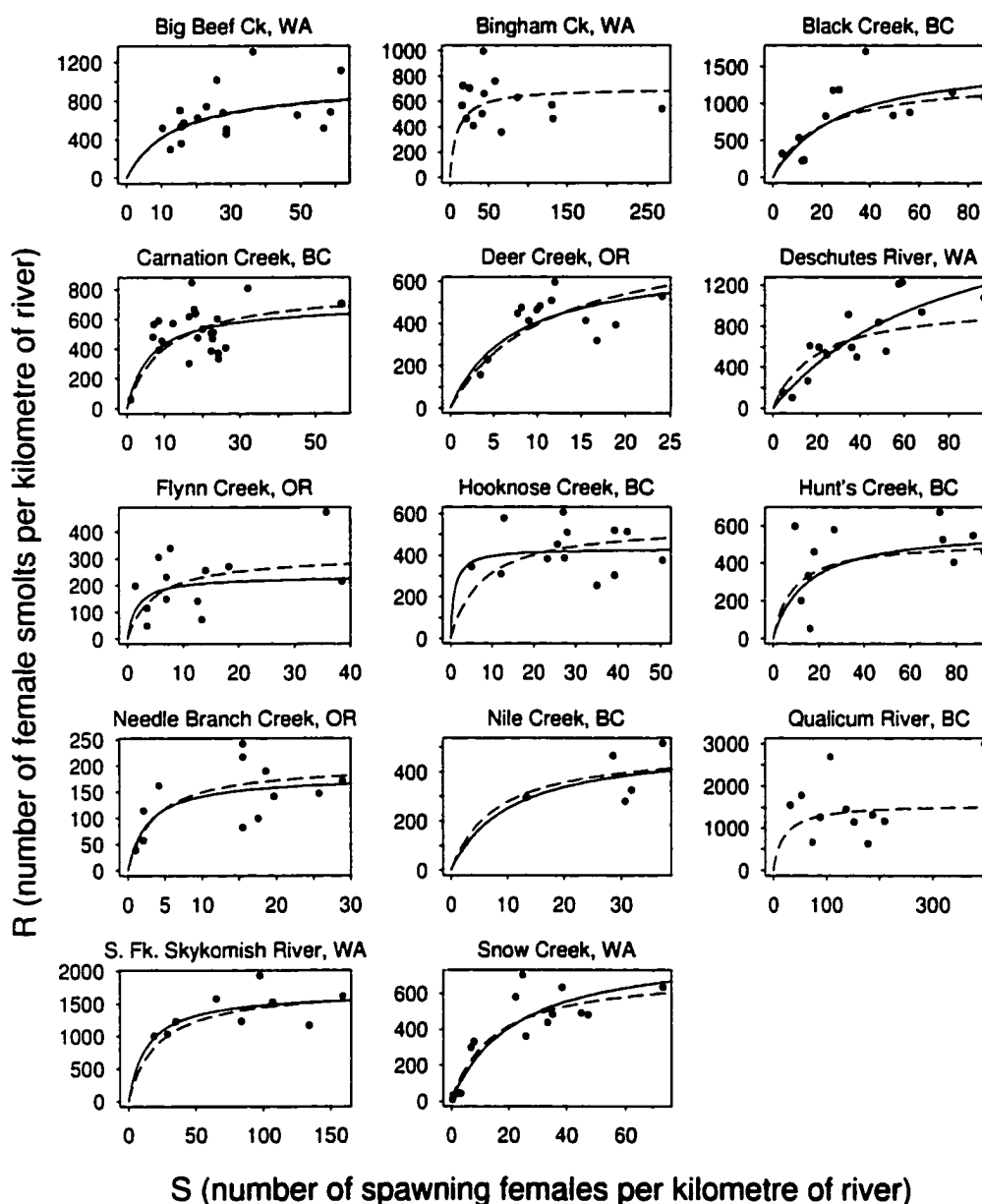


Figure 5.2: Coho salmon spawner-recruitment data with superimposed fitted Beverton-Holt curves. Solid lines are individual maximum likelihood fits; dashed lines are empirical Bayes curves from the parameter estimates obtained using the Lindstrom and Bates algorithm.

From Figure 5.2 we see that when an individual data set determines the fit well (e.g. for Big

Beef Creek), the empirical Bayes curve is very close to the individual maximum likelihood fit. For the two stocks with no individual maximum likelihood fits, the empirical Bayes curves seem quite reasonable. In cases where the data are relatively uninformative (e.g. for Hooknose Creek, where the slope at the origin is poorly determined, or for Deschutes River, where the asymptotic level is poorly determined), the empirical Bayes curves provide plausible fits by borrowing strength from the other stocks.

5.2.2 Wolfinger's derivation of an approximate REML likelihood

Recall from Chapter 4 that the REML likelihood for a linear mixed effects model can be derived by placing a flat prior on β and obtaining the marginal density of the data:

$$p(y|\zeta) = \int p(y|u, \beta, R)p(u|\bar{D})p(\beta)du d\beta, \quad (5.20)$$

with $p(\beta) \equiv 1$. Using this approach, Wolfinger (1993) showed that the REML version of the Lindstrom-Bates algorithm can also be derived using Laplace's approximation to the above integral:

$$\int e^{nf(\xi)} d\xi = (2\pi/n)^{p/2} | -f''(\hat{\xi}) |^{1/2} e^{nf(\hat{\xi})} \{1 + O(1/n)\},$$

where ξ is a p -dimensional vector and $\hat{\xi}$ maximizes $e^{nf(\xi)}$. Note that Laplace's approximation replaces integration by maximization, a strategy we have previously seen in two other contexts. One was in the elimination of nuisance parameters in Chapter 2, where profile likelihood may be used instead of integrated likelihood. The second was in the empirical Bayes strategy of replacing integration over the posterior distribution of the hyperparameter by substitution of the maximum likelihood estimate of the hyperparameter.

Here $\xi = (\beta^T, u^T)^T$ and $e^{nf(\xi)} = p(y|u, \beta, R)p(u|\bar{D})$, which is the joint distribution of y and u , as in (5.8). Thus $\hat{\xi}$ is the posterior mode $(\hat{\beta}, \hat{u})$. Using the Gauss-Newton approximation (5.13) followed by some matrix manipulations,

$$\begin{aligned} | -f''(\hat{\xi}) | &\approx \left| \begin{array}{cc} \hat{X}^T R^{-1} \hat{X} & \hat{X}^T R^{-1} \hat{Z} \\ \hat{Z}^T R^{-1} \hat{X} & \hat{Z}^T R^{-1} \hat{Z} + \bar{D}^{-1} \end{array} \right| \\ &= | \hat{X}^T \hat{V}^{-1} \hat{X} |. \end{aligned}$$

Hence

$$-2 \log p(y|\zeta) \approx \log |\hat{V}| + (y - \eta(A\hat{\beta} + B\hat{u}))^T R^{-1} (y - \eta(A\hat{\beta} + B\hat{u})) + \log |\hat{X}^T \hat{V}^{-1} \hat{X}| + \hat{u}^T \bar{D}^{-1} \hat{u} + (n - q) \log 2\pi + (q + c - 1) \log n.$$

Substituting the expressions (5.11) for $\hat{\beta}$ and \hat{u} gives

$$\log p(y|\zeta) \approx -\frac{1}{2} \left[(w - \hat{X}\hat{\beta})^T \hat{V}^{-1} (w - \hat{X}\hat{\beta}) + \log |\hat{V}| + \log |\hat{X}^T \hat{V}^{-1} \hat{X}| + (n - q) \log 2\pi + (q + c - 1) \log n \right].$$

Up to an additive constant, this is the same as the approximate REML log likelihood in (5.19).

5.2.3 Another Laplacian approximation

Pinheiro and Bates (1995) obtained an approximate log likelihood using Laplace's approximation. Independently Vonesh (1996) obtained the same approximation. Their approach differs from that of Wolfinger (1993) in that β is not integrated out of the likelihood, i.e., the integrand is expanded only around \hat{u} . For the purposes of this section, we redefine \hat{Z} so that it is evaluated at $\hat{\beta}$ instead of $\hat{\beta}$, i.e.,

$$\hat{Z} = \left. \frac{\partial \eta}{\partial u^T} \right|_{\beta, \hat{u}} = \eta'(A\hat{\beta} + B\hat{u})B$$

and as before $\hat{V} = \hat{Z}\bar{D}\hat{Z}^T + R$. The approximate log likelihood is

$$\log p(y|\beta, \zeta) \approx -\frac{1}{2} \left[(w - \eta(A\hat{\beta} + B\hat{u}))^T \hat{V}^{-1} (w - \eta(A\hat{\beta} + B\hat{u})) + \log |\hat{V}| + n \log 2\pi \right],$$

with w redefined as $w = y + \hat{Z}\hat{b}$.

5.2.4 Asymptotic results

Vonesh (1996) examined the asymptotic properties of approximate maximum likelihood estimates obtained using the Laplacian approximation of Section 5.2.3. He found that the estimates are consistent to order

$$O_p \left[\max \left\{ m^{-1/2}, \min(n_i)^{-1} \right\} \right].$$

In other words, the order of accuracy depends both on the number of studies, m , and on the number of observations per study, n_i . Intuitively, the dependence on the n_i comes about because for the i th study, we only have n_i observations providing information about the realized values of the random effects u_i . For small n_i , the estimate \hat{u}_i may be poor, so that Laplace's approximation may not work well. Vonesh's result tells us that for consistency, we must have both $n_i \rightarrow \infty$ for each i , and $m \rightarrow \infty$. For meta-analysis, this may be conceptually problematic. While we can imagine accumulating information from more and more studies, the smallest number of observations per study may always be limited. In real examples we often have *both* a small number of studies *and* a small number of observations per study. Sometimes our desire is explicitly to combine information from large and small studies.

5.2.5 Nuñez's approach (SPML)

A quite different approach to estimation for nonlinear mixed effects models has been proposed by Nuñez (1998). It has its origins in methods of simulated-based inference (McFadden 1989; Gourieroux and Monfort 1993) and pseudo-likelihood (Gourieroux, Monfort, and Trognon 1984) developed in the econometrics literature. Recall from the start of Section 5.2 that the ELS estimator minimizes

$$(y - \mu)^T V^{-1} (y - \mu) + \log |V|, \quad (5.21)$$

where $\mu = E(y)$ and $V = \text{Var}(y)$. When y is normally distributed, (5.21) is the kernel of the log likelihood function, but in general we do not expect y to be normal. The ELS estimator is an example of a *pseudo maximum likelihood* estimator, i.e., one obtained by maximizing a likelihood function associated with a family of probability distributions which does not

necessarily contain the data generating distribution (White 1982; Gouriéroux, Monfort, and Trognon 1984). White (1982) notes that “in many (if not most) circumstances, one may not have complete confidence” that one’s parametric probability model is correctly specified, or as Box (1979) pointed out “All Models Are Wrong But Some Are Useful.” This means that almost any application of maximum likelihood estimation is in fact pseudo maximum likelihood.

The Kullback-Leibler Information Criterion

We now formalize these ideas, following White (1994, Section 2.3). The presentation is quite general here; details more specific to the estimation of nonlinear mixed effects models are subsequently developed.

In statistical modeling, a key idea is that of approximating the data generating distribution as closely as possible. This requires some criterion for evaluating the discrepancy between probability distributions. One such criterion is the *Kullback-Leibler Information Criterion* (KLIC). If f and g are probability densities, the KLIC is

$$I(f : g) \equiv E_f \left[\log \frac{f}{g} \right],$$

where E_f denotes the expectation with respect to f . An important property of the KLIC is known as the *information inequality*: $I(f : g) \geq 0$ and $I(f : g) = 0$ if and only if $f = g$ almost everywhere. From an information-theoretic perspective, the KLIC can be interpreted as the surprise experienced on average when we believe that g describes a given phenomenon and we are then informed that in fact the phenomenon is described by f .

Suppose $y_i \sim f_i$ independently for $i = 1, \dots, m$ and consider a family of distributions $\{p(y_i|\lambda) : \lambda \in \Lambda\}$. Note that the family of distributions is generally chosen to make computations feasible. Let $f^m = \prod_{i=1}^m f_i$. The pseudo likelihood $p^m = \prod_{i=1}^m p(y_i|\lambda)$ (called a quasi likelihood by White) can be viewed as an approximation to f^m . One way to measure

the adequacy of the approximation is by using the KLIC,

$$\begin{aligned}
 I(f^m : p^m; \lambda) &\equiv E_{f^m} \left[\log \frac{f^m(y)}{p^m(y|\lambda)} \right]. \\
 &= \int \log \frac{f^m(y)}{p^m(y|\lambda)} f^m(y) dy \\
 &= \int \log[f^m(y)] f^m(y) dy - \int \log[p^m(y|\lambda)] f^m(y) dy.
 \end{aligned}$$

Choosing λ to minimize I is equivalent to choosing λ to maximize the second term on the right hand side, which is

$$E_{f^m} [\log p^m(y|\lambda)]. \quad (5.22)$$

When $p^m(y|\lambda)$ is correctly specified in the sense that $p^m(y|\lambda_0) = f^m(y)$ for a unique vector λ_0 in Λ , then choosing λ to maximize (5.22) yields λ_0 by the information inequality.

In practice, however, λ cannot be chosen in this way because the expectation in (5.22) cannot be evaluated without knowing the data generating distribution f^m . Instead, by approximating the expectation using sample information, an approximate solution may be obtained. Note that maximizing (5.22) is equivalent to maximizing

$$E_{f^m} \left[\frac{1}{m} \log p^m(y|\lambda) \right].$$

Also, note that the pseudo likelihood is given by

$$\tilde{\ell}(\lambda) = \frac{1}{m} \log p^m(y|\lambda) = \frac{1}{m} \sum_{i=1}^m \log p(y_i|\lambda).$$

Provided that a law of large numbers applies to the sums on the right hand side, it follows that for sufficiently large m , $m^{-1} E[\log p^m(y|\lambda)]$ will be well approximated by $\tilde{\ell}(\lambda)$. Hence the value of λ that provides the best approximation to f^m might be well approximated by the value $\hat{\lambda}$ which maximizes $\tilde{\ell}(\lambda)$. We now return to our parametric model, to illustrate these ideas in the specific context of interest.

Pseudo maximum likelihood for the nonlinear mixed effects model

We will suppose that λ^* is a value of $\lambda = (\beta^T, \zeta^T)^T$ such that

$$E_{f^m}(y) = \mu(\lambda^*) \text{ and } \text{Var}_{f^m}(y) = V(\lambda^*).$$

In other words, at λ^* our model matches the first two moments of the observations. Consider the ELS criterion (5.21), expressed in terms of the m studies:

$$\bar{c}_m(\lambda) = \frac{1}{m} \sum_{i=1}^m c_i(\lambda) \quad (5.23)$$

where

$$c_i(\lambda) = [y_i - \mu_i(\lambda)]^T V_i(\lambda)^{-1} [y_i - \mu_i(\lambda)] + \log |V_i(\lambda)|,$$

with $\mu_i(\lambda) = E_{\lambda}(y_i)$ and $V_i(\lambda) = \text{Var}_{\lambda}(y_i)$. Note that

$$E_{\lambda^*}[\bar{c}_m(\lambda)] = \frac{1}{m} \sum_{i=1}^m E_{\lambda^*} \{ [y_i - \mu_i(\lambda)]^T V_i(\lambda)^{-1} [y_i - \mu_i(\lambda)] + \log |V_i(\lambda)| \}.$$

But because $[y_i - \mu_i(\lambda)]^T V_i(\lambda)^{-1} [y_i - \mu_i(\lambda)]$ is a quadratic function, its expectation depends only on the first two moments of y_i . Therefore

$$E_{\lambda^*} \{ [y_i - \mu_i(\lambda)]^T V_i(\lambda)^{-1} [y_i - \mu_i(\lambda)] + \log |V_i(\lambda)| \} = \int \log[\phi(y_i, \lambda)] \phi(y_i, \lambda^*) dy_i,$$

where $\phi(y_i, \lambda)$ denotes the multivariate normal pdf with mean $\mu_i(\lambda)$ and variance-covariance matrix $V_i(\lambda)$. But from the information inequality,

$$\int \log[\phi(y_i, \lambda)] \phi(y_i, \lambda^*) dy_i \geq \int \log[\phi(y_i, \lambda^*)] \phi(y_i, \lambda^*) dy_i.$$

Therefore,

$$E_{\lambda^*}[\bar{c}_m(\lambda)] \geq E_{\lambda^*}[\bar{c}_m(\lambda^*)]. \quad (5.24)$$

In other words, the expected ELS criterion is minimized by λ^* .

The difficulty remains that there are no explicit forms for the means

$$\mu_i(\lambda) = E_\lambda(y_i) = E_\lambda[\eta_i(A_i\beta + B_i u_i)]$$

and the variance-covariance matrices

$$V_i(\lambda) = \text{Var}_\lambda(y_i) = \text{Var}_\lambda[\eta_i(A_i\beta + B_i u_i)] + R_i(\lambda).$$

Simulated pseudo maximum likelihood

Nuñez (1998) proposed approximating $\mu_i(\lambda)$ and $V_i(\lambda)$ by the Monte-Carlo sums

$$\bar{\mu}_{i,g}(\lambda) = \frac{1}{g} \sum_{k=1}^g \eta_i(A_i\beta + B_i u_i^k)$$

and

$$V_{i,g}(\lambda) = \frac{1}{g-1} \sum_{k=1}^g \left[\eta_i(A_i\beta + B_i u_i^k) - \bar{\mu}_{i,g}(\lambda) \right] \left[\eta_i(A_i\beta + B_i u_i^k) - \bar{\mu}_{i,g}(\lambda) \right]^T + R_i(\lambda),$$

where $u_i^k = D^{1/2} e_i^k$ and the e_i^k are drawn independently from $N(0, I)$. The simulated objective function is thus

$$\bar{c}_m^g(\lambda) = \frac{1}{m} \sum_{i=1}^m c_i^g(\lambda) \quad (5.25)$$

where

$$c_i^g(\lambda) = [y_i - \bar{\mu}_{i,g}(\lambda)]^T V_{i,g}(\lambda)^{-1} [y_i - \bar{\mu}_{i,g}(\lambda)] + \log |V_{i,g}(\lambda)|.$$

Nuñez calls the estimator that minimizes (5.25) the *simulated pseudo maximum likelihood* (SPML) estimator.

Under a number of assumptions, Nuñez proved that, almost surely,

$$\bar{c}_m^g(\lambda) - E_{\lambda \cdot}[\bar{c}_m(\lambda)] \longrightarrow 0, \text{ as } m \rightarrow \infty \text{ and } g \rightarrow \infty \text{ uniformly on } \Lambda. \quad (5.26)$$

This (strong) uniform convergence is a critical step in Nuñez's argument: it shows that with large enough m and g , the simulated objective function gets arbitrarily close to the

expectation of the non-simulated objective function we are trying to approximate (with probability 1). There are in fact two parts to this: one part shows that for large enough g , it does not matter that we are actually simulating the objective function; the other part shows that for large enough m the objective function approaches its expectation. The required assumptions are that λ^* is an interior point of a compact metric space, that the observed vectors y_i are uniformly bounded, that the first four moments of $c_i^g(\lambda)$ are finite for all g , and that the η_i are almost surely twice continuously differentiable and square integrable.

The above assumptions are fairly reasonable in general and should be satisfied, for example, in the case of the Beverton-Holt spawner-recruitment model. Note however that the derivative of the hockey stick spawner-recruitment model is not continuous. Even for a single stock, estimation of the hockey stick model is problematic (Barrowman and Myers 2000). Difficulties encountered with fitting nonlinear mixed effects models using the hockey stick were one of the motivations for developing the generalized hockey stick models (Chapter 2). The quadratic hockey stick (2.14, p. 59) has a continuous first derivative, however its second derivative is discontinuous, thus Nuñez's asymptotics do not apply. On the other hand, the logistic hockey stick (2.17, p. 61) has all derivatives continuous. This is important, not just for asymptotic reasons: smoothness is essential in order to avoid numerical difficulties.

Denote the value of λ that minimizes $\bar{c}_m^g(\lambda)$ by $\hat{\lambda}_m^g$. To prove strong consistency, Nuñez made an additional assumption of second-order identifiability, i.e.,

$$\left. \begin{array}{l} \mu_i(\lambda^*) = \mu_i(\lambda) \\ V_i(\lambda^*) = V_i(\lambda) \end{array} \right\} \iff \lambda = \lambda^*$$

With this additional assumption, (5.24) and (5.26) allowed Nuñez to prove that $\hat{\lambda}_m^g$ is strongly consistent for λ^* .

Finally, to prove asymptotic normality, Nuñez made a further assumption that the gradient of $c_i^g(\lambda^*)$ satisfies the Liapounov condition. This condition is required in order to apply the Central Limit Theorem in cases where the summands are not identically distributed; for example Richardson and Welsh (1994) showed that this condition follows from their assumptions in order to prove asymptotic normality of REML estimators for linear mixed effects models. With this assumption, Nuñez proved that $\hat{\lambda}_m^g$ is asymptotically normal (provided that $m^{1/2}/g \rightarrow 0$ as m and g tend to infinity).

Comparison with the algorithm of Lindstrom and Bates

In Section 5.2.4, we noted that the asymptotic properties of algorithms based on using Laplace's approximation may be of concern. The algorithm of Lindstrom and Bates (1990) is closely related to methods involving the Laplace approximation. Since Lindstrom and Bates' algorithm involves estimation of realized values of the study-specific random effects, the number of observations per study, n_i , will strongly affect the accuracy of the estimates, so that the quality of the approximation to the marginal likelihood may be limited. The asymptotic properties of SPML seem preferable because there is no dependence on the study sizes n_i .

Núñez and Concordet (2000) conducted simulations to compare the performance of several estimation methods for some simple nonlinear mixed effects models. SPML consistently outperformed that of Lindstrom and Bates.

SPML does have several drawbacks compared to that of Lindstrom and Bates. It is computationally intensive since each iteration requires recomputing mean vectors and variance-covariance matrices. For example, for the 14 populations in the coho salmon dataset, a typical population has 15 observations, so we need to compute a 15×15 variance-covariance matrix based on, say, 500 simulated response vectors. This imposes a substantial computational burden.

Second, unlike SPML, Lindstrom and Bates' algorithm provides approximate BLUPs as a natural by-product of the estimation procedure. One possible approach to obtaining approximate BLUPs from SPML would be to substitute parameter estimates into the approximate BLUP formula $\hat{u} = \bar{D}\hat{Z}^T\hat{V}^{-1}(w - \hat{X}\hat{\beta})$ from the Lindstrom and Bates algorithm.

Finally, the Lindstrom and Bates algorithm leads to a natural definition of an approximate REML likelihood. It is not clear how to define REML estimation using SPML.

Example: Coho salmon

In Section 5.2.1, we illustrated the Lindstrom and Bates algorithm by fitting a Beverton-Holt nonlinear mixed effects model to the coho salmon data. We now try fitting the same model using SPML. As the number of simulated values, g , increases, we expect to see convergence of the parameter estimates. We call a graph of the parameter estimate versus g a *convergence profile*. Convergence profiles for each parameter in the model are shown

in Figure 5.3.

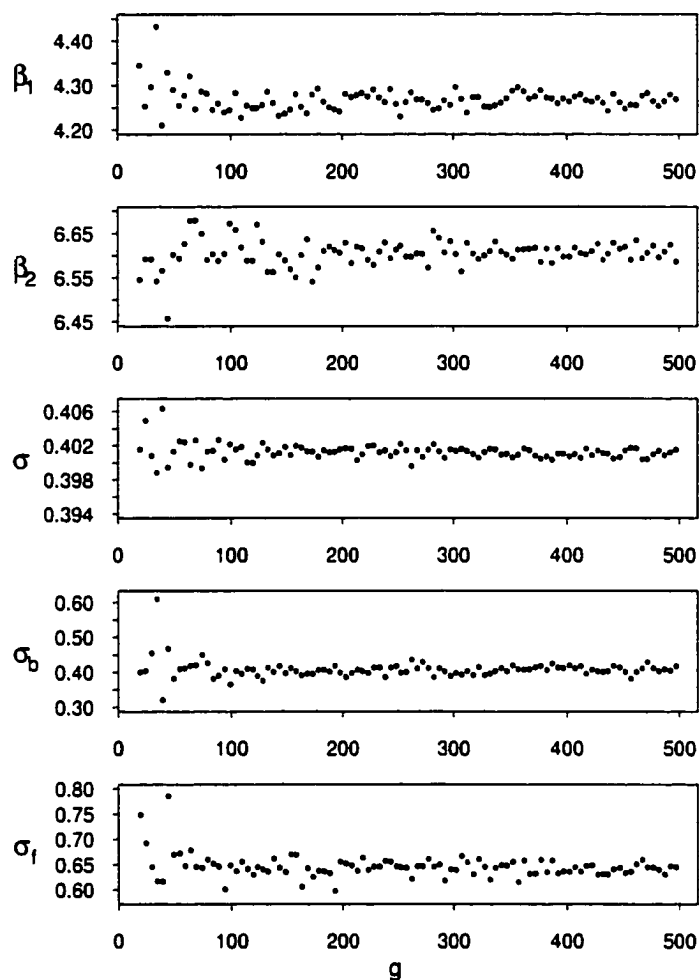


Figure 5.3: Convergence profiles for each parameter in a Beverton-Holt nonlinear mixed effects model for the coho salmon data.

Note that the parameters seem to be converging, though perhaps a little slowly. In one of the models considered by Nuñez and Concordet (2000), they find that a value of $g = 250$ provides acceptable accuracy. In Figure 5.3, we see that even for $g \approx 500$, there is some variability in the estimates. The explanation for this is that each data point represents a separate estimate using different simulated random numbers. The estimation algorithm therefore injects additional variability into the estimates, which is undesirable.

5.2.6 A modification of SPML: Stylized normal samples

To remove the additional variability we saw in the estimates in Figure 5.3, we propose the use of *stylized normal samples* instead of simulated normal samples. In the one-dimensional case, a stylized normal sample is the set of normal quantiles that are used in quantile-quantile plots. The k th element of a stylized normal sample of size g is

$$\Phi^{-1}\left(\frac{k-a}{g+1-2*a}\right),$$

where Φ is the cdf of the standard normal distribution, and a is a continuity correction commonly taken as $\frac{1}{2}$ or $\frac{3}{8}$.

For the Beverton-Holt model, we have two random effects, and therefore need to generalize the notion of stylized normal samples to the bivariate case. To this end, suppose X and Y are random variables whose joint distribution is standard bivariate normal. Their density is

$$f_{XY}(x, y) = \frac{1}{2\pi} e^{-(x^2/2)-(y^2/2)}. \quad (5.27)$$

Switching to polar coordinates by defining $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$, the random variables R and Θ have joint density

$$f_{R\Theta}(r, \theta) = \frac{1}{2\pi} r e^{-r^2/2}. \quad (5.28)$$

Thus R and Θ are independent, Θ is uniformly distributed on $[0, 2\pi]$, and R has the Rayleigh density

$$f_R(r) = r e^{-r^2/2}. \quad (5.29)$$

The cdf of R is

$$F_R(r) = \int_0^r u e^{-u^2/2} du = 1 - e^{-r^2/2}. \quad (5.30)$$

The inverse of the cdf of R is obtained by solving

$$p = 1 - e^{-r^2/2} \quad (5.31)$$

for r , giving

$$F_R^{-1}(p) = \sqrt{-2 \log(1 - p)}. \quad (5.32)$$

A stylized bivariate normal sample with $g = t^2$ elements may thus be obtained by selecting t angles

$$\theta = \frac{2\pi k}{t}, \quad k = 1, \dots, t, \quad (5.33)$$

and for each angle, spacing points along a radial line from the origin, with spacings

$$F_R^{-1}\left(\frac{1-a}{t+1-2*a}\right), F_R^{-1}\left(\frac{2-a}{t+1-2*a}\right), \dots, F_R^{-1}\left(\frac{t-a}{t+1-2*a}\right). \quad (5.34)$$

We call the points along a radial line “arms” and refer to this as the radial approach.

The stylized bivariate normal sample with $a = \frac{3}{8}$ for $t = 10$ (i.e. $g = 100$) is shown in Figure 5.4, together with a normal probability plot of the x -component of the sample.

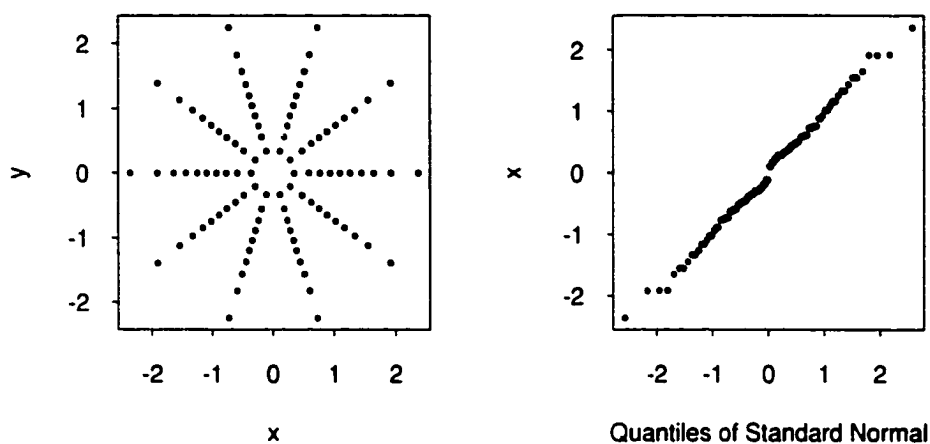


Figure 5.4: Left panel: Stylized bivariate normal sample of size 100. Points lie on ten radial lines emanating from the origin. Along each line the points are spaced according to the quantiles of a Rayleigh distribution. Right panel: Normal probability plot of the x -component of the stylized sample shown in the left panel.

The normal probability plot shows that marginally the stylized sample is not quite right: there is a vertical gap at 0 and slightly erratic behaviour in the tails. It seems that the stylized bivariate normal sample doesn't cover the plane "evenly enough." To improve this, I also tried generating "spiraling" arms by smoothly changing the angle along each arm. That is, the angle for the p th point along the k th arm would be

$$\theta = \frac{2\pi(k + p/t)}{t}.$$

The result is shown in Figure 5.5.

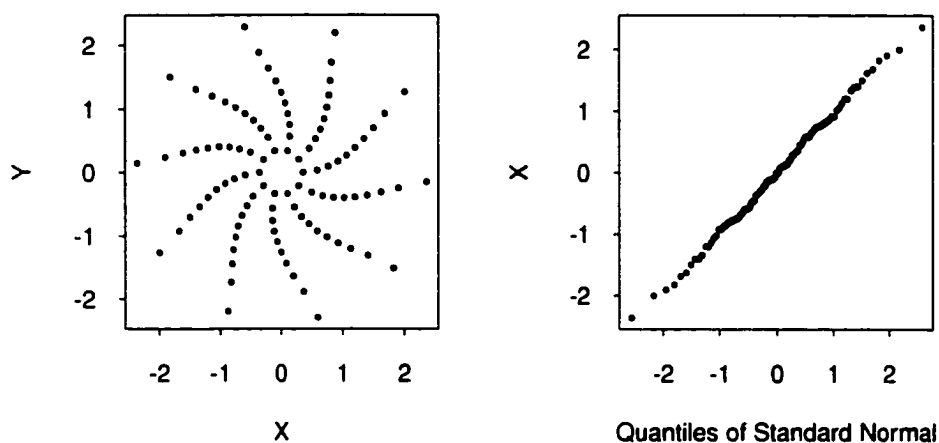


Figure 5.5: Left panel: Alternative stylized bivariate normal sample of size 100. Points lie on ten radial curves emanating from the origin. Along each curve the points are spaced according to the quantiles of a Rayleigh distribution. Right panel: Normal probability plot of the x -component of the stylized sample shown in the left panel.

For small g , the alternative approach seems to consistently give more linear marginal normal probability plots. For large g , both approaches result in “shoulders” at the extremes.

Example: Coho salmon

Using the stylized bivariate normal samples discussed above, SPML shows much less variability. Convergence profiles for each parameter in the model are shown in Figure 5.6.

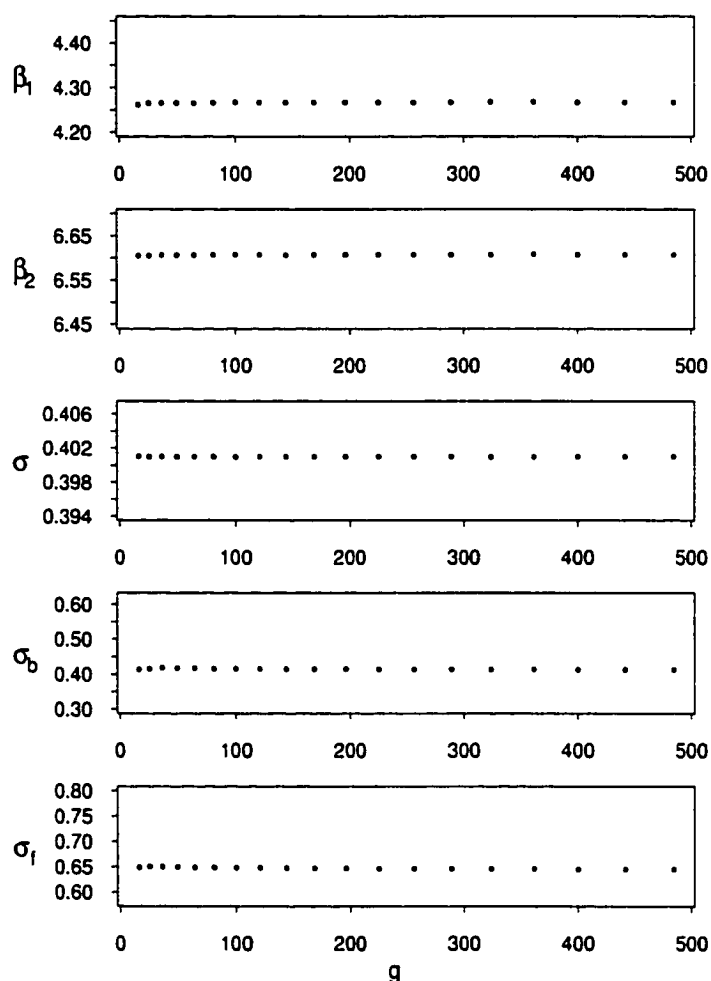


Figure 5.6: Convergence profiles for each parameter in a Beverton-Holt nonlinear mixed effects model for the coho salmon data. The vertical scale in each panel is the same as in Figure 5.3.

5.2.7 Uncertainty estimates

Asymptotic theory for SPML provides estimates of standard errors for the parameter estimates. Approximate standard errors are also available in the Lindstrom and Bates algorithm using the Hessian of the approximate likelihood function. Lindstrom and Bates (1990) note, however, that these uncertainty estimates can be quite inaccurate. Furthermore, uncertainty estimates for the robust methods developed later in this chapter may be difficult to obtain.

We therefore propose to use resampling methods to estimate standard errors.

Parametric bootstrap

A simple approach is to use a *parametric bootstrap* procedure: we take the independent variables (e.g. for the coho salmon data, the observed spawner quantities) as fixed, and build “pseudo observations” by repeatedly sampling from the model distribution using the parameter estimates as the “true” values. In our context, the model distribution involves variation at both the between-studies and the within-studies levels. The parametric bootstrap approach provides a rough assessment of the uncertainty in our parameter estimates, but depends on the model assumptions, e.g., that the random effects and errors are normally distributed. Nonparametric resampling approaches therefore seem more attractive.

Nonparametric resampling approaches

Davison and Hinkley (1997) discuss nonparametric resampling approaches for repeated measures data using the simple example of a balanced one-way ANOVA model. They note that “it may be important to take careful account of the two (or more) sources of variation when setting up a resampling scheme.” Paraphrasing their approach in meta-analytic language, they suggest first resampling studies and then resampling observations within studies either with or without replacement. Consideration of the first two moments of the resampled data leads Davison and Hinkley to conclude that resampling of observations within studies should be performed *without* replacement to best reproduce within-study correlations. Simulations performed by Wong (1999), however, suggest that there may be little to choose between the two procedures. For more complex datasets, e.g. in our context, unbalanced longitudinal nonlinear regression data, such multi-level resampling is more complex and little has been published on the subject. Das and Krishen (1999) discuss nonparametric resampling approaches for estimating standard errors in nonlinear mixed effects models for repeated measures data, however they do not deal with the issue of multiple sources of variation. Below are outlined several nonparametric resampling approaches that may be considered.

Jackknife

The earliest resampling method is the *jackknife*, proposed by Quenouille (1949), in which observations are dropped one at a time from the dataset, and the parameter estimates re-computed using each deleted dataset in turn. Efron (1996) used the jackknife to estimate standard errors in his empirical Bayes approach (Chapter 3), noting that this “amounts to thinking of the cases ... as being randomly sampled from some superpopulation.” In a meta-analytic context, the cases are studies, and the superpopulation is all “similar” studies that we can imagine having been performed. However this approach only deals with variability at the between-studies level.

Multi-halver jackknife

Generalizations of the jackknife that involve multiple case deletion have also been proposed. For example, *half-sampling* methods (McCarthy 1969) are based on taking structured sub-samples (blocks) of data sets. Tukey (1987) has advocated the use of these methods, calling them the *multihalver jackknife*. The idea is that if the data can be divided into blocks of two observations each then a subsample may be formed by leaving out one of the observations in each pair. For example consider a data set consisting of 8 observations occurring in pairs, denoted

$$(AB)(CD)(EF)(GH).$$

The subsample *ADEG* has one observation from each pair. By taking the other observation from each pair, we obtain the subsample *BCFH*. We call *ADEG* the *left-hand half* and *BCFH* the *right hand half*. This halving can be represented by the sequence

$$+ - ++$$

where “+” means take the first observation in a pair in the left-hand half and the second observation in a pair in the right-hand half, and “-” means the converse. Instead of using all possible halvings, Tukey suggests using orthogonal sequences of +’s and -’s, e.g. the

four rows

++++
 ++--
 +-+-
 +--+

have the property that any two rows agree in two positions and disagree in two positions. Jackknife formulas based on this approach can be used to estimate uncertainty of parameter estimates. For longitudinal studies such as the wolf and salmon datasets, this seems a reasonable approach to resampling while preserving the structure of the data.

Marginal bootstrap

In the jackknife approaches described above, the observed data are resampled in a structured way. An attractive alternative is to resample residuals rather than observed data. However, in the mixed model setting the definition of residuals is not entirely clear. Recall that our model for the i th study is

$$y_i = \eta_i(\theta_i) + \varepsilon_i,$$

with θ_i depending on fixed and random effects:

$$\theta_i = A_i\beta + B_iu_i.$$

The marginal mean and variance-covariance of y_i is

$$\mu_i(\lambda) = E_\lambda(y_i) = E_\lambda[\eta_i(A_i\beta + B_iu_i)]$$

and

$$V_i(\lambda) = \text{Var}_\lambda(y_i) = \text{Var}_\lambda[\eta_i(A_i\beta + B_iu_i)] + R_i.$$

Using the SPML algorithm, we can obtain an estimate, $\hat{\lambda}$, of λ . Denote $\hat{\mu}_i = \mu_i(\hat{\lambda})$ and $\hat{V}_i = V_i(\hat{\lambda})$.

For the i th study, a naive way to define a residual vector is

$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i.$$

These residuals can be bootstrapped as follows. Within each study, sample the elements in $\hat{\varepsilon}_i$ with replacement to obtain $\hat{\delta}_i^*$. Then compute the bootstrap sample for that study as

$$y_i^* = \hat{\mu}_i + \hat{\varepsilon}_i^*.$$

However, this approach is logically flawed. To see this, note that

$$\hat{\varepsilon}_i = y_i - \hat{\mu}_i = \eta_i(A_i\beta + B_i u_i) - \hat{\mu}_i + \varepsilon_i,$$

so these residuals include both random effect variability (from the u_i 's) and error variability (from the ε_i 's). Resampling these residuals ignores their marginal variance-covariance structure.

The solution is to define a *standardized* marginal residual vector

$$\hat{\omega}_i = \hat{V}_i^{-1/2}(y_i - \hat{\mu}_i),$$

which is an estimate of

$$\omega_i = V_i^{-1/2}(y_i - \mu_i).$$

Note that the marginal expectation of ω_i is zero and the marginal variance-covariance of ω_i is the identity matrix.

A nonparametric bootstrap procedure can be formulated as follows. Within each study, sample the elements in $\hat{\omega}_i$ with replacement to obtain $\hat{\omega}_i^*$. Then compute the bootstrap sample for that study as

$$y_i^* = \hat{\mu}_i + \hat{V}_i^{1/2} \hat{\omega}_i^*.$$

Note that since V_i is the *marginal* variance-covariance, it incorporates both within-study and between-study variation. We refer to it as a “marginal” bootstrap since it uses the estimated marginal moments of the data.

A modification of the above procedure would be to pool the $\hat{\omega}_i$'s across the studies to obtain $\hat{\omega}$ and then resample the $\hat{\omega}_i^*$'s from $\hat{\omega}$. It is not clear whether this would be preferable.

A different modification of the above procedure is to incorporate resampling of studies as well as within-study residuals. Each bootstrap sample is obtained by first sampling with replacement from the studies and then within each selected study, resampling the standardized marginal residuals as described above. Resampling the studies treats the configuration of each study (including factors such as sample size and covariates) as coming from a distribution rather than being fixed. Depending on the context, this "type-II marginal bootstrap" might be more appropriate.

Resampling approaches for the Coho salmon models

Following the discussion above, we propose to assess the uncertainty of parameter estimates for the Coho salmon models using five different approaches: (1) a parametric bootstrap approach; (2) a jackknife approach in which studies are left out one at a time; (3) a multi-halver jackknife approach in which the data are repeatedly halved; (4) a marginal bootstrap approach in which standardized marginal residuals are resampled; and (5) a "type-II" marginal bootstrap approach which also resamples rivers. Though none of these approaches may be ideal, together they should give some indication of the uncertainty of the estimates. We now comment on each of these approaches in the context of the coho salmon data, commenting on some of their shortcomings.

In each step of the parametric bootstrap procedure, for each stock, random effects are sampled from their estimated distribution. Next, a vector of errors is sampled from their estimated distribution. The spawner observations are held fixed and simulated recruitments are constructed from the sampled random effects and errors. As discussed above, this is strongly model-dependent. For robust estimation, this approach is logically flawed in that we do not believe that the model distribution holds exactly.

In each step of the jackknife approach, a different study is left out of the complete data set, so that in all 14 different sets of parameter estimates are obtained, from which standard errors may be computed using jackknife formulas. This procedure ignores within-stock variability, however.

Within a stock, the issue of how to perform nonparametric resampling is a special case

of the problem of nonparametric resampling for regression models. For an ordinary regression model, one approach is to pair the dependent and independent variables and bootstrap these observations. This is the approach applied with the multi-halver jackknife, since it preserves the structure of the data. We apply the multi-halver jackknife across stocks by treating the data on the 14 stocks as a single data set with a total of 185 observations. There are 128 possible pairs of orthogonal halvings. This approach ignores the between-stock variability.

The other approach for nonparametric resampling for regression models is to obtain residuals from a fitted regression, resample these residuals, and reconstruct pseudo observations. This is the approach applied in the marginal bootstraps described above. The marginal bootstraps seem the most satisfactory from a theoretical perspective since they incorporate both the within-stock and between-stock sources of variability.

A final approach, not implemented here, is nevertheless of interest. Given empirical Bayes estimates of stock-specific parameters, such as are naturally output by the Lindstrom and Bates algorithm, we can define “empirical Bayes” residuals. These residuals could be resampled and added to the empirical Bayes estimates to form pseudo-observations. In contrast to the “marginal” bootstraps described above, this might be termed a “conditional” bootstrap.

A careful simulation study would be required to determine which procedures are most satisfactory—and under what conditions. In the present work, we simply compare results for the observed data.

5.2.8 Example: Coho salmon Beverton-Holt mixed model

For the Beverton-Holt model, SPML gave very similar results to Lindstrom and Bates’ algorithm.

Algorithm	β_1	β_2	σ	σ_b	σ_f
Lindstrom & Bates	4.27 (.18)	6.58 (.18)	0.40	0.43	0.64
SPML	4.27	6.61	0.40	0.41	0.64
parametric bootstrap	(.18)	(.19)	(.02)	(.24)	(.15)
jackknife	(.20)	(.20)	(.04)	(.15)	(.16)
multi-halver	(.17)	(.08)	(.04)	(.18)	(.06)
marginal bootstrap	(.13)	(.09)	(.03)	(.18)	(.09)
type-II marginal bootstrap	(.14)	(.15)	(.04)	(.21)	(.13)

Table 5.1: Point estimates (and standard errors in parentheses) for Beverton-Holt nonlinear mixed effects model for coho salmon. The estimates and standard errors for the Lindstrom & Bates algorithm were obtained from the Splus function `nlme`. Note that `nlme` does not report standard errors for the variance components. For SPML standard errors are reported from: a parametric bootstrap procedure with 100 replications; a leave-out-one-study jackknife procedure; a multi-halver jackknife procedure; a marginal bootstrap; and a “type-II” marginal bootstrap in which populations *and* residuals were resampled.

The similarity of the parameter estimates from the Lindstrom and Bates algorithm and the SPML algorithm provides some reassurance that, at least in this case, the approximations underlying the two algorithms have a similar effect. The standard errors from the different resampling approaches are roughly similar, although the multihalver and marginal bootstrap standard errors are generally smaller than the others.

The leave-one-out estimates on which the jackknife standard errors are based are of interest in themselves, since they tell us something about the influence of individual data sets.

Population omitted	β_1	β_2	σ	σ_b	σ_f
none	4.27	6.61	0.40	0.41	0.64
Big Beef Ck, Washington	4.26	6.59	0.41	0.43	0.67
Bingham Ck. Washington	4.20	6.62	0.41	0.36	0.69
Black Creek, BC	4.31	6.54	0.39	0.48	0.63
Carnation Creek, BC	4.18	6.62	0.42	0.35	0.68
Deer Creek, Oregon	4.24	6.60	0.41	0.42	0.67
Deschutes River, Washington	4.38	6.52	0.40	0.36	0.64
Flynn Creek, Oregon	4.25	6.70	0.38	0.46	0.61
Hooknose Creek, BC	4.21	6.64	0.41	0.38	0.68
Hunt's Creek, BC	4.32	6.63	0.38	0.47	0.67
Needle Branch Creek, Oregon	4.25	6.72	0.40	0.42	0.51
Nile Creek, BC	4.28	6.64	0.40	0.42	0.66
Qualicum River, BC	4.24	6.55	0.39	0.42	0.63
S. Fork Skykomish River, WA	4.23	6.54	0.41	0.40	0.61
Snow Creek, Washington	4.36	6.59	0.40	0.34	0.68

Table 5.2: Leave-one-out SPML parameter estimates for the coho salmon data.

Examination of Table 5.2 reveals that the leave-one-out effects are sometimes substantial: for example, leaving out the Deschutes River data boosts the estimated mean slope at the origin by 12%. The Flynn Creek and Hunt's Creek data sets seem to affect the estimate of σ ; these are data sets exhibiting a large amount of recruitment variability, so that omitting them lowers the estimate of σ . The Needle Branch Creek data set seems to affect the estimate of σ_f ; this data set exhibits a very low asymptotic level (e.g. see Figure 2.12 on p. 56), so that omitting it lowers the estimated variability in the asymptotic level.

Sensitivity of estimates

Since one of the goals of this work is to consider methods for robust estimation of nonlinear mixed effects models, it is important to consider the sensitivity of parameter estimates to unusual or outlying data. As discussed earlier, it is important to consider the effects of both outlying studies (data sets) and outlying observations within studies.

We begin by considering the sensitivity of estimates to outlying studies (data sets). A realistic example is furnished by the data entry error mentioned in the Introduction (Section 1.3, p. 10). Recall that the river length for the Hunt's Creek data set was originally entered as 1.4 km instead of the correct value of 5.4 km. Considering this data set alone, this mistake would inflate the estimate of the asymptotic recruitment level by a factor of $5.4/1.4 \approx 3.9$, but the estimate of the slope at the origin would be unaffected. Table 5.3 shows the effect of this error on estimates of the nonlinear mixed effects model.

Data set	β_1	β_2	σ	σ_b	σ_f
Correct	4.27	6.61	0.40	0.41	0.64
Incorrect	4.27	6.72	0.40	0.41	0.70

Table 5.3: SPML parameter estimates for the correct data and for the incorrect data (with the river length for the Hunt's Creek data set equal to 1.4 km instead of 5.4 km).

Relative to the standard errors reported in Table 5.1, the changes in the parameter estimates are not large. Nevertheless, the estimate of β_2 has increased by 0.11, corresponding to an increase in the estimated mean asymptotic level of 12% and the standard deviation component σ_f has been inflated. The effect of such an error might have been more pronounced if it had occurred for a different stock. Examination of the raindrop plot of Figure 2.12 (p. 56) shows that the asymptotic-level raindrop for Hunt's Creek is roughly in the middle of the other raindrops. Scaling it up by a factor of $5.4/1.4 \approx 3.9$ (which is a shift of 1.4 on a log scale) still keeps it near the other raindrops.

Suppose instead that the Deschutes River data set had been incorrectly scaled by the same factor (i.e. 3.9). Table 5.4 shows the effect such an error would have.

Data set	β_1	β_2	σ	σ_b	σ_f
Correct	4.27	6.61	0.40	0.41	0.64
Incorrect	4.21	6.83	0.41	0.37	0.85

Table 5.4: SPML parameter estimates for the correct data and for incorrect data in which the Deschutes River data set is scaled up by a factor of 3.9.

The effects are considerably more pronounced than was the case for Hunt's Creek because the modification pushes the Deschutes River raindrop away from the other raindrops. The estimate of σ_f in particular reflects this.

Next, we consider the sensitivity of estimates to outlying observations within data sets. Consider the Qualicum River data set. The largest observed spawner quantity also happens to correspond to the largest observed recruitment, 3000 female smolts per kilometre of river. Suppose that due to a data entry error this was recorded as 30000 instead of 3000. Table 5.5 shows the effect of this on the parameter estimates.

Data set	β_1	β_2	σ	σ_b	σ_f
Correct	4.27	6.61	0.40	0.41	0.64
Incorrect	4.17	6.68	0.46	0.32	0.69

Table 5.5: SPML parameter estimates for the correct data and for data with the largest recruitment for Qualicum River incorrectly recorded as 30000.

This single data-entry error has resulted in substantial changes in the parameter estimates. For example, the β_1 parameter has decreased by 0.10, corresponding to a 10% decrease in the estimated mean slope at the origin. The estimated variance components have also

changed considerably. A single spurious observation can have drastic effects on parameter estimates.

The sensitivity of estimates of nonlinear mixed effects models to outlying observations and/or studies provides motivation for the development of estimation methods that are robust to contamination of the assumed distributions of the random effects and the errors. We return to this in Section 5.3, but for now we move on to consider another model for the coho data.

5.2.9 Example: Coho salmon hockey-stick mixed model

Recall from the Introduction, the hockey stick spawner-recruitment model:

$$R_{i,j} = \alpha_i \min(S_{ij}, S_i^*), = \begin{cases} \alpha_i S_{ij} & \text{if } S_{ij} < S_i^* \\ \alpha_i S_i^* & \text{if } S_{ij} \geq S_i^*. \end{cases}$$

Note that $\alpha_i S_i^*$ is the maximum recruitment, and we therefore write $R_i^{\max} = \alpha_i S_i^*$. The model of Section 5.1.1 (p. 107) is easily adapted for the hockey stick model. Letting $y_{ij} = \log(R_{ij}/S_{ij})$, we have

$$y_{ij} = \begin{cases} \log \alpha_i & \text{if } S_{ij} < S_i^* \\ \log(R_i^{\max}/S_{ij}) & \text{if } S_{ij} \geq S_i^*. \end{cases}$$

As for the Beverton-Holt model, we set $\alpha_i = e^{\theta_{i1}}$ and $R_i^{\max} = e^{\theta_{i2}}$ with $\theta_{i1} = \beta_1 + b_i$ and $\theta_{i2} = \beta_2 + f_i$. Thus

$$\eta_i(\theta_i) = \begin{cases} \theta_{i1} & \text{if } S_{ij} < S_i^* \\ \theta_{i2} - \log S_{ij} & \text{if } S_{ij} \geq S_i^*. \end{cases}$$

We adopt the same distributional assumptions we made for the Beverton-Holt mixed model. As noted in Section 2.6.3, the hockey stick model can lead to problems because the likelihood surface is not smooth and may have multiple local maxima. Particularly with the Lindstrom & Bates algorithm we noted some difficulties with starting values for this model. Parameter estimates are given in Table 5.6.

Algorithm	β_1	β_2	σ	σ_b	σ_f
Lindstrom & Bates	3.71 (.22)	6.28 (.66)	0.40	0.22	0.66
SPML	4.01 (.14)	6.37 (.36)	0.40 (.05)	0.29 (.14)	0.78 (.28)

Table 5.6: Point estimates (and standard errors) for hockey-stick nonlinear mixed effects model for coho salmon. The estimates and standard errors for the Lindstrom & Bates algorithm were obtained from the Splus function `n1me`. Note that `n1me` does not report standard errors for the variance components. The standard errors reported for SPML were computed using a parametric bootstrap procedure with 50 replications.

Comparing these results with those of Table 5.1, we see that the hockey-stick model generally gives lower estimates of the slope at the origin than the Beverton-Holt model. Barrowman and Myers (2000) studied the hockey-stick and its generalizations in a fixed effects context. They showed that as $S \downarrow 0$ the Beverton-Holt model extrapolates the survival R/S above observed levels.

5.3 Approaches to robust estimation

There do not seem to be any proposals for robust estimation of nonlinear mixed effects models in the literature. That this is the case is hardly surprising. Robust estimation of linear mixed effects models has been a fairly recent development, and the structure of those models makes robustification relatively straightforward. For example, as described in Chapter 4, Huggins (1993b) replaced the marginal likelihood of y with an objective function that is less sensitive to extreme observations. In contrast, for the nonlinear model we have no closed-form expression for the marginal likelihood or even for the mean $\mu = E(y)$ and variance $V = \text{Var}(y)$.

We therefore consider modifications to two of the estimation algorithms that we described above.

5.3.1 Modifying Lindstrom and Bates' algorithm

One fairly obvious approach is to try to robustify each of the two steps in the Lindstrom and Bates (1990) algorithm. Recall that the pseudo-data step was based on maximizing the joint probability density of y and u in β and u . This distribution may be interpreted as being proportional to the posterior for β and u . The terms that depend on β and u are

$$\begin{aligned} & (y - \eta(A\beta + Bu))^T R^{-1} (y - \eta(A\beta + Bu)) + u^T \bar{D}^{-1} u \\ & = \left[R^{-1/2} (y - \eta(A\beta + Bu)) \right]^T \left[R^{-1/2} (y - \eta(A\beta + Bu)) \right] + \\ & \quad \left[\bar{D}^{-1/2} u \right]^T \left[\bar{D}^{-1/2} u \right], \end{aligned} \quad (5.35)$$

and thus maximizing the joint density is equivalent to solving a nonlinear least squares problem. A straightforward approach to robustifying this replaces each of the two sum of squares terms in (5.35) by a less rapidly growing function. Thus we replace (5.35) by

$$\rho_1 \left[R^{-1/2} (y - \eta(A\beta + Bu)) \right] + \rho_2 \left[\bar{D}^{-1/2} u \right], \quad (5.36)$$

where, as in the previous chapter, we abuse notation slightly so that $\rho_1(r_i) = \sum_{j=1}^{n_i} \rho_1(r_{ij})$, and similarly for ρ_2 . We refer to this robustified version of Lindstrom and Bates' pseudo-data step as the *joint step* in our algorithm.

There is an interesting connection between this approach and the algorithm proposed by Fellner (1986) for robust estimation of linear mixed effects models. We mentioned Fellner's approach to robust estimation of realized random effects in Section 4.6.1 (p. 102). As noted there, Fellner's algorithm is based on the Henderson-Harville algorithm (Henderson 1963; Henderson 1973; Harville 1977) for variance components estimation. The Henderson-Harville algorithm uses a by-now-familiar tactic: replacing integration by maximization via solution of the mixed model equations (4.5). Welsh and Richardson (1997) show that equations similar to (5.36) are closely related to Fellner's algorithm.

To robustify the linear mixed effects step of Lindstrom and Bates' algorithm, we start

with the approximate marginal log likelihood (5.18):

$$-\frac{1}{2} [(w - \hat{X}\beta)^T \hat{V}^{-1} (w - \hat{X}\beta) + \log |\hat{V}|].$$

We then robustify this as in Robust ML I

$$\sum_{i=1}^m \left(\rho \left[\hat{V}_i^{-1/2} (y_i - \hat{X}_i \beta) \right] + \frac{\kappa}{2} \log |\hat{V}_i| \right). \quad (5.37)$$

We refer to this robustified version of Lindstrom and Bates' linear mixed effects step as the *marginal step* in our algorithm.

We thus have an alternating two-step algorithm defined by the choice of functions ρ_1 , ρ_2 , and ρ , together with consistency correction κ . Given the k th estimate of ζ , denoted $\zeta^{(k)}$, perform the following steps:

Step 1. Joint step. With $\zeta = \zeta^{(k)}$, minimize (5.36) to obtain $\beta^{(k)}$ and $u^{(k)}$.

Step 2. Marginal step. With $\hat{\beta} = \beta^{(k)}$, $\hat{u} = u^{(k)}$, $\hat{X} = \hat{X}^{(k)}$, and $\hat{Z} = \hat{Z}^{(k)}$, minimize (5.37) to obtain $\beta^{(k+1)}$ and $\zeta^{(k+1)}$.

There are disadvantages to basing an estimation algorithm on the algorithm of Lindstrom and Bates (1990). First, as noted on page 5.2.5, the asymptotics may be compromised. Our proposed algorithm also features three different ρ -functions and a consistency correction term. Presumably the different ρ -functions provide control over different aspects of the sensitivity of the estimator to outlying values. However, it is not clear how to choose the ρ functions and how to set the consistency correction term. Finally, the asymptotic properties of the proposed algorithm may not be easy to determine. A more straightforward approach may be to modify Nuñez's SPML method, and we will focus on this method. Nevertheless, we do report on some results from the robustified Lindstrom and Bates algorithm below.

Weights corresponding to a robust fit

In Section 2.4.1 (p. 39), we defined weights for robust fits as $w_{ij} = \psi(r_{ij})/r_{ij}$, where r_{ij} are residuals from a robust fit. Here we have directly optimized the robustified objective

function, so an appropriate definition of weights is

$$w_{ij} = \sqrt{\frac{\rho(r_{ij})}{r_{ij}^2/2}}.$$

When $\rho(r_{ij}) = \frac{1}{2}r_{ij}^2$, we have $w_{ij} = 1$.

5.3.2 Example: Coho salmon Beverton-Holt mixed model

The robustified Lindstrom and Bates algorithm was implemented in order to obtain some empirical evidence of its performance. Following Huggins (1993b), Tukey's ρ -function was used with a tuning constant of 4 and a consistency correction of 0.6835546. It is not clear that this consistency correction is appropriate. Square roots of the \hat{V}_i matrix were performed using Cholesky decompositions.

We note that the robust fit gave an increased estimate of the mean of $\log \alpha$ of 4.40 instead of 4.27. While this is not significant (according to the approximate standard errors), it is suggestive.

Figure 5.7 shows empirical Bayes curves based on the robust fit on a log survival scale for two of the coho salmon populations. The whisker-weight scatterplot of Section 2.4.2 is used to display the robust weights in these cases.

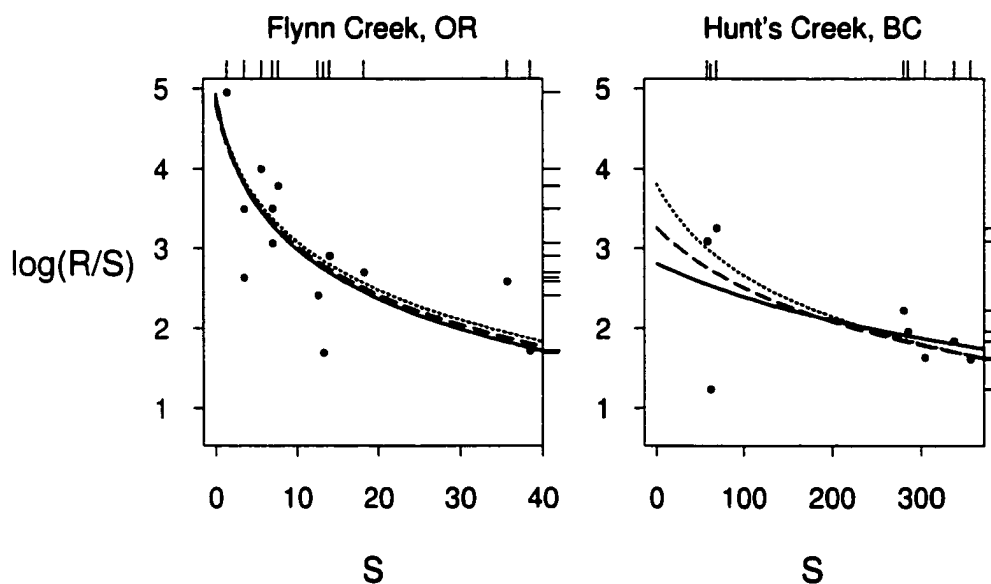


Figure 5.7: Log survival versus spawners for two of the coho salmon populations. Superimposed on each plot are Beverton-Holt log survival curves obtained from the individual fits (solid), the empirical Bayes curve from the Lindstrom and Bates algorithm (dashed), and the empirical Bayes curve from the robustified Lindstrom and Bates algorithm (dotted). On the top margins of each panel are whiskers representing the individual weights from the robust fit. The whiskers are repeated on the right-hand margin to help distinguish weights of nearby points.

The Flynn Creek fits are essentially unaffected by the robust modification. The Hunt's Creek fit shows the strongest effect of the robust modification. Note, however, that as we saw for the wolf data in the previous chapter, the mixed effects model seems to deal with the unusual observation to some extent.

5.3.3 Modifying SPML

Once again, recall that the ELS estimator minimizes

$$\sum_{i=1}^m (y_i - \mu_i)^T V_i^{-1} (y_i - \mu_i) + \log |V_i|, \quad (5.38)$$

where $\mu_i = E(y_i)$ and $V_i = \text{Var}(y_i)$. If we could compute these moments, a natural candidate for a robust estimator would be an analogue of the Robust ML I estimator, i.e., the minimizer of

$$\sum_{i=1}^m \left(\rho \left[V_i^{-1/2} (y_i - \mu_i) \right] + \frac{\kappa}{2} \log |V_i| \right).$$

Of course we cannot, in general, compute the μ_i 's and the V_i 's. However, SPML provides an approximation to (5.38), namely

$$\sum_{i=1}^m [y_i - \bar{\mu}_{i,g}(\lambda)]^T V_{i,g}(\lambda)^{-1} [y_i - \bar{\mu}_{i,g}(\lambda)] + \log |V_{i,g}(\lambda)|,$$

where $\bar{\mu}_{i,g}(\lambda)$ and $V_{i,g}(\lambda)$ are Monte Carlo estimates of μ_i and V_i based on the parameter vector λ . Therefore a computable analogue of the Robust ML I estimator is the minimizer of

$$\sum_{i=1}^m \left(\rho \left[V_{i,g}(\lambda)^{-1/2} (y_i - \bar{\mu}_{i,g}(\lambda)) \right] + \frac{\kappa}{2} \log |V_{i,g}(\lambda)| \right), \quad (5.39)$$

which we call “robust SPML” or *RSPML*. In notation compatible with what we used for SPML, we can write a normalized version of this simulated objective function as

$$\bar{d}_m^g(\lambda) = \frac{1}{m} \sum_{i=1}^m d_i^g(\lambda)$$

where

$$d_i^g(\lambda) = 2\rho \left[V_{i,g}(\lambda)^{-1/2} (y_i - \bar{\mu}_{i,g}(\lambda)) \right] + \kappa \log |V_{i,g}(\lambda)|.$$

When $\rho(r_{ij}) = \frac{1}{2} r_{ij}^2$ (with, as before, $\rho(r_i) = \sum_{j=1}^{n_i} \rho(r_{ij})$) and $\kappa = 1$, we have $\bar{d}_m^g(\lambda) = \bar{c}_m^g(\lambda)$, the SPML objective function.

Of course the robustness of *RSPML* depends on the choice of a suitable ρ -function. As for the linear mixed effects model in the previous chapter, for the influence function to be bounded, the terms in the estimating equations must be bounded. To obtain estimating

equations, we equate the derivatives of (5.39) to zero:

$$\sum_{i=1}^m \left\{ \frac{\partial}{\partial \lambda^T} \left[V_{i,g}(\lambda)^{-1/2} (y_i - \bar{\mu}_{i,g}(\lambda)) \right] \psi \left[V_{i,g}(\lambda)^{-1/2} (y_i - \bar{\mu}_{i,g}(\lambda)) \right] + \frac{\kappa}{2} \text{tr} \left[V_{i,g}(\lambda)^{-1} \left(\frac{\partial V_{i,g}(\lambda)}{\partial \lambda^T} \right) \right] \right\} = 0.$$

For robustness, the summand above must be bounded in y_i . As for the linear mixed effects model, this means that not only must $\psi(r)$ be bounded but also $r\psi(r)$ must be bounded. A redescending ψ function will thus guarantee the robustness of RSPML.

Asymptotic properties of RSPML

If we knew the marginal variances $V_i(\lambda)$ and means $\mu_i(\lambda)$, then we would not need to use simulation to compute the objective function $d_m^{\bar{g}}(\lambda)$. Instead, we could optimize

$$\bar{e}_m(\lambda) = \frac{1}{m} \sum_{i=1}^m e_i(\lambda)$$

where

$$e_i(\lambda) = 2\rho \left[V_i(\lambda)^{-1/2} (y_i - \mu_i(\lambda)) \right] + \kappa \log |V_i(\lambda)|.$$

We will call this *RPML* (Robust Pseudo Maximum Likelihood). In this section we consider the asymptotic properties of RPML. In practice we use RSPML, but the number of simulations, g , can be made arbitrarily large so that $\mu_i(\lambda)$ and $V_i(\lambda)$ can be estimated to arbitrary precision.

In order to obtain asymptotic results in a mixed effects model context, we need to account for the fact that although the observation vectors y_i are assumed independent, they are not identically distributed. We abbreviate this i.n.i.d. For example, Hoadley (1971) obtained results for maximum likelihood estimates in the i.n.i.d. case. Beal (1984) generalized Hoadley's results for estimators obtained by minimizing an objective function. We will obtain sufficient conditions for consistency and asymptotic normality of RPML by showing that they satisfy Beal's conditions.

In Section 5.2.5, we saw that SPML estimates of λ are consistent for λ^* , the parameter vector that minimizes the KLIC, which is also the value of λ for which the model matches

the first two moments of the observations. While RPML estimates of λ can be shown to be consistent for λ^* , this λ^* does not have the attractive moment-matching property. For this reason, we may wish to consider some form of bias correction. One way to do this is through the choice of the consistency correction, κ .

Consistency

We begin with some notation following Beal and Hoadley, but with modifications where necessary to reduce conflict with our notation. Define the following extended random variables:

$$R_i(\lambda) = \begin{cases} e_i(\lambda^*) - e_i(\lambda), & \text{if } e_i(\lambda^*) < \infty \\ 0 & \text{otherwise.} \end{cases}$$

$$R_i(\lambda, q) = \sup_t \{R_i(t) : \|t - \lambda\| \leq q\}.$$

$$W_i(r) = \sup_{\lambda} \{R_i(\lambda) : \|\lambda\| > r\}.$$

Let $r_i(\lambda) = E[R_i(\lambda)]$, $r_i(\lambda, q) = E[R_i(\lambda, q)]$, and $w_i(r) = E[W_i(r)]$. Let $\bar{r}_m(\lambda) = \frac{1}{m} \sum r_i(\lambda)$, and $\bar{w}_m(r) = \frac{1}{m} \sum w_i(r)$.

Beal's conditions for consistency (expressed in our notation) are:

C1. $\lambda \in \Lambda$, a closed subset of R^p .

C2. $e_i(\lambda)$ is almost surely an upper semicontinuous function, uniformly in i .

C3'. There exists $q^* = q(\lambda) > 0$, $r > 0$, and $K > 0$ for which

(i) $E[R_i(\lambda, q)^2] \leq K$ for all i and $0 \leq q \leq q^*$.

(ii) $E[W_i(r)^2] \leq K$ for all i .

C4'. (i) $\lim \bar{r}_m(\lambda) < 0$, $\lambda \neq \lambda^*$.

(ii) $\lim \bar{w}_m(r) < 0$.

C5. (i) $R_i(\lambda, q)$ is a measurable function of y_i .

(ii) $W_i(r)$ is a measurable function of y_i .

Beal points out that if Λ is assumed to be compact then conditions C3'(ii), C4'(ii), and C5(ii) are unnecessary.

Beal's primary interest was in extended least squares estimation of nonlinear mixed effects models. In this context, he provided conditions under which the above conditions for consistency are satisfied. I will adapt Beal's conditions and argument for my case.

The critical condition for consistency is C3'(i). In our case, the presence of a bounded ρ -function (e.g., Tukey's ρ) simplifies the task of showing that C3'(i) holds. We still need to bound the determinants $|V_i(\lambda)|$ and to that end, we now introduce some notation. We begin with a notational device that is required because the number of observations, n_i , (and hence the dimension of V_i) varies from study to study. Let $s = \max\{n_1, \dots, n_m\}$ be the maximum number of observations on a study, assumed bounded. Let $f \in \{1, \dots, s\}$. For any sequence $\{X_i\}$, define a subsequence $\{X_{fk}\}$ as being all terms X_i such that $n_i = f$. For conformable matrices A and B , we write $B > A$ (or $A < B$) if $B - A$ is positive definite.

To prove consistency in our case, we adopt the following assumptions:

- A1. Λ is compact.
- A2. μ_i and V_i are both continuous functions uniformly in i .
- A3. ρ is continuous and bounded, i.e. for any z , $\rho(z) < c$.
- A4. For each λ , there exists a sequence $(A_1, B_1), \dots, (A_s, B_s)$ such that for each $f \in \{1, \dots, s\}$,
 - (i) A_f and B_f are positive definite matrices
 - (ii) $A_f < V_{fk} < B_f$ for each k .

Theorem 1. If conditions A1–A4 and condition C4'(i) are satisfied, then $\hat{\lambda} \rightarrow_p \lambda^*$ as $m \rightarrow \infty$.

Proof. C2 is satisfied by A2 and A3. C1 is satisfied by A1. Also by A1, conditions C3'(ii), C4'(ii), and C5(ii) are unnecessary, so we need only show C3'(i), and C5(i). To show C3'(i), first note that by A4, for all i , $|V_i(\lambda)| > d = \min_f |A_f(\lambda)|$. Then, by the continuity of V_i , which is uniform in i , we have for some $q^* > 0$, for all $q \leq q^*$, and for all i ,

$$\sup \{ -\log |V_i(t)| : \|t - \lambda\| \leq q \} \leq -\log d.$$

Also by A4, for all i , $|V_i(\lambda)| < e = \max_f |B_f(\lambda^*)| < \infty$. Hence for all $q \leq q^*$ and for all i ,

$$R_i(\lambda, q) \leq \kappa \log(e/d) + 2c.$$

By a similar argument,

$$R_i(\lambda, q) \geq \kappa \log(g/h) - 2c,$$

where $g = \min_f |A_f(\lambda^*)|$ and $h = \max_f |B_f(\lambda)|$. Therefore

$$R_i(\lambda, q)^2 \leq [\kappa \log(e/d) + 2c]^2 + [\kappa \log(g/h) - 2c]^2$$

Therefore for each λ , $E[R_i(\lambda, q)^2] \leq K$ for all i and $0 \leq q \leq q^*$, satisfying C3'(i).

To satisfy C5(i), by A2, A3, and Lemma 2 of Jennrich (1969), there exists a Borel-measurable function, $\tilde{\lambda}$, on R^s into Λ such that $R_i(\lambda, q) = R_i(\tilde{\lambda}(y_i))$. By A2 and A3, the function defined on R^s , given by $y \rightarrow e_i(y, \lambda^*) - e_i(y, \tilde{\lambda}(y))$ is Borel-measurable. Therefore C5(i) holds.

Q.E.D.

Condition C4'(i) defines λ^* , the parameter value for which our estimator is consistent. Beal (1984) provided a scenario under which λ^* is defined in "finite terms." The idea is as follows: for a given data set (e.g., spawner-recruitment data for a set of rivers), we call $\{(n_1, \mu_1, V_1), \dots, (n_m, \mu_m, V_m)\}$ the "explanatory sequence" of the data set. One may then define a point, $\tilde{\lambda}$, as the unique minimizer of $E(\bar{e}_m(\lambda))$. If we imagine an infinite number of data sets with identical explanatory sequences, then concatenating these data sets and concatenating their explanatory sequences gives "an infinite sequence of observations whose associated infinite explanatory sequences have a repeating character." Condition C4'(i) is then satisfied with $\lambda^* = \tilde{\lambda}$.

Asymptotic normality

For consistency with the notation of Beal (and in turn Hoadley), we define

$$\Phi_i(y_i, \lambda) = -e_i(\lambda).$$

Also, let

$$\dot{\Phi}_i(y_i, \lambda) = -\frac{\partial e_i}{\partial \lambda} \quad \text{and} \quad \ddot{\Phi}_i(y_i, \lambda) = -\frac{\partial^2 e_i}{\partial \lambda \lambda^T}.$$

Beal's conditions for asymptotic normality (expressed in our notation) are:

N1. λ^* is an interior point of Λ .

N2. $\hat{\lambda} \rightarrow_P \lambda^*$.

N3. $\dot{\Phi}_i(y_i, \lambda)$ and $\ddot{\Phi}_i(y_i, \lambda)$ exist almost surely.

N4. $\dot{\Phi}_i(y_i, \cdot)$ is a continuous function uniformly in i , almost surely, and $\dot{\Phi}_i(\cdot, \lambda)$ is a measurable function.

N5. (i) $\pi_i = E [\dot{\Phi}_i(y_i, \lambda^*)]$ exists.

(ii) $m^{1/2} \bar{\pi}_m = m^{-1/2} \sum \pi_i \rightarrow \pi$.

N6. (i) $\Gamma_i(\lambda) = -E [\ddot{\Phi}_i(y_i, \cdot)]$ exists.

(ii) $\Delta_i = E [(\dot{\Phi}_i(y_i, \lambda^*) - \pi_i)(\dot{\Phi}_i(y_i, \lambda^*) - \pi_i)^T]$ exists.

N7. (i) $\bar{\Gamma}_m(\lambda) = m^{-1} \sum \Gamma_i(\lambda) \rightarrow \bar{\Gamma}(\lambda)$.

(ii) $\bar{\Delta}_m = m^{-1} \sum \Delta_i \rightarrow \bar{\Delta}$, and $\bar{\Delta}$ is positive definite.

N8'. There exists $K > 0$ such that $E [|\dot{\Phi}_{i,k}(y_i, \lambda^*) - \pi_{i,k}|^3] \leq K$ for all i, k .

N9. There exists $K > 0$, $\delta > 0$, and $\varepsilon > 0$ and random variables, $B_{i,jk}(y_i)$ such that

(i) $\sup \{ |\dot{\Phi}_{i,jk}(y_i, t)| : \|t - \lambda^*\| \leq \varepsilon \} \leq B_{i,jk}(y_i)$.

(ii) $E |B_{i,jk}(y_i)|^{1+\delta} \leq K$ for all i, j, k .

To determine sufficient conditions for asymptotic normality in our case, we need to examine the first and second derivatives of the e_i 's. For the k th element of the derivative of e_i , we have

$$\frac{\partial e_i}{\partial \lambda_k} = 2b_{ik}^T \Psi \left[V_i(\lambda)^{-1/2} (y_i - \mu_i(\lambda)) \right] + \text{tr} \left[V_i(\lambda)^{-1} \frac{\partial V_i}{\partial \lambda_k} \right] \quad (5.40)$$

where

$$b_{ik} = \frac{\partial}{\partial \lambda_k} \left\{ V_i(\lambda)^{-1/2} (y_i - \mu_i(\lambda)) \right\} = -V_i(\lambda)^{-1/2} \frac{\partial \mu_i}{\partial \lambda_k} + \frac{\partial V_i^{-1/2}}{\partial \lambda_k} (y_i - \mu_i(\lambda)).$$

It is useful here to recall that provided that a matrix A is invertible, the derivative of A exists if and only if the derivative of A^{-1} exists, and that the following relationship holds:

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1} \frac{\partial A}{\partial x} A^{-1}.$$

For the (k, ℓ) th element of the second derivative of e_i , we have

$$\begin{aligned} \frac{\partial^2 e_i}{\partial \lambda_k \partial \lambda_\ell} &= 2b_{ik}^T \psi' \left[V_i(\lambda)^{-1/2} (y_i - \mu_i(\lambda)) \right] b_{i\ell} + \\ &2 \left[\frac{\partial^2 \mu_i^T}{\partial \lambda_k \partial \lambda_\ell} V_i(\lambda)^{-1/2} - \frac{\partial \mu_i^T}{\partial \lambda_k} \frac{\partial V_i^{-1/2}}{\partial \lambda_\ell} - \frac{\partial \mu_i^T}{\partial \lambda_\ell} \frac{\partial V_i^{-1/2}}{\partial \lambda_k} + (y_i - \mu_i(\lambda))^T \frac{\partial^2 V_i^{-1/2}}{\partial \lambda_k \partial \lambda_\ell} \right] \times \\ &\psi \left[V_i(\lambda)^{-1/2} (y_i - \mu_i(\lambda)) \right] + \\ &\text{ctr} \left[V_i(\lambda)^{-1} \frac{\partial^2 V_i}{\partial \lambda_k \partial \lambda_\ell} + \frac{\partial V_i(\lambda)^{-1}}{\partial \lambda_\ell} \frac{\partial V_i}{\partial \lambda_k} \right] \end{aligned} \quad (5.41)$$

In the context of extended least squares estimation, Beal provided conditions under which the above general conditions for asymptotic normality are satisfied. In addition to A1–A4, we shall require the following conditions:

- A5. λ^* is an interior point of Λ .
- A6. $\mu_i(\lambda)$ and $V_i(\lambda)$ are both twice continuously differentiable uniformly in i .
- A7. ρ is twice continuously differentiable.
- A8. All third order moments of y_i exist and are bounded.

Theorem 2. If conditions A1–A8, condition C4'(i), and conditions N5(ii) and N7–N9 are satisfied, and provided $\bar{\Gamma}$ is non-singular,

$$m^{1/2}(\hat{\lambda} - \lambda^*) \rightarrow_D N(\bar{\Gamma}^{-1}\mu, \bar{\Gamma}^{-1}\bar{\Delta}\bar{\Gamma}^{-1}).$$

Proof. We need to show that conditions N1–N4, N5(i) and N6 are satisfied. Condition N1 is satisfied by A5. Condition N2 is satisfied by Theorem 1. Conditions N3 and N4 follow from A6 and A7 and equations (5.40) and (5.41). Conditions N5(i) and N6 follow from A8 and equations (5.40) and (5.41).

Q.E.D.

Comments. As noted above, λ^* may not be the parameter value we seek, and some form of bias correction may be necessary. In the robustness literature, perhaps the most common approach to this is to ensure Fisher consistency, i.e. to ensure that $\pi_i = 0$ for $i = 1, \dots, m$, typically by careful choice of κ , the consistency correction. We will thus assume Fisher consistency, satisfying condition N5(ii).

Recall from page 5.3.3 that λ^* can be defined in finite terms by envisioning infinite repetitions of an experiment. Under this scenario, condition N7 will be satisfied.

We now comment on the applicability of the remaining conditions for asymptotic normality for the spawner-recruitment data and models. Condition A5 is a standard assumption that requires that we are not close to a boundary of the parameter space. This does not appear to be a concern for the spawner-recruitment models considered here, although we have observed boundary estimates for some three-parameter models. Tukey's ρ -function is twice continuously differentiable, satisfying condition A7.

Condition A6 is satisfied for suitably smooth nonlinear functions η . To see this, consider

$$\frac{\partial \mu_{ij}}{\partial \lambda_k} = \frac{\partial}{\partial \lambda_k} \int \eta_{ij}(\theta_i) \phi_i(\theta_i) d\theta_i, \quad (5.42)$$

where ϕ_i is the multivariate normal pdf with mean vector $A_i\beta$ and variance-covariance matrix $B_i DB_i^T$. We wish to show that $\frac{\partial \mu_{ij}}{\partial \lambda_k}$ is well behaved. To this end we wish to exchange the order of differentiation and integration in (5.42). By Lebesgue's Dominated Convergence Theorem and the mean value theorem, we can do this provided that there exists a function $g(\theta_i, \lambda_k)$ and a constant $\delta_0 = \delta_0(\lambda_k) > 0$ such that

$$\left| \frac{\partial}{\partial \lambda_k} \eta_{ij}(\theta_i) \phi_i(\theta_i) \Big|_{\lambda_k = \lambda'_k} \right| \leq g(\theta_i, \lambda_k) \quad \text{for all } \lambda'_k \text{ such that } |\lambda'_k - \lambda_k| \leq \delta_0$$

with

$$\int g(\theta_i, \lambda_k) d\theta_i < \infty.$$

Note that

$$\frac{\partial}{\partial \lambda_k} \eta_{ij}(\theta_i) \phi_i(\theta_i) = \eta_{ij}(\theta_i) \frac{\partial \phi_i(\theta_i)}{\partial \lambda_k}.$$

Consider our Beverton-Holt model, where

$$\eta_{ij}(\theta_i) = -\log(e^{-\theta_{i1}} + S_{ij}e^{-\theta_{i2}})$$

and ϕ_i has mean $(\beta_1, \beta_2)^T$ and variance-covariance matrix $\text{diag}(\sigma_b^2, \sigma_f^2)$, so that

$$\phi_i(\theta_i) = (2\pi\sigma_b\sigma_f)^{-1} \exp\left(-\frac{1}{2} \left\{ (\theta_{i1} - \beta_1)^2/\sigma_b^2 + (\theta_{i2} - \beta_2)^2/\sigma_f^2 \right\}\right).$$

For example,

$$\frac{\partial \phi_i(\theta_i)}{\partial \beta_1} = -\phi_i(\theta_i)(\theta_{i1} - \beta_1)/\sigma_b^2.$$

Thus

$$\left| \frac{\partial}{\partial \beta_1} \eta_{ij}(\theta_i) \phi_i(\theta_i) \right| = \left| \log(e^{-\theta_{i1}} + S_{ij}e^{-\theta_{i2}}) \left(\frac{\theta_{i1} - \beta_1}{\sigma_b^2} \right) \right| \phi_i(\theta_i).$$

Provided $\sigma_b^2 > 0$, the above expression is bounded by an integrable function, as required.

Thus the derivative with respect to β_1 can be taken inside the integral, and

$$\frac{\partial \mu_{ij}}{\partial \beta_1} = \int \log(e^{-\theta_{i1}} + S_{ij}e^{-\theta_{i2}}) \left(\frac{\theta_{i1} - \beta_1}{\sigma_b^2} \right) \phi_i(\theta_i) d\theta_i. \quad (5.43)$$

This derivative is thus continuous. Next, consider the second derivative

$$\frac{\partial^2 \mu_{ij}}{\partial \beta_1^2}.$$

Once again, to take the derivative inside the integral, we need to show that the derivative of the integral is bounded by an integrable function. The absolute value of the derivative of

the integrand in (5.43) is

$$\left| \log(e^{-\theta_{i1}} + S_{ij}e^{-\theta_{i2}}) \left[\left(\frac{\theta_{i1} - \beta_1}{\sigma_b^2} \right) - 1/\sigma_b^2 \right] \right| \phi_i(\theta_i).$$

Again, provided $\sigma_b^2 > 0$, the above expression is bounded by an integrable function. The derivative can be taken inside the integral sign, and we conclude that the second derivative is also continuous. In a similar way, we can show that all of the first and second derivatives of the moments are continuous.

Condition A8 (existence and boundedness of third moments) holds for coho salmon spawner-recruitment data provided recruitment is strictly positive. To see this, recall from Section 2.3 that y_{ij} can be interpreted as log survival, and is thus bounded above. Furthermore, since unstandardized recruitment observations are non-zero counts, and the rivers are of finite length, y_{ij} is bounded below. Since y_{ij} is bounded above and below, its third moments must exist and be bounded. (Note that since y_{ij} is in fact bounded above and below, our model of additive normal errors is only a convenient approximation.) The boundedness of y_{ij} also shows, together with (5.40), that N8' holds and together with (5.41) and A6 that N9 holds.

5.3.4 Example: Coho salmon Beverton-Holt mixed model

We ran RSPML for the coho salmon Beverton-Holt mixed model using Huber's ρ -function with a tuning constant of 2; the results were virtually identical using Tukey's ρ -function with a tuning constant of 8. The results are given in Table 5.7 together with the SPML results for comparison.

Algorithm	β_1	β_2	σ	σ_b	σ_f
SPML	4.27 (.18)	6.61 (.19)	0.40 (.02)	0.41 (.24)	0.64 (.15)
RSPML	4.41 (.21)	6.59 (.20)	0.36 (.05)	0.40 (.20)	0.59 (.16)

Table 5.7: Point estimates (and standard errors) for Beverton-Holt nonlinear mixed effects model for coho salmon estimated using SPML and RSPML using Huber's ρ -function with a tuning constant of 2. The reported standard errors were computed using a parametric bootstrap procedure with 100 replications.

As for the estimates from the robustified Lindstrom and Bates algorithm, the mean of $\log \alpha$, β_1 , is estimated to be higher than in the nonrobust fit. In fact, the RSPML estimates were virtually identical to those from the robustified Lindstrom and Bates algorithm.

Sensitivity of estimates

We now return to the examples used to assess the sensitivity of SPML results for the coho salmon Beverton-Holt case, to see how RSPML performs. The first example considered sensitivity to outlying studies (data sets) using an actual data entry error: the error in recording the river length for Hunt's Creek. Table 5.8 shows the effect of this error on RSPML estimates.

Data set	β_1	β_2	σ	σ_b	σ_f
Correct	4.41	6.59	0.36	0.40	0.59
Incorrect	4.41	6.70	0.36	0.42	0.62

Table 5.8: Point estimates for Beverton-Holt nonlinear mixed effects model for coho salmon estimated using RSPML using Huber's ρ -function with a tuning constant of 2 for the correct data and for the incorrect data (with the river length for the Hunt's Creek data set equal to 1.4 km instead of 5.4 km).

RSPML seems to be no less sensitive than SPML to this error. This is perhaps not too surprising. We have just 14 stocks and a robust procedure cannot be expected to perform very well in the case of a relatively modest deviation affecting an entire stock. However, when much larger changes (e.g. by a factor of 50) were introduced, RSPML showed less sensitivity than SPML.

The second example considered sensitivity of estimates to outlying observations within data sets. A modification of a single observation in the Qualicum River data set (changing a recruitment from 3000 to 30000) produced fairly substantial changes in the SPML parameter estimates. Table 5.9 shows the effect of this modification on the RSPML estimates.

Data set	β_1	β_2	σ	σ_b	σ_f
Correct	4.41	6.59	0.36	0.40	0.59
Incorrect	4.37	6.60	0.38	0.38	0.60

Table 5.9: Point estimates for Beverton-Holt nonlinear mixed effects model for coho salmon estimated using RSPML using Huber's ρ -function with a tuning constant of 2 for the correct data and for data with the largest recruitment for Qualicum River incorrectly recorded as 30000.

The RSPML estimates moved considerably less than the SPML estimates did in the same situation (Table 5.5). For example, the β_1 parameter decreased by 0.04, corresponding to a 4% decrease in the estimated mean slope at the origin instead of the 10% decrease observed for SPML. Table 5.10 shows the percentage change in the estimate of each parameter for SPML and RSPML.

Algorithm	Percentage change in estimates				
	$E(\alpha)$	$E(R^{max})$	σ	σ_b	σ_f
SPML	-10	7	15	-22	8
RSPML	-4	1	6	-5	2

Table 5.10: Percentage change in estimates due to altering the largest recruitment for Qualicum River from 3000 to 30000 for SPML and RSPML. (From Tables 5.5 and 5.9.)

5.3.5 Comparison of the two proposals for robust estimation

The RSPML approach to robust estimation of nonlinear mixed effects models seems preferable to the modification of Lindstrom and Bates' algorithm for several reasons. First, as noted on page 5.2.5, the asymptotics of the Lindstrom and Bates algorithm may be compromised. Second, our proposed modification of the Lindstrom and Bates algorithm includes three different ρ -functions and a consistency correction term. Presumably the different ρ -functions provide control over different aspects of the sensitivity of the estimator to outlying values. However, it is not clear how to choose the ρ functions and how to set the consistency correction term. Nor is it clear what the convergence properties of such an algorithm might be.

The robustification of SPML is conceptually more straightforward. First of all, the algorithm consists of simply optimizing a nonlinear function—albeit a complex one. Second, the SPML objective function resembles the likelihood of a linear mixed effects model, for which careful work on robust estimation has been published.

Nevertheless, preliminary results suggest that, at least in some situations, the performance of the two algorithms may be similar. Further comparison of the two procedures

both theoretically and using simulations is clearly needed.

Chapter 6

Conclusions

This thesis has explored a number of topics in hierarchical modeling from a meta-analytic perspective. In Section 6.1, we review the new methods developed here. Questions involving spawner-recruitment modeling motivated this work, and in Section 6.2 we discuss the context of our methods in conservation biology and fisheries management. Finally, in Section 6.3, we discuss some interesting avenues of future inquiry.

6.1 Summary

In this work, we have investigated hierarchical models for raw-data meta-analysis. Since meta-analysis introduces an additional level of modeling, the hazards are many and a cautious approach is called for. Our focus has therefore been on graphical and robust methods.

6.1.1 Graphical methods

Graphical methods can help orient the user and reveal structures in the data that may be of central importance. The *raindrop plot*, introduced in Section 2.6, is a generalization of the usual meta-analytic plot showing study-specific point estimates and error bars. As we have seen, in some cases, individual point estimates may not exist. Even if they do, the normal theory that often underlies confidence intervals may be a poor approximation. Raindrop plots overcome these difficulties and provide a compact, informative display. An extension of the raindrop plot may be used to display probability distributions such as random

effect distributions, posterior and predictive distributions, and bootstrap distributions. For multi-parameter problems, raindrop plots for the parameter of interest may be based on approximate likelihoods with nuisance parameters eliminated, as discussed in Section 2.5. However, approximate likelihoods remain a topic of research, and in some cases may be quite misleading. In two-dimensional problems, for example, joint likelihood contours may be preferable.

Several other novel graphical displays were also introduced. The whisker-weights display of weights (e.g., from a robust regression analysis) in the margins of a scatter plot is, as far as I know, a new idea. An alternative is to plot variable-size points to indicate the relative weights, however this may distract the viewer from the relationship in the scatter plot.

The “mushroom plot” of Figure 4.5 is a novel way of considering the consequences of parametric estimates of random effect distributions.

6.1.2 Robust methods

Statistical models are just that: models; we do not expect that our assumptions will be met exactly. For this reason, good statistical procedures should not be highly sensitive to slight deviations from assumptions. Robust approaches to estimation for linear mixed effects models have recently been developed. In this work, we tried to extend these approaches to nonlinear mixed effects models. There is much still to be done in this area, however the two proposed approaches seem promising. The conceptually simpler approach is our proposed modification to Nuñez’s SPML for estimating nonlinear mixed effects models. Incorporating of a ρ -function in the pseudo likelihood may result in an asymptotic bias of unknown magnitude. However the results of (admittedly limited) simulations suggest that this is not a large problem. The Lindstrom and Bates algorithm and variants are widely implemented and less computationally intensive than SPML. However the asymptotic properties of this approach and related approaches based on Laplace approximations may be of concern. Our proposed robustification of their alternating two-stage algorithm involves three different ρ -functions and a consistency correction. Choice of these functions is a concern and properties of the resulting estimates unclear.

6.1.3 Other contributions

We have made two other contributions in this work. One is the developments of generalizations of the “hockey stick” segmented regression model for spawner-recruitment modeling. The abrupt bend in the ordinary hockey stick is problematic in several respects. First, it is hard to believe that a natural population would actually follow so abrupt a relationship. Second, the bend can cause numerical difficulties in estimation, such as multiple local maxima in the likelihood surface. Third, the abruptness of the bend makes the hockey-stick model highly nonlinear, and seems to cause difficulties for the Lindstrom and Bates algorithm. We proposed two “generalized hockey sticks”: the quadratic hockey stick, which has a simple form, and the logistic hockey stick, which is smoother but less easily interpreted.

Another contribution is our proposed modification of Nuñez’s algorithm, namely the use of stylized normal samples in place of simulated normal samples. As intended, this reduces the variability in the estimates.

6.2 Mixed effects models for spawner-recruitment data

Our use of mixed effects models for raw-data meta-analysis was originally motivated by interest in fish population dynamics and the possibility to investigate previously unanswered questions using the large spawner-recruitment database compiled by Myers, Bridson, and Barrowman (1995).

Our main conclusion is that appropriate statistical methods can use information from other studies to estimate spawner-recruitment parameters in local situations where these are poorly defined. In the coho salmon example, this method allows spawner-recruitment parameters to be fitted for a given stock even though the data for that stock by itself give implausible estimates. The empirical Bayes estimates are clearly superior in these cases to the individual fits, but it is possible that they may shrink the estimates too close to the mean in some cases. In other words, too much weight may be given to the population estimate for the random effect, and not enough weight to the individual stock estimate. Nonetheless, this finding should be useful when designing management strategies for stocks whose population dynamics are too variable to allow reliable resolution of spawner-recruitment curves.

We view the statistical methodology as part of a process, the endpoint of which is a

scientific or management decision. For example, the output of the statistical component may be used as input to a risk assessment. Here we briefly consider the application of results obtained using our methods to extinction and management models.

6.2.1 Application to extinction models

Considerable effort has been devoted to the development of both analytical and simulation models that estimate extinction probabilities of natural populations, (Lande 1993; Ludwig 1996; Fagan, Meir, and Moore 1999). On the whole, these models suffer from the lack of plausible parameter values, as there is often very little data available. Instead, parameters are drawn from arbitrary distributions and the conclusions remain broad (Foley 1994; Johst and Wissel 1997). This problem is particularly acute for parameters that describe population dynamics at low population sizes. Moreover, it is the dynamics at low sizes that are of greatest import when estimating extinction risk.

Meta-analytic techniques provide a way to obtain estimates for populations where little is known. In the case of the coho salmon, I was able to obtain estimates of α , the maximum reproductive rate at low population size (and thus critical to predicting extinction) for each stream, as well as an estimate of the variance of α . This information could be incorporated directly into an extinction model, thus overcoming the difficulties highlighted by Routledge and Irvine (1999) about imprecise predictions.

6.2.2 Application to management models

Recent international and national management regimes rely heavily on the goal of fishing such that the biomass (i.e. the quantity of spawners) is at or above the level needed to produce *maximum sustainable yield*—the maximum harvest that can be removed from a population on a sustainable basis. For example, the UN Convention of Straddling Fish Stocks and Highly Migratory Fish Stocks requires states that sign the agreement to rebuild stocks to the “biomass that would produce maximum sustainable yield.” Similar language is in the Sustainable Fisheries Act passed by the US Congress. It thus becomes crucial to estimate the biomass level at which maximum yield can be sustained, and thus the capacity of ecosystems to produce fish, in the case of the coho salmon, the R_i^{\max} for a stream. Traditional methods that rely on oversimplified models that apply to only single stocks

usually produce very unreliable, and uncertain estimates (Hilborn and Walters 1992). The approach described here allows much improved estimates of R_i^{\max} and α which are needed to estimate maximum sustainable yield.

Our analysis also provides estimates of the maximum reproductive rate, α , which forms the basis of most calculations of management reference points (Mace and Doonan 1988; Mace 1994; Hilborn and Walters 1992). These estimates are certainly far superior to the individual point estimates, but it remains to be seen if they are generally superior to the estimates obtained from a meta-analysis using the Ricker model (Myers, Bowen, and Barrowman 1999). Simulation studies are required to determine this.

6.3 Future research

This work raises a considerable number of questions of both a statistical and a biological nature.

The graphical methods we have introduced deserve further study. For example, several variants of the raindrop plot were suggested; it is not clear which is preferable, or whether another method might be better.

An interesting feature was observed in the analysis of both the wolf data and the coho salmon data: observations that appear to be quite unusual in fits to individual studies seem much less so in a mixed effects model. The proposed robust procedures for fitting linear and nonlinear mixed effects models resulted in very mild downweighting of such observations. It appears that in addition to the well-known advantage of shrinkage, mixed models seem to avoid “overfitting” individual studies, thereby mitigating the effects of unusual observations. This certainly deserves further study. A related question is how best to obtain robust estimates of realized random effects.

Our use of stylized normal samples instead of simulated normal samples in Nuñez’s algorithm raises some questions. First, is there an optimal way to generate stylized bivariate normal samples? The spiral approach of Figure 5.5 (p. 133) seemed to outperform the radial approach, but both were somewhat ad hoc.

In Section 5.2.7, we briefly described several resampling methods that could be used to estimate standard errors for parameter estimates in nonlinear mixed effects models for repeated measures data. In practice we used five procedures: a parametric bootstrap, a simple

jackknife in which each study in turn was left out, a multihalver jackknife method, and two types of “marginal” bootstraps. Little has been published concerning resampling methods for hierarchical data and an investigation of the performance of the various methods would be very useful.

Another topic of interest is model building for nonlinear mixed effects models. This includes issues like how to select covariates, whether to use correlated or uncorrelated random effects, and so forth. Another question concerns whether study-specific effects should be fixed or random. Kiefer and Wolfowitz (1956) showed that the use of random effects instead of study-specific fixed effects is generally preferable. Jiang (1996) showed that the asymptotic properties of estimates for models with study-specific fixed effects are very different from the usual asymptotics.

Our proposed robust modifications to algorithms for estimation of nonlinear mixed effects models raise a number of questions. For example, the asymptotic properties of the proposed modifications deserve more attention. From a practical standpoint, it would be useful to obtain bias corrections.

Other questions concern analogues of REML estimators for nonlinear mixed effects models. Lindstrom and Bates defined approximate REML estimators for nonlinear mixed effects models. Richardson and Welsh (1995) developed Robust REML methods (I and II) for linear mixed effects models. It would be interesting to try to adapt these robust REML methods to the Lindstrom and Bates algorithm.

In this work, I did not consider robustness against the effect of leverage points. Richardson (1997) developed estimation methods for linear mixed effects models that provide robustness against leverage points and it would be useful to adapt these methods to nonlinear mixed effects models.

In addition to the statistical issues raised by this work, there are a number of interesting biological questions. As is well known, statistically significant results are not always biologically important. For example, fits to two spawner recruitment models may be significantly different but have essentially identical management consequences. Conversely, fits may be quite similar but have radically different management consequences.

One extension of the spawner-recruitment models considered in this thesis is the introduction of *depensation*, a decrease in recruitment per spawner at low levels of spawners. Depensation is predicted by several biological models and could have drastic consequences

for populations reduced to low levels by overfishing. Myers, Barrowman, Hutchings, and Rosenberg (1995) performed a fixed-effects raw-data meta-analysis of spawner-recruitment datasets for 128 fish populations using a modified Beverton-Holt model and found little evidence of depensation. However, a mixed model analysis seems more appropriate and recent work suggests there may be depensation, albeit to a fairly limited extent.

Bibliography

- Agresti, A. 1990. *Categorical Data Analysis*. Wiley, New York.
- Barndorff-Nielsen, O. 1983. On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**: 343–365.
- Barrowman, N. J., and Myers, R. A. 2000. Still more spawner-recruitment curves: The hockey stick and its generalizations. *Canadian Journal of Fisheries and Aquatic Sciences* **57**: 665–676.
- Bates, D. M., and Watts, D. G. 1988. *Nonlinear regression analysis and its applications*. Wiley, New York.
- Beal, S. L. 1984. Asymptotic properties of optimization estimators for the independent not identically distributed case with applications to extended least squares estimators. Technical report, University of California, San Francisco, CA.
- Beamish, R. J., Mcfarlane, G. A., and Thomson, R. E. 1999. Recent declines in the recreational catch of coho salmon (*Oncorhynchus kisutch*) in the Strait of Georgia are related to climate. *Canadian Journal of Fisheries and Aquatic Sciences* **56**: 506–516.
- Berger, J. O., Liseo, B., and Wolpert, R. L. 1999. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**: 1–28.
- Bijnens, L., Collette, L., Ivanov, A., Boes, G. H., and Sylvester, R. 1996. Optimal graphical display of the results of meta-analyses of individual patient data. 1996 Cochrane Colloquium.
- Birnbaum, A. 1961. A unified theory of estimation, I. *Annals of Mathematical Statistics* **32**: 112–135.

- Box, G. E. P. 1979. Robustness is the strategy of scientific model building. In R. Launer and G. Wilkinson (Eds.), *Robustness in Statistics*. Academic Press.
- Box, G. E. P., and Tiao, G. C. 1973. *Bayesian Inference in Statistical Analysis*. Wiley, New York.
- Bradford, M. J. 1999. Temporal and spatial trends in the abundance of coho salmon smolts from western North America. *Transactions of the American Fisheries Society* **128**: 840–846.
- Bradford, M. J., Myers, R. A., and Irvine, J. R. 2000. Reference points for coho salmon harvest rates and escapement goals based on freshwater production. *Canadian Journal of Fisheries and Aquatic Sciences* **57**: 677–686.
- Breslow, N. E., and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9–25.
- Carlin, B. P., and Gelfand, A. E. 1990. Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association* **85**: 105–114.
- Cleveland, W. S. 1985. *The elements of graphing data*. Wadsworth Advanced Books and Software, Monterey, California.
- Cooper, H., and Hedges, L. V. 1994. *The Handbook of Research Synthesis*. Russell Sage Foundation, New York.
- Cox, D. R. 1982. Combination of data. In S. Kotz, N. L. Johnson, and C. B. Read (Eds.), *Encyclopedia of Statistical Sciences, Volume 2*. Wiley, New York, pp. 45–53.
- Cressie, N., and Lahiri, S. N. 1993. The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis* **45**: 217–233.
- Daniels, M. J. 1999. A prior for the variance in hierarchical models. *The Canadian Journal of Statistics* **27**: 567–578.
- Das, S., and Krishen, A. 1999. Some bootstrap methods in nonlinear mixed-effect models. *Journal of Statistical Planning and Inference* **75**: 237–245.
- Davidian, M., and Giltinan, D. M. 1995. *Nonlinear models for repeated measurement data*. Chapman and Hall, New York.

- Davison, A. C., and Hinkley, D. V. 1997. *Bootstrap methods and their application*. Cambridge University Press, Cambridge.
- Deely, J. J., and Lindley, D. V. 1981. Bayes empirical Bayes. *Journal of the American Statistical Association* **76**: 833–841.
- Efron, B. 1996. Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association* **91**: 538–565.
- Eysenck, H. J. 1978. An exercise in mega-silliness. *American Psychologist* **33**: 517.
- Fagan, W. F., Meir, E., and Moore, J. L. 1999. Variation thresholds for extinction and their implications for conservation strategies. *American Naturalist* **154**: 510–520.
- Fellner, W. H. 1986. Robust estimation of variance components. *Technometrics* **28**: 51–60.
- Foley, P. 1994. Predicting extinction times from environmental stochasticity and carrying capacity. *Conservation Biology* **8**: 124–137.
- Galbraith, R. F. 1988. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine* **7**: 889–894.
- Gourieroux, C., and Monfort, A. 1993. Simulation-based inference: a survey with special reference to panel data models. *Journal of Econometrics* **59**: 5–33.
- Gourieroux, C., Monfort, A., and Trognon, A. 1984. Pseudo maximum likelihood methods: theory. *Econometrica* **52**: 681–700.
- Hampel, F. R., Rochetti, E. M., Rousseeuw, P. J., and Stahel, W. A. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.
- Hartley, H. O., and Rao, J. N. K. 1967. Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**: 93–108.
- Harville, D. A. 1974. Bayesian inference for variance components using only error contrasts. *Biometrika* **61**: 383–385.
- Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**: 320–340.

- Hedges, L. V., and Olkin, I. 1980. Vote-counting methods in research synthesis. *Psychological Bulletin* **88**: 359–369.
- Hedges, L. V., and Olkin, I. 1985. *Statistical Methods for Meta-analysis*. Academic Press, San Diego.
- Henderson, C. R. 1963. Selection index and expected genetic advance. In W. D. Hanson and H. F. Robinson (Eds.), *Statistical Genetics and Plant Breeding*. National Academy of Sciences and National Research Council Publication No. 982, Washington, D.C., pp. 141–163.
- Henderson, C. R. 1973. Maximum likelihood estimation of variance components.
- Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. N. 1959. Estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**: 192–218.
- Hilborn, R., and Mangel, M. 1997. *The Ecological Detective*. Princeton University Press, Princeton, New Jersey.
- Hilborn, R., and Walters, C. J. 1992. *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty*. Chapman and Hall, New York.
- Hintze, J. L., and Nelson, R. D. 1998. Violin plots: a box plot-density trace synergism. *The American Statistician* **52**: 181–184.
- Hoadley, B. 1971. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Annals of Mathematical Statistics* **42**: 1977–1991.
- Hobbes, T. 1651. *Leviathan*.
- Hobert, J. P., and Casella, G. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed MOdels. *Journal of the American Statistical Association* **91**: 1461–1473.
- Huggins, R. M. 1993a. On the robust analysis of variance components models for pedigree data. *The Australian Journal of Statistics* **35**: 43–57.
- Huggins, R. M. 1993b. A robust approach to the analysis of repeated measures. *Biometrics* **49**: 715–720.

- Hutchings, J. A., and Morris, D. W. 1985. The influence of phylogeny, size and behaviour on patterns of covariation in salmonid life histories. *Oikos* **45**: 118–124.
- Hyndman, R. J. 1996. Computing and graphing highest density regions. *The American Statistician* **50**: 120–126.
- Jennrich, R. 1969. Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics* **40**: 633–643.
- Jiang, J. 1996. REML estimation: asymptotic behavior and related topics. *Annals of Statistics* **24**: 255–286.
- Johst, K., and Wissel, C. 1997. Extinction risk in a temporally correlated fluctuating environment. *Theoretical Population Biology* **52**: 91–100.
- Kass, R. E., and Steffey, D. 1989. Approximate Bayesian inference in conditional independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* **84**: 717–726.
- Kiefer, J., and Wolfowitz, J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**: 887–906.
- Laird, N. M., and Louis, T. A. 1987. Empirical Bayes confidence intervals based on bootstrap samples. with discussion and with a reply by the authors. *Journal of the American Statistical Association* **82**: 739–757.
- Laird, N. M., and Ware, J. H. 1982. Random effects models for longitudinal data. *Biometrics* **38**: 963–974.
- Lande, R. 1993. Risks of population extinction from demographic and environmental stochasticity and random catastrophes. *American Naturalist* **142**: 911–927.
- Larivière, S., Jolicoeur, H., and Crête, M. 2000. Status and conservation of the gray wolf (*Canis lupus*) in wildlife reserves of Québec. *Biological Conservation* **94**: 143–151.
- Lee, J. J., and Tu, Z. N. 1997. A versatile one-dimensional distribution plot: the BLiP plot. *The American Statistician* **51**: 353–358.
- Lerman, P. M. 1980. Fitting segmented regression models by grid search. *Applied Statistics* **29**: 77–84.

- Liang, K.-Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
- Light, R. J., Singer, J. D., and Willett, J. B. 1994. The visual presentation and interpretation of meta-analyses. In H. Cooper and L. V. Hedges (Eds.), *The Handbook of Research Synthesis*. Russell Sage Foundation, New York, pp. 439–453.
- Lindstrom, M. J., and Bates, D. M. 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**: 673–687.
- Ludwig, D. 1996. Uncertainty and the assessment of extinction probabilities. *Ecological Applications* **6**: 1067–1076.
- Mace, P. M. 1994. Relationships between common biological reference points used as threshold and targets of fisheries management strategies. *Canadian Journal of Fisheries and Aquatic Sciences* **51**: 110–122.
- Mace, P. M., and Doonan, I. J. 1988. A generalized bioeconomic simulation model for fish population dynamics. *N. Z. Fish. Assess. Res. Doc.* **88/4**.
- McCarthy, P. J. 1969. Pseudo-replication: half samples. *Review of the International Statistical Institute* **37**: 239–264.
- McFadden, D. 1989. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* **57**: 995–1026.
- Miettinen, O. S. 1985. *Theoretical epidemiology*. Wiley, New York.
- Miller, J. J. 1977. Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Annals of Statistics* **5**: 746–762.
- Morris, C. N. 1996. Comment on “Empirical Bayes Methods for Combining Likelihoods”. *Journal of the American Statistical Association* **91**: 555–558.
- Myers, R. A., Barrowman, N. J., Hutchings, J. A., and Rosenberg, A. A. 1995. Population dynamics of exploited fish stocks at low population levels. *Science* **269**: 1106–1108.
- Myers, R. A., Bowen, K. G., and Barrowman, N. J. 1999. Maximum reproductive rate of fish at low population sizes. *Canadian Journal of Fisheries and Aquatic Sciences* **56**: 2404–2419.

- Myers, R. A., Bridson, J., and Barrowman, N. J. 1995. Summary of worldwide stock and recruitment data. *Canadian Technical Report of Fisheries and Aquatic Sciences* **2024**: iv + 327.
- Myers, R. A., and Mertz, G. 1998. The limits of exploitation: a precautionary approach. *Ecological Applications* **8, Supplement 1**: S165–S169.
- Myers, R. A., Mertz, G., and Barrowman, N. J. 1995. Spatial scales of variability in cod recruitment in the North Atlantic. *Canadian Journal of Fisheries and Aquatic Sciences* **52**: 1849–1862.
- Myers, R. A., Mertz, G., and Bridson, J. M. 1997. Spatial scales of interannual recruitment variations of marine, anadromous, and freshwater fish. *Canadian Journal of Fisheries and Aquatic Sciences* **54**: 1400–1407.
- Normand, S.-L. T. 1999. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**: 321–359.
- Nuñez, O. 1998. Propriétés asymptotiques d'un estimateur dans un modèle non linéaire à effets mixtes. *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique* **326**: 381–384.
- Nuñez, O., and Concordet, D. 2000. A simulated pseudo-maximum likelihood estimator for nonlinear mixed effects models. *Biometrika* **In press**.
- Olkin, I. 1999. Diagnostic statistical procedures in medical meta-analysis. *Statistics in Medicine* **18**: 2331–2341.
- Patterson, H. D., and Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545–554.
- Patterson, H. D., and Thompson, R. 1974. Maximum likelihood estimation of components of variance. In *Proceedings of the 8th International Biometric Conference*. pp. 197–207.
- Pinheiro, J. C. 1994. Topics in Mixed Effects Models. Ph. D. thesis, University of Wisconsin, Madison.
- Pinheiro, J. C., and Bates, D. M. 1995. Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics* **4**: 12–35.

- Platt, R. W. 1998. Estimation using the modified profile likelihood in log odds ratio regression analysis. *Communications in Statistics — Simulation* **27**: 905–919.
- Platt, R. W., Leroux, B. G., and Breslow, N. 1999. Generalized linear mixed models for meta-analysis. *Statistics in Medicine* **18**: 643–654.
- Quenouille, M. H. 1949. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society series B* **11**: 68–84.
- Ramos, R. Q., and Pantula, S. G. 1995. Estimation of nonlinear random coefficient models. *Statistics & Probability Letters* **24**: 49–56.
- Richardson, A. M. 1997. Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association* **92**: 154–162.
- Richardson, A. M., and Welsh, A. H. 1994. Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *The Australian Journal of Statistics* **36**: 31–43.
- Richardson, A. M., and Welsh, A. H. 1995. Robust restricted maximum likelihood in mixed linear models. *Biometrics* **51**: 1429–1439.
- Robinson, G. K. 1991. That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**: 15–51. With comments and a rejoinder by the author.
- Routledge, R. D., and Irvine, J. R. 1999. Chance fluctuations and the survival of small salmon stocks. *Canadian Journal of Fisheries and Aquatic Sciences* **56**: 1512–1519.
- Sacks, H. S., Chalmers, T. C., Blum, A. L., Berrier, J., and Pagano, D. 1990. Endoscopic hemostasis, an effective therapy for bleeding peptic ulcers. *Journal of the American Medical Association* **264**: 494–499.
- Searle, S. R., Casella, G., and McCulloch, C. E. 1992. *Variance Components*. Wiley, New York.
- Sheiner, L. B., and Beal, S. L. 1980. Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics* **8**: 553–571.
- Smith, A. F. M. 1983. Comment on: “Bayes methods for combining the results of cancer studies in humans and other species” [*J. Amer. Statist. Assoc.* **78** (1983), 293–308]

- by William H. DuMouchel and Jeffrey E. Harris. *Journal of the American Statistical Association* **78**: 310–311.
- Smith, T. C., Spiegelhalter, D. J., and Thomas, A. 1995. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14**: 2685–2699.
- Stahel, W. A., and Welsh, A. 1997. Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference* **57**: 295–319.
- Staudte, R. G., and Sheather, S. J. 1990. *Robust estimation and testing*. Wiley, New York.
- Steimer, J.-L., Mallet, A., Golmard, J., and Boisvieux, J. 1984. Alternative approaches to estimation of population pharmacokinetic parameters: comparison with the non-linear mixed effect model. *Drug Metabolism Reviews* **15**: 265–292.
- Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press, Cambridge, MA.
- Tishler, A., and Zang, I. 1981. A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association* **76**: 980–987.
- Tukey, J. W. 1987. Kinds of bootstraps and kinds of jackknives, discussed in terms of a year of weather-related data. Technical Report 292, Princeton University.
- van Houwelingen, H. C., and Zwinderman, K. H. 1993. A bivariate approach to meta-analysis. *Statistics in Medicine* **12**: 2273–2284.
- Verbyla, A. P. 1990. A conditional derivation of residual maximum likelihood. *The Australian Journal of Statistics* **32**: 227–230.
- Vonesh, E. F. 1996. A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika* **83**: 447–452.
- Vonesh, E. F., and Carter, R. L. 1992. Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* **48**: 1–17.
- Walters, C. J. 1993. Where have all the coho gone? In L. Berg and P. Delaney (Eds.), *Proceedings of the Coho Workshop, Nanaimo, B.C., May 26-28, 1992*. pp. 1–8.

- Walters, C. J., and Ward, B. 1998. Is solar radiation responsible for declines in marine survival rates of anadromous salmonids that rear in small streams? *Canadian Journal of Fisheries and Aquatic Sciences* **55**: 2533–2538.
- Welsh, A. H., and Richardson, A. M. 1997. Approaches to the robust estimation of mixed models. In G. S. Maddala and C. R. Rao (Eds.), *Handbook of Statistics*, Vol. 15. Elsevier Science B. V., pp. 343–384.
- Westfall, P. H. 1986. Asymptotic normality of the ANOVA estimates of components of variance in the nonnormal, unbalanced hierarchical mixed model. *Annals of Statistics* **14**: 1572–1582.
- White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* **50**: 1–25.
- White, H. 1994. *Estimation, inference and specification analysis*. Cambridge University Press, Cambridge.
- Wolfinger, R. 1993. Laplace's approximation for nonlinear mixed models. *Biometrika* **80**: 791–795.
- Wong, L. C. 1999. Estimating genetic variance components with uncertainty in full-sibling relationships. Master's thesis, Dalhousie University, Halifax, Nova Scotia.