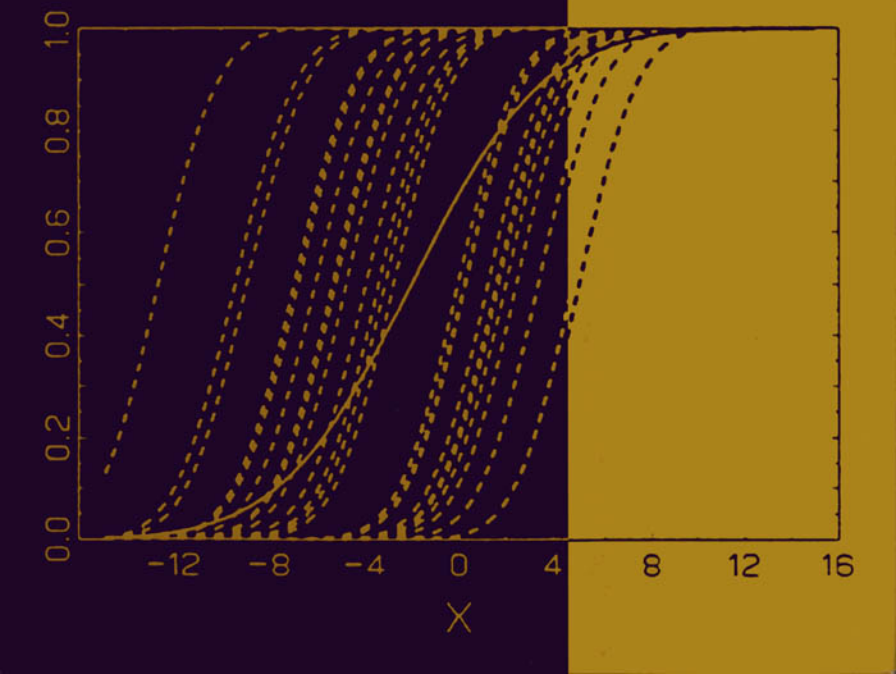
Wiley Series in Probability and Statistics

Generalized, Linear, and Mixed Models

Charles E. McCulloch, Shayle R. Searle



Generalized, Linear, and Mixed Models

WILEY SERIES IN PROBABILITY AND STATISTICS
TEXTS, REFERENCES, AND POCKETBOOKS SECTION

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, David W. Scott, Bernard W. Silverman, Adrian F. M. Smith, Jozef L. Teugels; Vic Barnett, Emeritus, Ralph A. Bradley, Emeritus, J. Stuart Hunter, Emeritus, David G. Kendall, Emeritus*

A complete list of the titles in this series appears at the end of this volume.

Generalized, Linear, and Mixed Models

CHARLES E. McCULLOCH
SHAYLE R. SEARLE
Departments of Statistical Science and Biometrics
Cornell University

A WILEY-INTERSCIENCE PUBLICATION
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Library of Congress Cataloging-in-Publication Data

McCulloch, Charles E.
Generalized, linear, and mixed models / Charles E. McCulloch, Shayle R. Searle.
p. cm. — (Wiley series in probability and statistics. Texts and references section)
Includes bibliographical references and index.
ISBN 0-471-19364-X (cloth : alk. paper)
1. Linear models (Statistics). I. Searle, S. R. (Shayle R.), 1928–. II. Title. III. Wiley series in probability and statistics. Texts and references section.
QA279.M3847 2000
519.5'35—dc21 00-059430

Printed in the United States of America.

10 9 8 7 6 5 4 3

List of Chapters

Contents

Preface

The last thirty or so years have been a time of enormous development of analytic results for the linear model (LM). This has generated extensive publication of books and papers on the subject. Much of this activity has focused on the normal distribution and homoscedasticity. Even for unbalanced data, many useful, analytically tractable results have become available. Those results center largely around analysis of variance (ANOVA) procedures, and there is abundant computing software which will, with wide reliability, compute those results from submitted data.

Also within the realm of normal distributions, but permitting heterogeneity of variance, there has been considerable work on linear mixed models (LMMs) wherein the variance structure is based on random effects and their variance components. Algebraic results in this context are much more limited and complicated than with LMs. However, with the advent of readily available computing power and the development of broadly applicable computing procedures (e.g., the EM algorithm) we are now at a point where models such as the LMM are available to the practitioner. Furthermore, models that are nonlinear and incorporate non-normal distributions are now feasible. It is to understanding these models and appreciating the available computing procedures that this book is directed.

We begin by reviewing the basics of LMs and LMMs, to serve as a starting point for proceeding to generalized linear models (GLMs), generalized linear mixed models (GLMMs) and some nonlinear models. All of these are encompassed within the title "Generalized, Linear, and Mixed Models."

The progress from easy to difficult models (e.g. from LMs to GLMMs) necessitates a certain repetition of basic analysis methods, but this is appropriate because the book deals with a variety of models and the application to them of standard statistical methods. For example, max-

imum likelihood (ML) is used in almost every chapter, on models that get progressively more difficult as the book progresses. There is, indeed, purposeful concentration on ML and, very noticeably, an (almost complete) absence of analysis of variance (ANOVA) tables.

Although analysis of variance methods are quite natural for fixed effects linear models with normal distributions, even in the case of linear mixed models with normal distributions they have much less appeal. For example, with unbalanced data from mixed models, it is not clear what the "appropriate" ANOVA table should be. Furthermore, from a theoretical viewpoint, any such table represents an over-summarization of data: except in special cases, it does not contain sufficient statistics and therefore engenders a loss of information and efficiency. And these deficiencies are aggravated if one tries to generalize analysis of variance to models based on non-normal distributions such as, for example, the Poisson or binomial. To deal with these we therefore concentrate on ML procedures.

Although ML estimation under non-normality is limited in yielding analytic results, we feel that its generality and efficiency (at least with large samples) make it a natural method to use in today's world. Today's computing environment compensates for the analytic intractability of ML and helps makes ML more palatable.

As prelude to the application of ML to non-normal models we often show details of using it on models where it yields easily interpreted analytic results. The details are lengthy, but studying them engenders a confidence in the ML method that hopefully carries over to non-normal models. For these, the details are often not lengthy, because there are so few of them (as a consequence of the model's inherent intractability) and they yield few analytic results. The brevity of describing them should not be taken as a lack of emphasis or importance, but merely as a lack of neat, tidy results. It is a fact of modern statistical practice that computing procedures are used to gain numerical information about the underlying nature of algebraically intractable results. Our aim in this book is to illuminate this situation.

The book is intended for graduate students and practicing statisticians. We begin with a chapter in which we introduce the basic ideas of fixed and random factors and mixed models and briefly discuss general methods for the analysis of such models. Chapters 2 and 3 introduce all the main ideas of the remainder of the book in two simple contexts (one-way classifications and linear regression) with a minimum of em-

phasis on generality of results and notation. These three chapters could form the core of a quarter course or, with supplementation, the basis of a semester-long course for Master's students. Alternatively, they could be used to introduce generalized mixed models towards the end of a linear models class.

Chapters 4, 5, 6 and 8 cover the main classes of models (linear, generalized linear, linear mixed, and generalized linear mixed) in more generality and breadth. Chapter 7 discusses some of the special features of longitudinal data and shows how they can be accommodated within LMMs. Chapter 9 presents the idea of prediction of realized values of random effects. This is an important distinction introduced by considering models containing random effects. Chapter 10 covers computing issues, one of the main barriers to adoption of mixed models in practice. Lest the reader think that everything can be accommodated under the rubric of the generalized linear mixed model, Chapter 11 briefly mentions nonlinear mixed models. And the book ends with two short appendices, M and S, containing some pertinent results in matrices and statistics.

For students with some training in linear models, the first 10 chapters, with light emphasis on Chapters 1 through 4 and 6, could form a "second" course extending their linear model knowledge to generalized linear models. Of course, the book could also be used for a semester long course on generalized mixed models, although in-depth coverage of all of the topics would clearly be difficult.

Our emphasis throughout is on modeling and model development. Thus we provide important information about the consequences of model assumptions, techniques of model fitting and methods of inference which will be required for data analysis, as opposed to data analysis itself. However, to illustrate the concepts we do also include analysis or illustration of the techniques for a variety of real data sets.

The chapters are quite variable in length, but all of them have sections, subsections and sub-subsections, each with its own title, as shown in the Table of Contents. At times we have sacrificed the flow of the narrative to make the book more accessible as a reference. For example, Section 2.1d is basically a catalogue of results with titles that make retrieval more straightforward, particularly because those titles are all listed in the table of contents.

Ithaca, NY
September 2000

Charles E. McCulloch
Shayle R. Searle

This page intentionally left blank

Chapter 1

INTRODUCTION

1.1 MODELS

a. Linear models (LM) and linear mixed models (LMM)

In almost all uses of statistics, major interest centers on averages and on variation about those averages. For more than sixty years this interest has frequently manifested itself in the widespread use of analysis of variance (ANOVA), as originally developed by R. A. Fisher. This involves expressing an observation as a sum of a mean plus differences between means, which, under certain circumstances, leads to methods for making inferences about means or about the nature of variability. The usually-quoted set of circumstances which permits this is that the mean of each datum be taken as a linear combination of unknown parameters, considered as constants; and that the data be deemed to have come from a normal distribution. Thus the linear requirement is such that the expected value (i.e., mean), μ_{ij}, of an observation y_{ij} can be, for example, of the form $\mu_{ij} = \mu + \alpha_i + \beta_j$ where μ, α_i and β_j are unknown constants—unknown, but which we are interested in estimating. And the normality requirement would be that y_{ij} is normally distributed with mean μ_{ij}. These requirements are the essence of what we call a *linear model*, or LM for short. By that we mean that the model is linear in the parameters, so "linearity" also includes being of the form $\mu_{ij} = b_0 + b_1 x_{1ij} + b_2 x_{2ij}^2$, for example, where the xs are known and there can be (and often are) more than two of them.

A variant of LMs is where parameters in an LM are treated not as constants but as (realizations of) random variables. To denote this different meaning we represent parameters treated as random by Roman

rather than Greek letters. Thus if the αs in the example were to be considered random, they would be denoted by as, so giving $\mu_{ij} = \mu + a_i + \beta_j$. With the βs remaining as constants, μ_{ij} is then a mixture of random and constant terms. Correspondingly, the model (which is still linear) is called a *linear mixed model*, or LMM. Until recently, most uses of such models have involved treating random a_is as having zero mean, being homoscedastic (i.e., having equal variance) with variance σ_a^2 and being uncorrelated. Additionally, normality of the a_is is usually also invoked.

There are many books dealing at length with LMs and LMMs. We name but a few: Graybill (1976), Seber (1977), Arnold (1981), Hocking (1985), Searle (1997), and Searle et al. (1992).

b. Generalized models (GLMs and GLMMs)

The last twenty-five years or so have seen LMs and LMMs extended to *generalized linear models* (GLMs) and to *generalized linear mixed models* (GLMMs). The essence of this generalization is two-fold: one, that data are not necessarily assumed to be normally distributed; and two, that the mean is not necessarily taken as a linear combination of parameters but that some function of the mean is. For example, count data may follow a Poisson distribution, with mean λ, say; and $\log \lambda$ will be taken as a linear combination of parameters. If all the parameters are considered as fixed constants the model is a GLM; if some are treated as random it is a GLMM.

The methodology for applying a GLM or a GLMM to data can be quite different from that for an LM or LMM. Nevertheless, some of the latter is indeed a basis for contributing to analysis procedures for GLMs and GLMMs, and to that extent this book does describe some of the procedures for LMs (Chapter 4) and LMMs (Chapter 6); Chapters 5 and 8 then deal, respectively, with GLMs and GLMMs. Chapters 2 and 3 provide details for the basic modeling of the one-way classification and of regression, prior to the general cases dealt with later.

1.2 FACTORS, LEVELS, CELLS, EFFECTS AND DATA

We are often interested in attributing the variability that is evident in data to the various categories, or classifications, of the data. For example, in a study of basal cell epithelioma sites (akin to Abu-Libdeh et al., 1990), patients might be classified by gender, age-group and

Table 1.1: A Format for Summarizing Data

	Low Exposure to Sunshine			High Exposure to Sunshine		
Gender	Age Group			Age Group		
	A	B	C	A	B	C
Male						
Female						

Table 1.2: Summarizing Exam Grades

	English			Geology		
Gender	Section			Section		
	A	B	C	A	B	C
Male						
Female						

extent of exposure to sunshine. The various groups of data could be summarized in a table such as Table 1.1.

The three classifications, gender, age, and exposure to sunshine, which identify the source of each datum are called *factors*. The individual classes of a classification are the *levels* of a factor (e.g., male and female are the two levels of the factor "gender"). The subset of data occurring at the "intersection" of one level of every factor being considered is said to be in a *cell* of the data. Thus with the three factors, gender (2 levels), age (3 levels) and sunshine (2 levels), there are $2 \times 3 \times 2 = 12$ cells.

Suppose that we have student exam grades from each of three sections in English and geology courses. The data could be summarized as in Table 1.2, similar to Table 1.1. Although the layout of Table 1.2 has the same appearance as Table 1.1, sections in Table 1.2 are very different from the age groups of Table 1.1. In Table 1.2 section A of English has no connection to (and will have different students from) section A of geology; in the same way neither are sections B (or C) the same in the two subjects. Thus the section factor is *nested* within the subject factor. In contrast, in Table 1.1 the three age groups are the same for both low and high exposures to sunshine. The age and sunshine factors are said to be *crossed*.

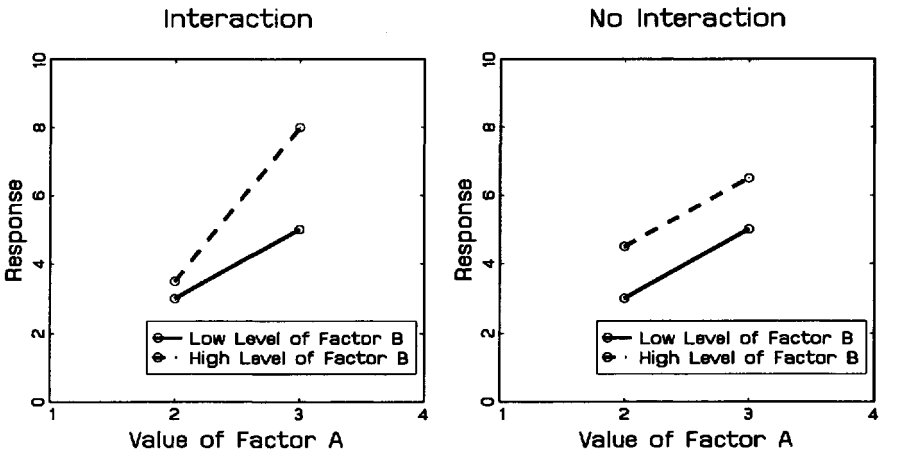
In classifying data in terms of factors and their levels, the feature of interest is the extent to which different levels of a factor affect the variable of interest. We refer to this as the *effect* of a level of a factor on that variable. The effects of a factor are always one or other of two kinds, as introduced in Section 1.1 in terms of parameters. First is the case of parameters being considered as fixed constants or, as we henceforth call them, *fixed effects*. These are the effects attributable to a finite set of levels of a factor that occur in the data and which are there because we are interested in them. In Table 1.1 the effects for all three factors are fixed effects.

The second case corresponds to what we earlier described as parameters being considered random, now to be called *random effects*. These are attributable to a (usually) infinite set of levels of a factor, of which only a random sample are deemed to occur in the data. For example, four loaves of bread are taken from each of six batches of bread baked at three different temperatures. Since there is definite interest in the particular baking temperatures used, the statistical concern is to estimate those temperature effects; they are fixed effects. No assumption is made that the temperature effects are random. Indeed, even if the temperatures themselves were chosen at random, it would not be sensible to assume that the temperature *effects* were random. This is because temperature is defined on a continuum and, for example, the effect of a temperature of 450.11° is almost always likely to be a very similar to the effect of a 405.12° temperature. This nullifies the idea of temperature effects being random.

In contrast, batches are not defined on a continuum. They are real objects, just as are people, or cows, or clinics and so, depending on the circumstance, it can be perfectly reasonable to think of their effects as being random. Moreover, we can do this even if the objects themselves have not been chosen as a random sample—which, indeed, they seldom are. So we assume that batch effects are random, and then interest in them lies in estimating the variance of those effects. Thus data from this experiment would be considered as having two sources of random variation: batch variance and, as usual, error variance. These two variances are known as *variance components*: for linear models their sum is the variance of the variable being observed.

Models in which the only effects are fixed effects are called *fixed effects models*, or sometimes just *fixed models*. And those having (apart from a single, general mean common to all observations) only random

Figure 1.1: Examples of Interaction and No Interaction.



effects are called *random effects models* or, more simply, *random models.* Further examples and properties of fixed effects and of random effects are given in Sections 1.3 and 1.4.

When there are several factors, the effect of the combination of two or more of them is called an *interaction effect.* In contrast, the effect of a single factor is called a *main effect.* The concept of interaction is as follows: If the change in the mean of the response variable between two levels of factor A is the same for different levels of factor B, we say that there is no interaction; but if that change is different for different levels of B, we say that there is an interaction. Figure 1.1 illustrates the two cases.

Details of the application of analysis of variance techniques to LMs and LMMs depend in many cases on whether every cell (as defined above) of the data has the same number of observations. If that is so the data are said to be *balanced data*; if not, the data are *unbalanced data*, in which case they are either *all-cells-filled data* (where every cell contains data) or *some-cells-empty data* (where some cells have no data). Implications of these descriptions are discussed at length in Searle et al. (1992, Sec. 1.2).

1.3 FIXED EFFECTS MODELS

Fixed effects and random effects have been specified and described in general terms. We now illustrate the nature of these effects using

real-life examples and emphasizing the particular properties of random effects.

a. Example 1: Placebo and a drug

A clinical trial of treating epileptics with the drug Progabide is described in Diggle et al. (1994). We ignore the baseline period of the experiment, and consider a response which is the number of seizures after patients were randomly allocated to either the placebo or the drug. If y_{ij} is the number of seizures experienced by patient j receiving treatment i ($i = 1$ for placebo and $i = 2$ for Progabide), a possible model for y_{ij} could be based upon starting from the expected value

$$\mathrm{E}[y_{ij}] = \mu_i,$$

where μ_i is the mean number of seizures expected from someone receiving treatment i. If we wanted to write $\mu_i = \mu + \alpha_i$ we would then have

$$\mathrm{E}[y_{ij}] = \mu_i = \mu + \alpha_i \tag{1.1}$$

where μ is a general mean and α_i is the effect on the number of seizures due to treatment i.

In this modeling of the expected value of y_{ij}, each μ_i (or μ and each α_i) is considered as a fixed unknown constant, the magnitudes of which we wish, in some general sense, to estimate; that is, we want to estimate μ_1, μ_2, and $\mu_1 - \mu_2$. And having estimated that difference we would want to test if it is less than zero (i.e., to test if the drug is reducing the number of seizures). In doing this the μ_is (or the α_is) correspond to the two different treatments being used. They are the only two being used, and in using them there is no thought for any other treatments. This is the concept of fixed effects. We consider just the treatments being used and no others, and so the effects are called *fixed effects*.

The manner in which data are obtained always affects inferences that can be drawn from them. We therefore describe a sampling process pertinent to this fixed effects model. The data are envisaged as being one possible set of data involving these same two treatments that could be derived from repetitions of the clinical trial, repetitions for which a different sample of people receiving each treatment would be used. This would lead on each occasion to a set of data that would be two random samples, one from a population of possible data having mean μ_1, and another from a population having μ_2.

The all-important feature of fixed effects is that they are deemed to be constants representing the effects on the response variable y of the various levels of the factor concerned, in this case the two treatments, placebo and drug. These treatments are the levels of the factor of particular interest, chosen because of interest in those treatments in the trial. But they could just as well be different fertilizers applied to a corn crop, different forage crops grown in the same region, different machines used in a manufacturing process, different drugs given for the same illness, and so on. The possibilities are legion, as are the varieties of models and their complexities. We offer two more brief examples.

b. Example 2: Comprehension of humor

A recent study (Churchill, 1995) of the comprehension of humor ("Did you get it?") involved showing three types of cartoons (visual only, linguistic only, and visual-linguistic combined) to two groups of adolescents (normal and learning disabled). Motivated by this study, suppose the adolescents record scores of 1 through 9, with 9 representing extremely funny and 1 representing not funny at all. Then with $\bar{y}_{ij\cdot}$ being the mean score from showing cartoon type i to people in group j, a suitable start for a model for $\bar{y}_{ij\cdot}$ could be

$$\mathrm{E}[\bar{y}_{ij\cdot}] = \mu_{ij} = \mu + \alpha_i + \beta_j, \tag{1.2}$$

where μ is a general mean, α_i is the effect on comprehension due to cartoon type i (= 1, 2 or 3) and β_j is the effect due to respondents being in adolescent group j (= 1 or 2). Because each of the same three cartoon types is shown to each of the two adolescent groups, this is an example of two crossed factors, cartoon type and adolescent group. Furthermore, since the three cartoon types and the two groups of people have been chosen as the only types and groups being considered, the α_is and β_js are fixed effects corresponding to the three types and two groups. They are the specific features of interest in this study, and under no circumstances can they be deemed to have been chosen randomly from a larger array of types and groups. Thus the α_is and β_js are fixed effects. This is just a simple extension of Example 1 which has one factor with two levels. In Example 2 there are two factors: one, type of cartoon, with three levels, and another, group of people, with two levels. After estimating effects for these levels we might want to make inferences about the extent to which the visual – linguistic cartoons were comprehended differently from the average of

the visual and linguistic ones; and we would also want to test if the learning-disabled adolescents differed in their cartoon comprehension from the non-disabled adolescents. (In actual practice this study involved 8 different cartoons within each type, and several people in each group; these two factors, cartoon-within-type of cartoon, and people-within-group, had also to be taken into account.)

c. Example 3: Four dose levels of a drug

Suppose we have a clinical trial in which a drug (e.g., Progabide of Example 1) is administered at four different dose levels. For y_{ij} being the datum for the jth person receiving dose i we could start with

$$\mathrm{E}[y_{ij}] = \mu_i = \mu + \alpha_i. \quad . \tag{1.3}$$

This is just like (1.1), only where $i = 1,\ 2,\ 3$ or 4, corresponding to the four dose levels. The μ_is (and α_is) are fixed effects because the four dose levels used in the clinical trial are the only dose levels being studied. They are the doses on which our attention is fixed. This is exactly like Example 1 which has only two fixed effects whereas here there are four, one for each dose level. And after collecting the data, interest will center on differences between dose levels in their effectiveness in reducing seizures.

No matter how many factors there are, if they are all fixed effects factors and the fixed effects are combined linearly as in (1.1), (1.2) and (1.3), the model is called a fixed model; or, more generally just a linear model (LM). And there can, of course, also be fixed effects in nonlinear models.

1.4 RANDOM EFFECTS MODELS

a. Example 4: Clinics

Suppose that the clinical trial of Example 3 was conducted at 20 different clinics in New York City. Consider just the patients receiving the dose level numbered 1. The model equation for y_{ij}, which represents the jth patient at the ith clinic, could then be

$$\mathrm{E}[y_{ij}] = \mu + a_i, \tag{1.4}$$

with $i = 1,\ 2,\ \ldots,\ 20$ for the 20 clinics. But now pause for a moment. It is not unreasonable to think of those clinics (as do Chakravorti and

Grizzle, 1975) as a random sample of clinics from some distribution of clinics, perhaps all the clinics in New York City.

Note that (1.4) is essentially the same algebraically as (1.3), save for having a_i in place of α_i. However, the underlying assumptions are different. In (1.3) each α_i is a fixed effect, the effect of dose level i on the number of seizures; and dose level i is a pre-decided treatment of interest. But in (1.4) each a_i is the effect on number of seizures of the observed patient having been in clinic i; and clinic i is just one clinic, the one from among the randomly chosen clinics that happened to be numbered i in the clinical trial. The clinics have been chosen randomly with the object of treating them as a representation of the population of all clinics in New York State, and inferences can and will be made about that population. This is a characteristic of random effects: they can be used as the basis for making inferences about populations from which they have come. Thus a_i is a *random effect.* As such, it is, indeed, a random variable, and the data will be useful for making an inference about the variance of those random variables; i.e., about the magnitude of the variation among clinics; and for predicting which clinic is likely to have the best reduction of seizures.

b. Notation

As indicated briefly in Section 1.1 we adopt a convention of μ for a general mean and, for purposes of distinction, Greek letters for fixed effects and Roman for random effects. Thus for fixed effects equations (1.1) and (1.3) have α_i, and (1.2) has α_i and β_j; but (1.4) has a_i for random effects.

– i. *Properties of random effects in LMMs*

With the a_is being treated as random variables, we must attribute probabilistic properties to them. There are two that are customarily employed; first, that all a_is are independently and identically distributed (i.i.d.); second, that they have zero mean, and then, that they all have the same variance, σ_a^2. We summarize this as

$$a_i \sim \text{i.i.d. } (0, \sigma_a^2) \quad \forall \quad i. \tag{1.5}$$

This means that

$$\mathrm{E}[a_i] \;=\; 0 \quad \forall \quad i, \tag{1.6}$$

$$\text{var}(a_i) = \text{E}[(a_i - \text{E}[a_i])^2] = \text{E}[a_i^2] = \sigma_a^2 \quad (1.7)$$

and

$$\text{cov}(a_i, a_k) = 0 \text{ for } i \neq k. \quad (1.8)$$

There are, of course, properties other than these that could be used such as the covariances of (1.8) being non-zero.

– ii. *The notation of mathematical statistics*

A second outcome of treating the a_is as random variables is that we must consider $\text{E}[y_{ij}] = \mu + a_i$ of (1.4) with more forethought, because it is really a mean calculated conditional on the value of a_i. To describe this situation carefully, we revert for a moment to the standard mathematical statistics notation which uses capital letters for random variables and lowercase letters for realized values. Since a random variable appears on the right-hand side of (1.4) the more precise way of writing (1.4) is

$$\text{E}[Y_{ij}|A_i = a_i] = \mu + a_i \quad (1.9)$$

from which, when the realized value of A_i is not known, we write

$$\text{E}[Y_{ij}|A_i] = \mu + A_i. \quad (1.10)$$

In fact, of course, (1.10) is the basic result from which (1.9) is the special case. And from (1.10) we get the standard result

$$\text{E}[Y_{ij}] = \text{E}_\text{A}\left[\text{E}[Y_{ij}|A_i]\right] = \text{E}_\text{A}[\mu + A_i] = \mu + \text{E}_\text{A}(A_i) = \mu \quad (1.11)$$

because we are taking $\text{E}_\text{A}[A_i]$, which is the expectation of A_i over the distribution of A, as being zero.

Note that assuming $\text{E}_\text{A}[A_i] = 0$ involves no loss of generality to the results (1.9), (1.10) or (1.11). This is because if instead of $\text{E}_\text{A}[A_i] = 0$ we took $\text{E}_\text{A}[A_i] = \tau$, say, then (1.11) would become

$$\text{E}[Y_{ij}] = \mu + \text{E}_\text{A}[A_i] = \mu + \tau = \mu', \text{ say.}$$

And then (1.10) would be

$$\text{E}[Y_{ij}|A_i] = \mu + A_i = \mu + \tau + A_i - \tau = \mu' + A_i'$$

for $A_i' = A_i - \tau$ with $\text{E}_\text{A}[A_i'] = \tau - \tau = 0$; and so (1.10) is effectively unaffected. Similarly, (1.9) would become

$$\text{E}[Y_{ij}|A_i = a_i] = \mu + a_i = \mu + \tau + a_i - \tau = \mu' + a_i'$$

which is (1.9) with μ' and a_i' in place of μ and a_i, respectively, and the form of (1.10) and (1.11) is retained.

Finally, if there is interest in the particular level of the random effect (e.g., in knowing how the ith clinic differs from the average) then we will be interested in predicting the realized value a_i. On the other hand, if interest lies in the population from which A_i is drawn, we will be interested in $\text{var}(A_i) = \sigma_a^2$.

From now on, for notational convenience, we judiciously ignore the distinction between a random variable and its realized value and let a_i do double duty for both; likewise for y_{ij}.

– iii. *Variance of y*

Having defined $\text{var}(a_i) = \sigma_a^2$, we now consider $\text{var}(y_{ij})$ by first considering the variation that remains in the data after accounting for the random factors. If the data were normally distributed we would typically define a residual error $y_{ij} - \text{E}[y_{ij}|a_i]$ and to it attribute a normal distribution. Equivalently we could simply assert that

$$y_{ij}|a_i \sim \text{ i.i.d. } N(\text{E}[y_{ij}|a_i],\ \sigma^2). \tag{1.12}$$

Either approach works perfectly well when assuming normality. But for non-normal cases (1.12) is more sensible.

In Example 1 each y_{ij} is a count of the number of seizures. It is therefore quite natural to think that y_{ij} should follow a Poisson model, and assert that

$$y_{ij}|a_i \sim \text{ i.i.d. Poisson}(\text{E}[y_{ij}|a_i]). \tag{1.13}$$

In doing this the "residual" variation is encompassed in the conditional distribution which in (1.13) is taken to be Poisson. If we tried to attribute a distribution to the residual $y_{ij} - \text{E}[y_{ij}|a_i]$ it would be much less natural since, e.g., $y_{ij} - \text{E}[y_{ij}|a_i]$ may not even take on integer values. Thus we would have an awkward-to-deal-with distribution.

– iv. *Variance and conditional expected values*

To obtain $\text{var}(y)$ we will often use the formula which relates variance to conditional expected values in order to partition variability:

$$\text{var}(y) = \text{var}(\text{E}[y|u]) + \text{E}[\text{var}(y|u)].$$

For the case of the homoscedastic linear model (1.4) through (1.7), this gives the usual *components of variance* breakdown:

$$\begin{aligned} \sigma_y^2 = \text{var}(y_{ij}) &= \text{var}(\text{E}[y_{ij}|a_i]) + \text{E}[\text{var}(y_{ij}|a_i)] \qquad (1.14) \\ &= \text{var}(\mu + a_i) + \text{E}[\sigma^2] \\ &= \sigma_a^2 + \sigma^2. \qquad (1.15) \end{aligned}$$

A similar formula holds for covariances:

$$\text{cov}(y, w) = \text{cov}_u(\text{E}[y|u], \text{E}[w|u]) + \text{E}_u[\text{cov}(y, w|u)], \qquad (1.16)$$

a derivation of which is to be found in Searle et al. (1992, p. 462). Applying this to the homoscedastic linear model gives

$$\begin{aligned} \text{cov}(y_{ij},\ y_{ij'}) &= \text{cov}(\mu + a_i,\ \mu + a_i) + \text{E}[0] \\ &= \sigma_a^2. \qquad (1.17) \end{aligned}$$

Thus σ_a^2 is the intra-class covariance, i.e., the covariance between every pair of observations in the same class; and $\sigma_a^2/(\sigma_a^2+\sigma^2)$ is the intra-class correlation coefficient.

c. Example 5: Ball bearings and calipers

Consider the problem of manufacturing ball bearings to a specified diameter that must be achieved with a high degree of accuracy. Suppose that each of 100 ball bearings is measured with each of 20 micrometer calipers, all of the same brand. Then a suitable model equation for y_{ij}, the diameter of the ith ball bearing measured with the jth caliper, could be

$$\text{E}[y_{ij}] = \mu + a_i + b_j. \qquad (1.18)$$

This is another example of two crossed factors as in Example 2, with the same model equation as in (1.2) except that the symbols a_i and b_j are used rather than α_i and β_j. But it is the equation of a different model because a_i and b_j are random effects corresponding, respectively, to the 100 ball bearings being considered as a random sample from the production line, and to the 20 calipers that are being considered as a random sample of calipers from some population of available calipers. Hence in (1.18) each a_i and b_j is treated in the same manner as a_i is treated in Example 4, with the additional property of stochastic independence of the a_is and b_js; thus

$$\text{cov}(a_i, b_j) = 0. \qquad (1.19)$$

In this case, inferences of interest will be those concerning the magnitudes of the variance among ball bearings and the variance among calipers.

1.5 LINEAR MIXED MODELS (LMMs)

a. Example 6: Medications and clinics

Another example of two crossed factors is to suppose that all four dose levels of Example 3 were used in all 20 clinics of Example 4, such that in each clinic each patient was randomly assigned to one of the dose levels. If y_{ijk} is the datum for patient k on dose level j in clinic i, then a suitable model equation for $\mathrm{E}[y_{ijk}]$ would be

$$\mathrm{E}[y_{ijk}] = \mu + a_i + \beta_j + c_{ij} \tag{1.20}$$

where a_i, β_j and c_{ij} are effects due to clinic i, dose j and clinic-by-dose interaction, respectively. Since, as before, the doses are the only doses considered, β_j is a fixed effect. But the clinics that have been used were chosen randomly, and so a_i is a random effect. Then, because c_{ij} is an interaction between a fixed effect and a random effect, it is a random effect, too. Thus the model equation (1.20) has a mixture of both fixed effects, the β_js, and random effects, the a_is and c_{ij}s. It is thus called a *mixed model.* It incorporates problems relating to the estimation of both fixed effects and variance components. Inferences of interest will be those concerning the effectiveness of the different doses and the variability (variance) among the clinics.

In application to real-life situations, mixed models have broader use than random models, because so often it is appropriate (by the manner in which data have been collected) to have both fixed effects and random effects in the same model. Indeed, every model that contains a μ is a mixed model, because it also contains unexplained variation, and so automatically has a mixture of fixed effects and random elements. In practice, however, the name *mixed model* is usually reserved for any model having both fixed effects (other than μ) and random effects, as well as the customary unexplained variation.

b. Example 7: Drying methods and fabrics

Devore and Peck (1993) report on a study for assessing the smoothness of washed fabric after drying. Each of nine different fabrics were

subjected to five methods of drying (line drying; line drying after brief machine tumbling; line drying after tumbling with softener; line drying with air movement; and machine drying). Clearly, method of drying is a fixed effects factor. But how about fabric? If those nine fabrics were specifically chosen as being the only fabrics under consideration, then fabric is a fixed factor. But if the nine fabrics just happened to be the fabrics occurring in a family wash, then it might be reasonable to think of those fabrics as just being a random sample of fabrics from some population of fabrics—and fabric would be a random effect.

Notice that it is what we think is the nature of a factor and of its levels occurring in the data that determines whether a factor is to be called fixed or random. This is discussed further in Section 1.6. As in many mixed models, inference is directed to differences between fixed effects and to the magnitude of the variance among random effects.

c. Example 8: Potomac River Fever

A study of Potomac River Fever in horses was conducted by sampling horses from social groups of horses within 522 farms in New York State (Atwill et al., 1996). The social groups were defined by whether the animals tended to be kept together (i.e., in the same barn or pasture). Breed, gender, and types of animal care (e.g., stall cleaning and frequency of spraying flies) are some of the fixed effects factors that might need to be reckoned with. But farm, and equine social group nested within farm, are clearly random factors.

d. Regression models

Customary regression analysis is based on model equations (for three predictor variables, for example) of the form

$$\mathrm{E}[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

where the βs are considered to be fixed constants which get estimated from data on y and the xs. But sometimes it is appropriate to think of some βs as being random. When this is so the model is often called a *random coefficients model.*

e. Longitudinal data

A common use of mixed models is in the analysis of longitudinal data, which are defined as data collected on each subject (broadly inter-

preted) on two or more occasions. Methods of analysis have typically been developed for the situation where the number of occasions is small compared to the number of subjects. Experiments with longitudinal data are widely used for at least three reasons: (1) to increase sensitivity by making within-subject comparisons, (2) to study changes through time, and (3) to use subjects efficiently once they are enrolled in a study.

The decision as to whether a factor should be fixed or random in a longitudinal study is often made on the basis of which effects vary with subjects. That is, subjects are regarded as a random sample of a larger population of subjects and hence any effects that are not constant for all subjects are regarded as random.

For example, suppose we are testing a blood pressure drug at each of two doses and a control dose (dose = 0) for each subject in our study. Individuals clearly have different average blood pressures, so our model must have a separate intercept for each subject. Similarly, the response of each subject to increasing dosage of the drug might vary from subject to subject, so we would model the slope for dose separately for each subject. To complete our model, we might also assume that blood pressure changes gradually with a subject's age, measured at the beginning of the study. If we let y_{ij} denote the blood pressure measurement taken on the ith subject (of age x_i) on occasion j at dose d_{ij}, we could then model $\mathrm{E}[y_{ij}]$ as

$$\mathrm{E}[y_{ij}] = a_i + b_i d_{ij} + \gamma x_i. \tag{1.21}$$

Since the a_i and b_i are specific to the ith subject, they would be declared random factors. Since γ is the same for all subjects, it is declared fixed.

If we are interested in the overall population response to the drug we can separate overall terms from the terms specific to each subject. To do so we rewrite (1.21) as

$$\mathrm{E}[y_{ij}] = (\alpha + a_i') + (\beta + b_i')d_{ij} + \gamma x_i, \tag{1.22}$$

where $a_i' = a_i - \alpha$ and $b_i' = b_i - \beta$, with α and β being averages over the population of subjects and are thus fixed effects. On the other hand, a_i' and b_i' are subject-specific deviations from these overall averages and so are treated as random effects with means zero and variances σ_a^2 and σ_b^2. Chapter 7 is devoted to the analysis of longitudinal data.

f. Model equations

Notice in the preceding examples that equations (1.1), (1.2), (1.3), (1.4), (1.18) and (1.20) are all described as *model equations*. Many writers refer to such equations as models; but this is not correct, because description of a model demands not only a model equation but also explanation of the nature of the terms in such an equation. For example, (1.1) and (1.4) are essentially the same model *equation*, $\mathrm{E}[y_{ij}] = \mu + \alpha_i$ and $\mathrm{E}[y_{ij}] = \mu + a_i$, but the models are not the same. The α_i is a fixed effect but the a_i is a random effect; and this difference, despite the sameness of the model equations, means that the models are different and the analysis of data in the two cases is accordingly different. Moreover, models being different but with the same right-hand sides of their model equation applies not just to whether effects are fixed or random, but can also apply to models that are even more different. For example, $\mu + \alpha_i$, which occurs so often in analysis of variance style models can also occur on the right-hand side of a model equation in a binomial model.

1.6 FIXED OR RANDOM?

Equation (1.2) for modeling cartoon types and groups of people (normal and disabled) is indistinguishable from (1.18) for modeling ball bearings and calipers. But the complete models in these cases are different because of the interpretation attributed to the effects: in the one case, fixed, and in the other, random. In these and the other examples most of the effects are clearly fixed or random; thus drugs and methods of drying are fixed effects, whereas clinics and farms are random effects. But such clear answers to the question "fixed or random?" are not necessarily the norm. Consider the following example.

a. Example 9: Clinic effects

A multicenter clinical trial is designed to judge the effectiveness of a new surgical procedure. If this procedure will eventually become a widespread procedure practiced at a number of clinics, then we would like to select a representative collection of clinics in which to test the procedure and we would then regard the clinics as a random effect.

However, suppose we change the situation slightly. Now assume that the surgical procedure is highly specialized and will be performed mainly at a very few referral hospitals. Also assume that all of those

referral hospitals are enrolled in the trial. In such a case we cannot regard the selected clinics as a sample from a larger group of clinics and we will be satisfied with making inferences only to the clinics in the study. We would therefore treat clinic as a fixed effect.

Is is clear that we could envision situations that are intermediate between the "treat clinics as random" scenario and the "treat clinics as fixed" scenario and making the decision between fixed and random would be very difficult. Thus it is that the situation to which a model applies is the deciding factor in determining whether effects are to be considered as fixed or random.

b. Making a decision

Sometimes, then, the decision as to whether certain effects are fixed or random is not immediately obvious. Take the case of year effects, for example, in studying wheat yields: are the effects of years on yield to be considered fixed or random? The years themselves are unlikely to be random, for they will probably be a group of consecutive years over which data have been gathered or experiments run. But the effects on yield may reasonably be considered random, subject, perhaps, to correlation between yields in successive years.

In endeavoring to decide whether a set of effects is fixed or random, the context of the data, the manner in which they were gathered and the environment from which they came are the determining factors. In considering these points the important question is: are the levels of the factor going to be considered a random sample from a population of values which have a distribution? If "yes" then the effects are to be considered as random effects; if "no" then, in contrast to randomness, we think of the effects as fixed constants and so the effects are considered as fixed effects. Thus when inferences will be made about a distribution of effects from which those in the data are considered to be a random sample, the effects are considered as random; and when inferences are going to be confined to the effects in the model, the effects are considered fixed.

Another way of putting it is to ask the questions: "Do the levels of a factor come from a probability distribution?" and "Is there enough information about a factor to decide that the levels of it in the data are like a random sample?" Negative answers to these questions mean that one treats the factor as a fixed effects factor and estimates the effects of the levels; and treating the factor as fixed indicates a more

limited scope of inference. On the other hand, affirmative answers mean treating the factor as a random effects factor and estimating the variance component due to that factor. In that case, when there is also interest in the realized values of those random effects that occur in the data, then one can use a prediction procedure for those values.

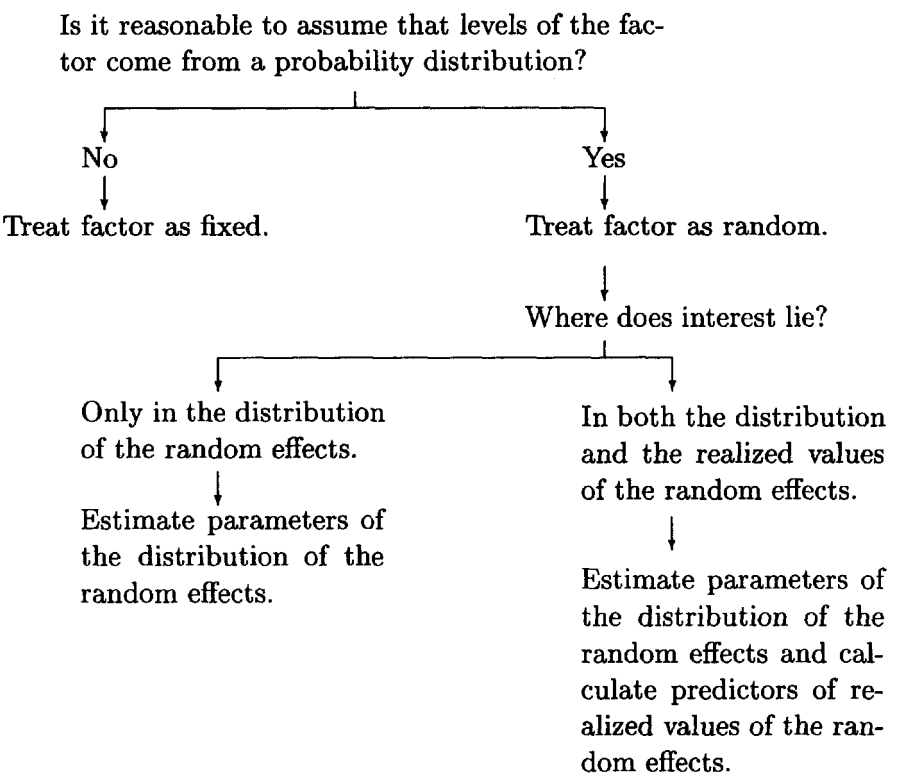
It is to be emphasized that the assumption of randomness does not carry with it the assumption of normality. Often this assumption *is* made for random effects, but it is a separate assumption made subsequent to that of assuming effects are random. Although many estimation procedures for variance components do not require normality, if distributional properties of the resulting estimators are to be investigated then normality of the random effects is often assumed.

For any factor, the decision tree shown in Figure 1.2 has to be followed in order to decide whether the factor is to be considered as fixed or random. Consider using Figure 1.2 for Example 7 of Section 1.5b where the two factors of interest are five methods of drying and nine different fabrics. To the question atop Figure 1.2, for methods of drying the answer is clearly "no." The five methods cannot be thought of as coming from a probability distribution. But for the other factor, fabrics, it might seem quite reasonable to answer that question with "Yes, the nine fabrics used in the experiment can be thought of as coming from a probability distribution insofar as their propensity for drying is concerned." Thus methods of drying would be treated as fixed effects and fabrics as random. On the other hand, suppose the nine fabrics were nine mixtures of Orlon and cotton being manufactured by one company for a shirt maker. Then those nine fabrics would be the only fabrics of interest to their manufacturer – and in no way would they be thought of as coming from a distribution of fabrics. So they would be treated as fixed.

1.7 INFERENCE

The essence of statistical analysis has three parts: collection of data, summarizing data, and making inferences. Data get considered as samples from populations, and from data one makes inferences about populations. These inferences might well be termed conclusions supported by probability statements. In contrast to conclusions derived by deductions as being rock-solid and immutable, one might say that conclusions drawn from inference are conclusions diluted by probability statements. In any case, that is where the use of statistics usually

Figure 1.2: Decision tree for deciding fixed versus random



leads us. We therefore briefly summarize goals we aim for when using inference, and some of the methods involved in doing that. Goals in the use of statistics are principally of three kinds: estimating (including confidence intervals), testing and predicting.

Inference in traditional linear models is based largely on least squares estimation for fixed effects, on analysis of variance sums of squares for estimating variances of random effects, and on normality assumptions for making tests of hypotheses, and for calculating confidence intervals, best predictors and prediction intervals. These procedures will always have their uses for LMs and LMMs. But for GLMs and GLMMs, where distribution assumptions different from normality are so often invoked, use is made of a broader range of procedures than those traditional to linear models. We list briefly here some of the inferential methodologies.

a. Estimation

In fixed and in mixed models we want to estimate both fixed effects and linear functions of them, particularly differences between the levels of any given factor. For instance, in Example 1 there would be interest in estimating the difference in mean seizure numbers between patients who received the drug Progabide and those who did not. And in Example 2 we would want to estimate the difference in humor comprehension between normal and learning-disabled people; and also the difference in the average of the visual-only plus verbal-only types of cartoon from the visual-and-verbal-combined type of cartoon. Similarly, in Example 7 we would be interested in estimating differences in fabric smoothness among the various methods of drying.

When random effects are part of a model we often want to estimate variances of the effects within a factor—and of covariances too, to the extent that they are part of the specification of the random effects. Thus in Example 4 estimating the variance of the clinic effects would be important because it would be an estimate of the variability within the entire population of clinics, not just within the 20 clinics used in the study. And in mixed models, as well as estimating fixed effects we also want to estimate variances of random effects just as in random models. Thus in Example 6 we would estimate differences in seizure numbers as between the various Progabide dose levels; and also estimate the variance among clinics.

– i. *Maximum likelihood (ML)*

The primary method of estimation we consider throughout this book is *maximum likelihood* (ML). If $\mathbf{y}$ is the data vector and $\boldsymbol{\theta}$ the vector of parameters in the distribution function of $\mathbf{y}$, we can represent that function as $f(\mathbf{y}|\boldsymbol{\theta})$, meaning that for some given value of $\boldsymbol{\theta}$ it is the density function of $\mathbf{y}$. But for $\boldsymbol{\theta}$ being simply the representation of any one of the possible values of $\boldsymbol{\theta}$ we could also rewrite the density function as $L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})$, which is called the *likelihood*. This is a function of $\boldsymbol{\theta}$ and ML is the process of finding that value of $\boldsymbol{\theta}$ which maximizes $L(\boldsymbol{\theta}|\mathbf{y})$. For mathematically tractable density functions this process can be quite straightforward, yielding a single, algebraic expression for the maximizing $\boldsymbol{\theta}$ as a function of $\mathbf{y}$. But for difficult functions it can demand iterative numerical methods and may not always yield a single value for the maximizing $\boldsymbol{\theta}$. Naturally, this presents problems when

applying ML to real data.

– ii. *Restricted maximum likelihood (REML)*

A method related to ML is *restricted* (or *residual*) *maximum likelihood* (REML), which involves the idea of applying ML to linear functions of $\mathbf{y}$, say $\mathbf{K}'\mathbf{y}$, for which $\mathbf{K}'$ is specifically designed so that $\mathbf{K}'\mathbf{y}$ contains none of the fixed effects which are part of the model for $\mathbf{y}$. So in ML, replace $\mathbf{y}$ with $\mathbf{K}'\mathbf{y}$ and one has REML. Historically REML was derived only for the case of linear mixed models (Patterson and Thompson, 1971) but has been generalized to nonlinear models (e.g., Schall, 1991).

Two valuable consequences of using REML are first, that variance components are estimated without being affected by the fixed effects. This means that the variance estimates are invariant to the values of the fixed effects. Second, in estimating variance components with REML, degrees of freedom for the fixed effects are taken into account implicitly, whereas with ML they are not. The simplest example of this is in estimating σ^2 from normally distributed data y_i, which we denote as $y_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$. With $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, the REML estimator of σ^2 is $S_{yy}/(n-1)$ whereas the ML estimator is S_{yy}/n. Further examples appear in Sections 2.1 and 3.1, and a full discussion of ML and REML is given in Chapter 6. Derivation of the preceding results involving S_{yy} is covered in Exercise E 1.6.

Beyond LMMs, the ideas of REML can be generalized directly to non-normal models where, in limited cases, a linear function of $\mathbf{y}$ can be constructed to contain none of the fixed effects. However, for non-normal and nonlinear models, other, alternative "definitions" of REML have been put forth. Two examples are the solutions of equations that equate quadratic forms of predicted random effects with their expected values, and the maximization of a likelihood after "integrating out" the fixed effects.

Whichever definition is adopted, note that REML does nothing about estimating fixed effects. This is because all REML methods are designed to be free of the fixed effects portion of a model.

– iii. *Solutions and estimators*

Estimators such as ML or REML are found by maximizing a function of the parameters (the likelihood or restricted likelihood) within the

bounds of the parameter space. In general this problem, the maximization of a nonlinear function within a constrained region, is quite difficult. For many models considered in this book, we will be able to do little more than point the reader in the direction of numerical methods for finding the numerical *estimates* for a particular data set. For other models we can be more explicit.

In cases where explicit, closed-form solutions exist for the maximizing values, an oft-successful method for finding those solutions is to differentiate the likelihood or restricted likelihood and set the derivatives equal to zero. From the resulting equations, solutions for the parameter symbols might well be thought of as ML or REML estimators, in which case they would be denoted by a "hat" or tilde over the parameter symbol. However, this is not a fail-safe method because, in some cases, solutions to these estimating equations may not be in the parameter space. For example, in some cases of ML or REML a solution for an estimated variance is such that it is possible for it to be negative; that is, it is possible for data to be such that the solution is a negative value (e.g., Section 2.2). Since, under ML and REML, negative estimates of positive parameters are not acceptable, we will often denote solutions with a dot above the parameter (e.g., $\dot{\sigma}_a^2$). We then proceed to adjust them (in accord with established procedures) to yield the ML estimators which will be denoted in the usual way with a "hat" above the parameter symbol. Thus for ν denoting a variance we can have

$$\begin{aligned} \nu &= \text{parameter} \\ \dot{\nu} &= \text{solution} \\ \hat{\nu} &= \text{estimator.} \end{aligned} \tag{1.23}$$

– iv. *Bayes theorem*

Even for fixed effects one school of thought in statistics is to assume that they are random variables with a distribution $\Pi(\boldsymbol{\theta})$. This is called the *prior distribution* of $\boldsymbol{\theta}$. Then, because

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta}) = \Pi(\boldsymbol{\theta}|\mathbf{y})f(\mathbf{y}), \tag{1.24}$$

we have

$$\Pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{1.25}$$

This is called the *posterior density* of $\boldsymbol{\theta}$; and the mean of $\boldsymbol{\theta}|\mathbf{y}$ derived from this density is an often-used estimator of $\boldsymbol{\theta}$—a Bayes estimator.

Suppose that $\Pi(\boldsymbol{\theta})$ and $\Pi(\boldsymbol{\theta}|\mathbf{y})$ themselves involve parameters $\boldsymbol{\varphi}$ so that

$$f(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|\boldsymbol{\varphi})}{\Pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\varphi})} \tag{1.26}$$

from (1.25). There can be cases where $\boldsymbol{\varphi}$ can be estimated as a function of y from (1.26) using, for example, marginal maximum likelihood. Then, on using those estimates in (1.25), the $\mathrm{E}[\boldsymbol{\theta}|\mathbf{y}]$ derived from that adaptation of $f(\boldsymbol{\theta}|\mathbf{y})$ is known as an empirical Bayes estimate of $\boldsymbol{\theta}$.

– v. *Quasi–likelihood estimation*

In many problems of statistical estimation we know some detail of the distribution governing the data, but may be unwilling to specify it exactly. This precludes the use of maximum likelihood, which requires exact specification of the distribution in order to construct the likelihood. The idea of quasi–likelihood, developed by Wedderburn (1974), addresses this concern. This is a method of estimation that requires only a model for the mean of the data and the relationship between the mean and the variance, yet in many cases retains full or nearly full efficiency compared to maximum likelihood. Since the input is minimal, the method is robust to misspecification of finer details of the model. Section 5.6 contains further details.

– vi. *Generalized estimating equations*

To capture some of the beneficial aspects of quasi–likelihood estimation in the context of correlated data models, Zeger and Liang (1986) and Liang and Zeger (1986) developed *generalized estimating equations* methods (GEEs). These methods are robust in the presence of misspecification of the variance-covariance structure of data. Estimates using GEEs are often easier to compute than maximum likelihood estimates. GEEs are discussed in Section 8.6a.

b. Testing

Insofar as testing is concerned one's usual interests are to test hypotheses about the parameters (and/or functions of them) which have been estimated as described above. With fixed effects, we test hypotheses

of the form that differences between levels of a factor are zero, or occasionally that they equal some pre-decided constant. And for random effects a useful hypothesis is that a variance component is zero—or, occasionally, that it equals some pre-decided value.

Ancillary to all these cases we often also want to use parameter estimates to establish confidence intervals for parameters, or for combinations of them.

– i. *Likelihood ratio test (LRT)*

Traditional analysis of variance methodology (under normality assumptions) leads to hypothesis tests involving F–statistics which are ratios of mean squares. These statistics can also be shown to be an outcome of the *likelihood ratio test* (LRT), first propounded by Neyman and Pearson (1928). This in its general form can be applied much more broadly than to traditional ANOVA, and is therefore useful for GLMs and GLMMs. That general form can be described as follows. Let $\hat{\boldsymbol{\theta}}$ be the maximizing $\boldsymbol{\theta}$ over the complete range of values of each element of $\boldsymbol{\theta}$. Similarly, let $\hat{\boldsymbol{\theta}}_0$ be the maximizing $\boldsymbol{\theta}$, limited (restricted or defined) by some hypothesis H pertaining to some elements of $\boldsymbol{\theta}$; and let $L(\hat{\boldsymbol{\theta}}_0)$ be the value of the likelihood using $\hat{\boldsymbol{\theta}}_0$ for $\boldsymbol{\theta}$. Then the likelihood ratio is $L(\hat{\boldsymbol{\theta}}_0)/L(\hat{\boldsymbol{\theta}})$; and it leads to a test statistic for the hypothesis H.

– ii. *Wald's procedure*

Another very general procedure for developing a hypothesis test, known as *Wald's test* (Wald, 1941), is that if $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ and $\mathbf{I}(\boldsymbol{\theta})$ is the information matrix for $\hat{\boldsymbol{\theta}}$, then $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)'[\mathbf{I}(\boldsymbol{\theta}_*)]^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)$ is a test statistic for the hypothesis $H\colon \boldsymbol{\theta} = \boldsymbol{\theta}_*$; and it has, under some conditions, approximately a χ^2_p distribution (p being the order of $\boldsymbol{\theta}$). For LMs this is exactly χ^2_p, or gets modified to exactly $\mathcal{F}$ when estimates are used in $\mathbf{I}(\boldsymbol{\theta})$. And for $p = 1$, the signed square root of this quadratic form in $\hat{\boldsymbol{\theta}}$ has approximately a normal distribution with zero mean and unit variance.

c. Prediction

Finally, there is prediction. When dealing with a random effects factor the random effects occurring in the data are realizations of a random variable. But they are unobservable. Nevertheless, in many situations we would like to use the data to put some sort of numerical values, or

predicted values on those realizations. They may be useful for selecting superior realizations: for example, picking superior clinics in our clinic example. It turns out (see Chapter 9) that the best predictor (minimum mean squared error) is a conditional mean. Thus if $\mathbf{y}$ represents the data the best predictor (BP) of a_i is $\mathrm{BP}(a_i) = \mathrm{E}[a_i|\mathbf{y}]$. And, similar to a confidence interval, we will at times also want to calculate a prediction interval for a_i using $\mathrm{BP}(a_i)$.

1.8 COMPUTER SOFTWARE

Nowadays there is a host of computer software packages that will compute some or many of the analysis methods described in the following chapters. Their existence is exceedingly important and useful to today's disciplines of statistics and data analysis. Nevertheless, this is not a book on software, its merits, its demerits, or the mechanics of using it. Software expands and (usually) improves so rapidly that whatever is written about it is somewhat outdated even before publication. So this is a statistics, not a software, user's manual. Very occasionally we mention SAS, which is a widely available package that can compute much of what we describe. In doing so, we give few or no details – and hope that our mentioning of only a single package is not construed as anything negative about the many other packages!

1.9 EXERCISES

E 1.1 Suppose a clothing manufacturer has collected data on the number of defective socks it makes. There are six subsidiary companies (factor C) that make knitted socks. At each company, there are five brands (B) of knitting machines with 20 machines of each brand at each company. All machines of all brands are used on the different types of yarn (Y) from which socks are made: cotton, wool, and nylon. At each company, data have been collected from just two machines (M) of each brand for operation by each of four locally resident workers, using each of the yarns. And on each occasion the number of defective socks in each of two replicate samples of 100 socks is recorded.

Which factors do you think should be treated as fixed and which as random? Give reasons for your decisions.

E 1.2 For $\mathrm{E}[y_{ij}|a_i] = \mu + a_i$ and $\mathrm{var}(y_{ij}|a_i) = \sigma^2$ use formulae (1.14) and (1.16) to derive $\mathrm{var}(y_{ij}) = \sigma_a^2 + \sigma^2$ and $\mathrm{cov}(y_{ij},\ y_{i\ell}) = \sigma_a^2$.

E 1.3 For $y_{ij} \sim \text{Poisson}(\lambda_i)$, repeat E 1.2; take $\lambda_i = e^{\mu+a_i}$, with $a_i \sim \mathcal{N}(0,\ \sigma_a^2)$. *Hint*: $\mathrm{E}[e^{a_i}]$ is the moment-generating function of a_i, i.e., $\mathrm{E}[e^{ta_i}]$, with t set equal to 1.

E 1.4 For $y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ with $i = 1,\ 2$, use (a) the LRT and (b) Wald's procedure to test $H_0 : \mu_1 = \mu_2$.

E 1.5 Derive (1.14), starting from

$$\begin{aligned} \mathrm{var}(y) &= \mathrm{E}\,(y - \mathrm{E}[y])^2 \\ &= \mathrm{E}\,(y - \mathrm{E}[y|u] + \mathrm{E}[y|u] - \mathrm{E}[y])^2 \\ &= \mathrm{E}_u\left[\mathrm{E}\,(y - \mathrm{E}[y|u] + \mathrm{E}[y|u] - \mathrm{E}[y])^2\,|u\right]. \end{aligned}$$

In doing so, explain why, in the expansion of the squared term, the cross-product is zero.

E 1.6 With $y_t \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2), t = 1, 2, \ldots, n$ and $\mathbf{y}' = [y_1, y_2, \ldots, y_n]$ we have $\mathbf{y} \sim \mathcal{N}(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n)$. Then the distribution function for $\mathbf{y}$ is

$$f(\mathbf{y}) = \exp\left[-\tfrac{1}{2}(\mathbf{y} - \mu\mathbf{1})'(\mathbf{y} - \mu\mathbf{1})/\sigma^2\right]/(2\pi\sigma^2)^{n/2}$$

and the log likelihood is

$$l_1 = -\tfrac{1}{2}(\mathbf{y} - \mu\mathbf{1})'(\mathbf{y} - \mu\mathbf{1})/\sigma^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi.$$

(a) By differentiating l_1 with respect to μ and σ^2 show that the ML estimators are

$$\hat{\mu} = \bar{y}_\cdot = \sum_{t=1}^{n} y_t/n$$

and

$$\tilde{\sigma}_1^2 = \sum_{t=1}^{n}(y_t - \bar{y})^2/n.$$

(b) For REML estimation of $\sigma^2, \mathbf{K}'$ for $\mathbf{K}'\mathbf{y}$ such that $\mathbf{K}'\mathbf{1} = \mathbf{0}$ can be taken as the first $n-1$ rows of $\mathbf{C}_n = \mathbf{I}_n - \mathbf{J}_n/n$. Thus

$$\mathbf{K}' = [\mathbf{I}_{n-1} \quad \mathbf{0}\mathbf{1}_{n-1}] - \frac{1}{n}\mathbf{J}_{(n-1)\times n}.$$

of order $(n-1) \times n$. Then

$$f(\mathbf{K}'\mathbf{y}) = \frac{\exp\left[-\frac{1}{2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}/\sigma^2\right]}{(2\pi\sigma^2)^{(n-1)/2} \mid \mathbf{K}'\mathbf{K} \mid^{1/2}}.$$

Thus the log likelihood of $\mathbf{K}'\mathbf{y}$ is

$$l_2 = -\frac{\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}}{2\sigma^2} - \frac{1}{2}\log \mid \mathbf{K}'\mathbf{K} \mid^{1/2} - \frac{n-1}{2}\log\sigma^2.$$

By differentiating this with respect to σ^2, show that the REML estimator of σ^2 is

$$\sigma^2_{\text{REML}} = \frac{\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}}{n-1}.$$

(c) Show that the REML estimator of σ^2 is $\sum_{t=1}^{n}(y_t - \bar{y})^2/(n-1)$, *Hint:* To do so, show that $\mathbf{K}'\mathbf{K} = \mathbf{I}_{n-1} - \mathbf{J}_{n-1}/n$ and $\mathbf{K}'\mathbf{y} = \left\{{}_c y_t - \bar{y}\right\}_{t=1}^{n-1}$.

Chapter 2

ONE-WAY CLASSIFICATIONS

We begin by describing fixed and random effects models for the one-way classification for both normally and Bernoulli (binary) distributed data. Not only do these constitute a convenient starting point for explaining many of the concepts described later in the book, they are also commonly employed in practice.

For example, in a modification of the comprehension of humor example (Section 1.3b) suppose that we have three cartoons, each of a different type (visual only, linguistic only, and visual–linguistic combined) and each is rated by separate people on a scale from 1 to 9, where 9 represents extremely funny and 1 represents not funny at all. We might consider the responses as approximately normally distributed and be interested in whether the mean rating is the same for the three cartoons. Alternatively or additionally, we might measure a yes/no response of whether the rater "got" the cartoon. The goal is the same: to compare the cartoons, but now, because of the binary nature of the data ("yes" or "no"), it is no longer valid to consider the data as approximately normally distributed. Statistical techniques acknowledging the binary nature of the data would be required.

If the inferential goal were to compare cartoon types, it would be insufficient to consider only a single cartoon of each type. A different modification of the humor example might have only visual cartoons, but would test, say, 15 different cartoons of this type. In this scenario, it is likely that we would regard cartoon as a random effect and the response could be either humor rating (from 1 to 9) or "got it?" (yes/no), or

both.

In this chapter we first consider normally distributed data with a fixed or random classification and then consider Bernoulli distributed data. In dealing with the 1-way classification throughout the chapter we let m be the number of classes and n_i the number of observations in the ith class. Then, with y_{ij} being the jth observation in the ith class, we have $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$.

2.1 NORMALITY AND FIXED EFFECTS

a. Model

Assuming normality and equal variances, a model for the responses y_{ij} is

$$\mathrm{E}[y_{ij}] = \mu_i \text{ with } y_{ij} \sim \text{indep. } \mathcal{N}(\mu_i, \sigma^2), \tag{2.1}$$

where "indep." means that the random variables are mutually independent and $\mathcal{N}(\mu_i, \sigma^2)$ indicates a normal distribution with mean μ_i and variance σ^2. An alternative but equivalent specification to (2.1) is the *overparameterized model*:

$$\mathrm{E}[y_{ij}] = \mu + \alpha_i \quad \text{with} \quad y_{ij} \sim \text{indep. } \mathcal{N}(\mu + \alpha_i, \sigma^2), \tag{2.2}$$

so-called because the mean of y is a function of more parameters than there are distinct values for the mean.

b. Estimation by ML

Derivation of maximum likelihood estimators of the parameters requires the likelihood. Using (2.1) it is

$$L = (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i\sum_j (y_{ij} - \mu_i)^2\right] \tag{2.3}$$

for $N \equiv \sum_{i=1}^m n_i$. Then

$$l = \log L = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i\sum_j (y_{ij} - \mu_i)^2.$$

The derivatives are

$$\begin{aligned} \frac{\partial l}{\partial \mu_k} &= -\frac{1}{2\sigma^2}\sum_j (y_{kj} - \mu_k)(-2) \qquad \text{and} \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i\sum_j (y_{ij} - \mu_i)^2. \end{aligned} \tag{2.4}$$

Setting these equal to zero gives solutions to the ML equations and, in this case, the ML estimators, which are denoted using "hats":

$$\begin{aligned} \hat{\mu}_i &= \bar{y}_{i\cdot} \\ \hat{\sigma}^2 &= \frac{1}{N}\sum_i\sum_j (y_{ij} - \hat{\mu}_i)^2 = \frac{1}{N}\sum_i\sum_j (y_{ij} - \bar{y}_{i\cdot})^2. \end{aligned} \tag{2.5}$$

These ML solutions are indeed ML estimators because they do not lie outside their corresponding parameter ranges (see Section 2.2b–iii), namely, $-\infty < \mu < \infty$ and $0 \leq \sigma^2$.

It is easily shown that $\hat{\mu}_i$ is unbiased; but $\hat{\sigma}^2$, the ML estimator of σ^2 in (2.5), is biased since

$$\mathrm{E}[\hat{\sigma}^2] = \mathrm{E}\left[\frac{1}{N}\sum_i (n_i - 1)s_i^2\right] = \frac{N-m}{N}\sigma^2 \neq \sigma^2, \tag{2.6}$$

where $s_i^2 = 1/(n_i - 1)\sum_j (y_{ij} - \bar{y}_{i\cdot})^2$ is the sample variance for the ith class. A common modification is to use an unbiased estimator:

$$s^2 = \frac{1}{N-m}\sum_i\sum_j (y_{ij} - \bar{y}_{i\cdot})^2.$$

This is also the REML (see Section 1.7a–ii) estimator. Derivation of this and some other estimators (see Sections 2.2b-vi and 3.2c) is left to Chapter 6 wherein a general equation is given for REML estimation of variances. And exercises in that chapter require using that general equation to obtain results for some simple cases such as those here.

The variances of these unbiased estimators are easily derived:

$$\mathrm{var}(\hat{\mu}_i) = \frac{\sigma^2}{n_i} \qquad \text{and} \qquad \mathrm{var}(s^2) = \frac{2\sigma^4}{N-m}. \tag{2.7}$$

These results follow easily from the standard result for the variance of a sample variance for a single sample from a normal distribution:

$$\mathrm{var}(s_i^2) = \frac{2\sigma^4}{n_i - 1}.$$

If we work with model (2.2) the derivatives of the log likelihood are

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_i\sum_j (y_{ij} - \mu - \alpha_i) \qquad \text{and} \\ \frac{\partial l}{\partial \alpha_k} &= \frac{1}{\sigma^2}\sum_j (y_{kj} - \mu - \alpha_k). \end{aligned} \tag{2.8}$$

Setting these equal to zero give the equations

$$\begin{aligned} \hat{\mu} + \hat{\alpha}_k &= \bar{y}_{k\cdot} \\ \hat{\mu} + \sum n_i \hat{\alpha}_i / N &= \bar{y}_{\cdot\cdot} \,. \end{aligned} \tag{2.9}$$

The latter equation is redundant since it equals the sum of each of the former equations multiplied by its n_k. Hence there is no unique solution to the equations. But we can get *a* solution by placing a constraint on the α_i. A commonly-used constraint, which clearly makes the equations easy to solve, is $\sum n_i \hat{\alpha}_i = 0$, which then yields

$$\hat{\mu} = \bar{y}_{\cdot\cdot} \qquad \text{and} \qquad \hat{\alpha}_k = \bar{y}_{k\cdot} - \bar{y}_{\cdot\cdot} \,. \tag{2.10}$$

However, some people find it distasteful to have a constraint which depends on the sample sizes; these may be, to some extent, random variables for a given experiment.

This gives the ML estimators of μ and α_i. For σ^2, (2.2) gives the same estimator as does (2.1).

c. Generalized likelihood ratio test

A starting point for many statistical investigations is to test the hypothesis

$$H_0 : \mu_i \textit{ all equal.}$$

Denoting the common value of μ_i under H_0 as μ, the ML estimator under H_0 is $\hat{\mu}_0 = \bar{y}_{\cdot\cdot}$ and the corresponding ML estimator of σ^2 is

$$\begin{aligned} \hat{\sigma}_0^2 &= \frac{1}{N} \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 \\ &= \frac{1}{N} \left[\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \right] . \end{aligned}$$

This gives a generalized likelihood ratio statistic (see Section 1.7b–i) of

$$\begin{aligned} \Lambda &= \frac{(2\pi\hat{\sigma}_0^2)^{-N/2} \exp\left(-N\hat{\sigma}_0^2 / 2\hat{\sigma}_0^2\right)}{(2\pi\hat{\sigma}^2)^{-N/2} \exp\left(-N\hat{\sigma}^2 / 2\hat{\sigma}^2\right)} \\ &= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-N/2} . \end{aligned} \tag{2.11}$$

Taking logarithms and multiplying by -2 gives

$$
\begin{aligned}
-2\log\Lambda &= N\log\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \\
&= N\log\left(\frac{\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot})^2+\sum_i\sum_j(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}{\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot})^2}\right) \\
&= N\log\left(1+\frac{m-1}{N-m}F\right), \qquad (2.12)
\end{aligned}
$$

where

$$
F=\frac{\sum_i\sum_j(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2/(m-1)}{\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot})^2/(N-m)}
$$

is the usual F-statistic from the analysis of variance for a one-way classification. Under H_0 : μ_i *all equal*, F has a central $\mathcal{F}$-distribution with numerator degrees of freedom $m-1$ and denominator degrees of freedom $N-m$. We denote this by $F \sim \mathcal{F}_{N-m}^{m-1}$.

Rejecting the null hypothesis when the likelihood ratio, Λ, is small ("the null hypothesis is unlikely") is equivalent to $-2\log\Lambda$ being large or $F \geq \mathcal{F}_{N-m,1-\alpha}^{m-1}$, the $100(1-\alpha)\%$ percentile of the $\mathcal{F}-$distribution (i.e., $P\{F_{N-m}^{m-1} \geq \mathcal{F}_{N-m,1-\alpha}^{m-1}\} = \alpha$) .

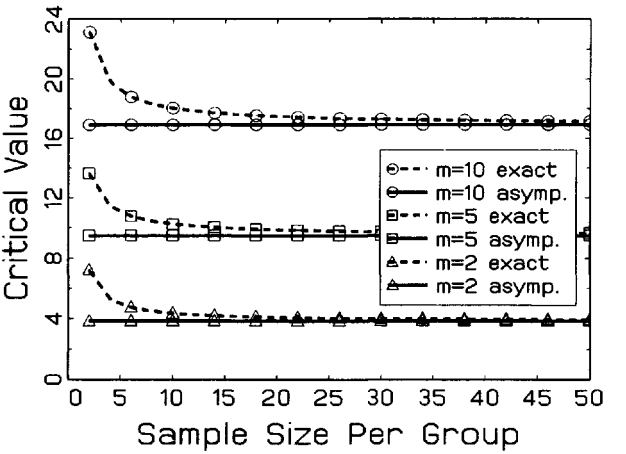
Asymptotic theory (Lehmann, 1986, p. 486) tells us that the large-sample distribution of $-2\log\Lambda$ under H_0 is chi-square with degrees of freedom equal to the difference in the number of parameters in the parameter space and the number under H_0. In this situation there are $m+1$ parameters $\mu_1, \mu_2, \ldots, \mu_m$ and σ^2 . Under H_0 there is a single μ-parameter and σ^2, so the difference is $m+1-2 = m-1$.

The large-sample test would therefore be to reject H_0 when $-2\log\Lambda \geq \chi^2_{m-1,1-\alpha}$. Exact distribution theory based on the $\mathcal{F}$-distribution is to reject H_0 when $-2\log\Lambda \geq N\log\left(1+\frac{m-1}{N-m}\mathcal{F}_{N-m,1-\alpha}^{m-1}\right)$. How do these compare? Figure 2.1 plots the χ^2- and $\mathcal{F}$-based critical values for $\alpha = 0.05$, $m = 2$, 5, and 10, versus N. For total sample sizes $N < 50$ the differences can be appreciable. For the impact of using the chi-square critical values see E 2.2.

d. Confidence intervals

Confidence intervals for μ_i, $\mu_i - \mu_k$, and σ^2 are easily derived and widely available (Snedecor and Cochran, 1989, Chapters 5 and 6). For completeness we list them here. In doing so we emphasize that these intervals are for fixed effects models only, not for mixed models. For

Figure 2.1: Critical values based on $\mathcal{F}$-distribution (dashed lines) and χ^2-distributions (solid lines) plotted versus N for $m = 10$ (top set of lines), 5 (middle set of lines), and 2 (bottom set of lines).



example, with $\mu_i = \mu + \alpha_i$, the interval shown here for $\mu_i - \mu_k = \alpha_i - \alpha_k$ is for the αs being fixed effects. The case of mixed models is considered in Chapter 6.

– i. ***For means***

A confidence interval for μ_i is given by

$$\bar{y}_{i\cdot} \pm t_{N-m,\alpha/2} \frac{s}{\sqrt{n_i}},$$

where $t_{\nu,\alpha}$ denotes the upper $100\alpha\%$ percentile of the $\mathcal{T}$-distribution on ν degrees of freedom, that is,

$$P\{\mathcal{T}_\nu > t_{\nu,\alpha}\} = \alpha.$$

– ii. ***For differences in means***

A confidence interval for $\mu_i - \mu_k$ is given by

$$\bar{y}_{i\cdot} - \bar{y}_{k\cdot} \pm t_{N-m,\alpha/2} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}.$$

– iii. ***For linear combinations***

A linear combination of the means is defined as $\sum_i c_i\mu_i$ where the c_i are known constants. A confidence interval for $\sum_i c_i\mu_i$ is given by

$$\sum_i c_i\bar{y}_{i\cdot} \pm t_{N-m,\alpha/2}\sqrt{s^2\sum_i \frac{c_i^2}{n_i}}\,.$$

– iv. ***For the variance***

For σ^2 the confidence interval is given by

$$\left(\frac{(N-m)s^2}{\chi^2_{N-m,1-\alpha/2}}, \frac{(N-m)s^2}{\chi^2_{N-m,\alpha/2}}\right).$$

We note that although this interval is exact under model (2.1) or (2.2), it is not robust to violations of the normality assumption (Snedecor and Cochran, 1989, p. 252) and should therefore be used with caution.

e. Hypothesis tests

Hypothesis tests concerning the means are straightforward. For example, to test

$$H_0 : \sum_i c_i\mu_i = \eta_0,$$

where η_0 is a hypothesized value, we use the t-statistic

$$t = \frac{\sum_i c_i\bar{y}_{i\cdot} - \eta_0}{s\sqrt{\sum_i c_i^2/n_i}}.$$

If the alternative is $H_A : \sum_i c_i\mu_i \neq \eta_0$, we reject when $|t| > t_{N-m,\alpha/2}$; if the alternative is $H_A : \sum_i c_i\mu_i > \eta_0$, we reject when $t > t_{N-m,\alpha}$; if the alternative is $H_A : \sum_i c_i\mu_i < \eta_0$, we reject when $t < -t_{N-m,\alpha}$.

2.2 NORMALITY, RANDOM EFFECTS AND ML

a. Model

In accordance with Section 1.1, when we assume a random effects classification we attribute a distribution to the effects of the levels of the factor.

A model corresponding to (2.1) is

$$\begin{aligned} \mathrm{E}[y_{ij}\,] &= \mu_i \\ y_{ij} &\sim \text{indep. } \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i &\sim \text{i.i.d. } \mathcal{N}(\mu, \sigma_\mu^2), \end{aligned} \tag{2.13}$$

where we have used the usual notation, i.i.d., to indicate that the random variables are mutually *independent with identical distributions.* Since μ_i appears in the expected value of y_{ij} but is later assumed to be random, (2.13) is a somewhat sloppy specification of the distributions. More precisely, the conditional distribution of y_{ij} given μ_i (we indicate the conditioning on μ_i with the vertical bar) is normal with mean μ and variance σ^2 and the distribution of μ_i is $\mathcal{N}(\mu, \sigma_\mu^2)$. This is written as follows:

$$\begin{aligned} \mathrm{E}[y_{ij}|\mu_i] &= \mu_i \\ y_{ij}\,|\mu_i &\sim \text{indep. } \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i &\sim \text{i.i.d. } \mathcal{N}(\mu, \sigma_\mu^2). \end{aligned} \tag{2.14}$$

An equivalent model, corresponding to (2.2), is traditionally written as

$$\begin{aligned} \mathrm{E}[y_{ij}|a_i] &= \mu + a_i \\ y_{ij}|a_i &\sim \text{indep. } \mathcal{N}(\mu + a_i, \sigma^2) \\ a_i &\sim \text{i.i.d. } \mathcal{N}(0, \sigma_a^2), \end{aligned} \tag{2.15}$$

where the notation $\sigma_a^2 = \sigma_\mu^2$ is now used in place of σ_μ^2.

It is appropriate at this stage to contrast the random effects model with a Bayesian approach. In a Bayesian approach the parameters μ_i would be assumed to have a distribution just as does the random effects model. However, the similarity ends there. In a true Bayesian approach the distribution of the μ_i would represent subjective information on the μ_i, not a distribution across tangible populations (e.g., across animals). The Bayesian approach would further hypothesize a distribution for all other unknown parameters (σ^2 in this case). The method of estimating the parameters would also be different.

– i. *Covariances caused by random effects*

A fundamental difference between the fixed and random effects models is that the observations, y_{ij}, in a random effects model are not independent. In fact, the assumption of a random factor can be viewed as a

convenient way to specify a variance-covariance structure. Essentially, observations with model equations that contain the same random effect are correlated. Using (1.6) yields

$$\begin{aligned}\operatorname{cov}(y_{ij}, y_{il}) &= \operatorname{cov}(\mathrm{E}[y_{ij}|a_i], \mathrm{E}[y_{il}|a_i]) + \mathrm{E}[\operatorname{cov}(y_{ij}, y_{il}|a_i)] \\ &= \operatorname{cov}(\mu + a_i, \mu + a_i) + 0 \\ &= \operatorname{cov}(a_i, a_i) = \sigma_a^2. \end{aligned} \tag{2.16}$$

Also

$$\begin{aligned}\operatorname{var}(y_{ij}) &= \operatorname{var}(\mathrm{E}[y_{ij}|a_i]) + \mathrm{E}[\operatorname{var}(y_{ij}|a_i)] \\ &= \operatorname{var}(\mu + a_i) + \mathrm{E}[\sigma^2] \\ &= \sigma_a^2 + \sigma^2. \end{aligned} \tag{2.17}$$

Thus we have an *intraclass correlation* of

$$\operatorname{corr}(y_{ij}, y_{il}) = \frac{\sigma_a^2}{\sqrt{(\sigma_a^2 + \sigma^2)(\sigma_a^2 + \sigma^2)}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}. \tag{2.18}$$

– ii. *Likelihood*

Since observations within the same level of a random effect are correlated the likelihood for the random model is more complicated than for the fixed effects model. For $\mathbf{y}_i = [y_{i1}\ y_{i2} \ldots y_{in_i}]'$ the model (2.15) has

$$\mathbf{y}_i \sim \mathcal{N}(\mu \mathbf{1}_{n_i}, \mathbf{V}_i),$$

where $\mathbf{V_i} = \sigma^2 \mathbf{I}_{n_i} + \sigma_a^2 \mathbf{J}_{n_i}$, $\mathbf{I}_n$ is the identity matrix of order n, $\mathbf{J}_n$ is an $n \times n$ matrix of all ones, and $\mathbf{1}_n$ is a column vector of all ones of order n. It is straightforward to show (see Section M.1 in Appendix M) that

$$\mathbf{V}_i^{-1} = \frac{1}{\sigma^2}\mathbf{I}_{n_i} - \frac{\sigma_a^2}{\sigma^2(\sigma^2 + n_i\sigma_a^2)}\mathbf{J}_{n_i}$$

and $|\mathbf{V}_i| = (\sigma^2 + n_i\sigma_a^2)(\sigma^2)^{n_i - 1}$. From these the likelihood is

$$L = \prod_{i=1}^{m} (2\pi)^{-n_i/2} |\mathbf{V}_i| \exp\{-\tfrac{1}{2}(\mathbf{y}_i - \mu \mathbf{1}_{n_i})' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mu \mathbf{1}_{n_i})\}, \tag{2.19}$$

or

$$\begin{aligned} l &= \log L \\ &= -\tfrac{1}{2} N \log 2\pi - \tfrac{1}{2} \textstyle\sum_i \log(\sigma^2 + n_i\sigma_a^2) - \tfrac{1}{2}(N - m) \log \sigma^2 \\ &\quad - \frac{1}{2\sigma^2} \sum_i \sum_j (y_{ij} - \mu)^2 + \frac{\sigma_a^2}{2\sigma^2} \sum_i \frac{(y_{i\cdot} - n_i\mu)^2}{\sigma^2 + n_i\sigma_a^2}. \end{aligned} \tag{2.20}$$

b. Balanced data

– i. *Likelihood*

Balanced data have $n_i = n \; \forall \; i$. This greatly simplifies $\log L$ so that it becomes

$$\begin{aligned} l = \log L &= -\tfrac{1}{2}N \log 2\pi - \tfrac{1}{2}m(n-1)\log \sigma^2 - \tfrac{1}{2}m[\log(\sigma^2 + n\sigma_a^2)] \\ &\quad - \frac{\Sigma_i \Sigma_j (y_{ij} - \mu)^2}{2\sigma^2} + \frac{n^2\sigma_a^2 \Sigma_i (\bar{y}_{i\cdot} - \mu)^2}{2\sigma^2(\sigma^2 + n\sigma_a^2)}. \end{aligned} \tag{2.21}$$

The last two terms in (2.21) can be rewritten after a little algebra (Searle et al., 1992, p. 80) to involve

$$\text{SSA} = \Sigma_i n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \quad \text{and} \quad \text{SSE} = \Sigma_i \Sigma_j (y_{ij} - \bar{y}_{\cdot\cdot})^2, \tag{2.22}$$

the familiar sums of squares for classes and error, respectively, in the usual analysis of variance of data from a 1-way classification. Further simplification comes from defining

$$\lambda = \sigma^2 + n\sigma_a^2, \tag{2.23}$$

so that l is then

$$l = -\tfrac{1}{2}N\log 2\pi - \tfrac{1}{2}m(n-1)\log\sigma^2 - \tfrac{1}{2}m\log\lambda - \frac{\text{SSE}}{2\sigma^2} - \frac{\text{SSA}}{2\lambda} - \frac{mn(\bar{y}_{\cdot\cdot} - \mu)^2}{2\lambda}. \tag{2.24}$$

Introducing λ in place of $\sigma^2 + n\sigma_a^2$ simplifies the ML estimation process. Then ML yields estimators of σ^2 and σ_a^2 through the standard property of ML estimation that the ML estimator of a function of parameters is that same function of ML estimators of the parameters.

– ii. *ML equations and their solutions*

The maximum likelihood equations are those equations obtained by equating to zero the partial derivatives of $\log L$ with respect to μ, σ^2 and λ:

$$\begin{aligned} l_\mu &= \frac{\partial l}{\partial \mu} = \frac{mn(\bar{y}_{\cdot\cdot} - \mu)}{\lambda}, \\ l_{\sigma^2} = \frac{\partial l}{\partial \sigma^2} &= \frac{-m(n-1)}{2\sigma^2} + \frac{\text{SSE}}{2\sigma^4} = \frac{-m(n-1)}{2\sigma^4}\left[\sigma^2 - \frac{\text{SSE}}{m(n-1)}\right], \\ l_\lambda &= \frac{\partial l}{\partial \lambda} = \frac{-m}{2\lambda} + \frac{\text{SSA}}{2\lambda^2} + \frac{mn(\bar{y}_{\cdot\cdot} - \mu)^2}{2\lambda^2} \\ &= \frac{-m}{2\lambda^2}\left(\lambda - \frac{\text{SSA}}{m}\right) + \frac{mn(\bar{y}_{\cdot\cdot} - \mu)^2}{2\lambda^2}. \end{aligned} \tag{2.25}$$

In equating these partial derivatives to zero we change the parameter symbols μ, σ_e^2 and λ to be $\dot{\mu}$, $\dot{\sigma}_e^2$ and $\dot{\lambda}$ representing solutions to those equations, and get those solutions as

$$\dot{\mu} = \bar{y}_{\cdot\cdot}, \quad \dot{\sigma}^2 = \text{MSE}, \quad \dot{\lambda} = \frac{\text{SSA}}{m} = \left(1 - \frac{1}{m}\right)\text{MSA}$$

and then

$$\dot{\sigma}_a^2 = \frac{\dot{\lambda} - \dot{\sigma}^2}{n} = \frac{(1 - 1/m)\text{MSA - MSE}}{n}, \tag{2.26}$$

where

$$\text{MSA} = \text{SSA}/(m-1) \quad \text{and} \quad \text{MSE} = \text{SSE}/m(n-1).$$

These are the solutions to the ML equations. But they are not necessarily the ML estimators, even though they maximize the likelihood function, L, for variation in μ, σ^2 and σ_a^2.

The theory of maximum likelihood tells us that solutions of ML equations do indeed maximize the likelihood function if the matrix of second derivatives (known as the *Hessian*) of the likelihood is negative definite when the parameters in the Hessian are replaced by the solutions. For $\dot{\mu}$, $\dot{\sigma}^2$ and $\dot{\gamma}$ this is left as E 2.17 at the end of this chapter.

– iii. *ML estimators*

The very definition of ML demands that the likelihood be maximized over the *parameter space.* And in the 1-way classification this space is, from the nature of the parameters, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$ and $0 \leq \sigma_a^2 < \infty$. Fortunately, in the 1-way classification $\dot{\sigma}_a^2$ is the only one of the three ML solutions $\dot{\mu}$, $\dot{\sigma}^2$ and $\dot{\mu}$ that is not necessarily in the parameter space.

We consider the solutions $\dot{\mu}$, $\dot{\sigma}^2$ and $\dot{\sigma}_a^2$ in turn. First, $\dot{\mu}$ does not depend on $\dot{\sigma}^2$ or $\dot{\sigma}_a^2$, and since $\dot{\mu} = \bar{y}_{\cdot\cdot}$ is clearly in the space of μ it is the MLE of μ:

$$\text{MLE}(\mu) = \hat{\mu} = \dot{\mu} = \bar{y}_{\cdot\cdot}\,.$$

Also $\dot{\sigma}^2 = \text{MSE}$ is in the parameter space for σ^2, since MSE is never negative (and we exclude the naive case where $y_{ij} = \bar{y}_{i\cdot}$ $\forall$ i and j, which would give $\dot{\sigma}^2 = 0$). But since $\dot{\sigma}_a^2$ depends on $\text{MSE} = \dot{\sigma}^2$, we must ensure not just that $\hat{\sigma}_a^2$ is in the parameter space for σ_a^2 but that the pair of estimators $(\hat{\sigma}^2, \hat{\sigma}_a^2)$ is in the 2-space defined by (σ^2, σ_a^2). As a result, we find that when $\dot{\sigma}_a^2 \leq 0$ then $\dot{\sigma}^2 = \text{MSE}$ is not the MLE of

σ^2. Establishing the ML estimators $\hat{\sigma}^2$ and $\hat{\sigma}_a^2$ from $\dot{\sigma}^2$ and $\dot{\sigma}_a^2$ through taking account of the possibility of $\dot{\sigma}_a^2$ being non-positive was first done by Herbach (1959) and is summarized in Searle et al. (1992, pp. 81–83). The consequences are that the MLEs of σ^2 and σ_a^2 are as follows:

$$\hat{\sigma}^2 = \begin{cases} \text{MSE} & \text{if } \left(1 - \frac{1}{m}\right)\text{MSA} \geq \text{MSE}, \\ \frac{\text{SST}}{mn} & \text{if } \left(1 - \frac{1}{m}\right)\text{MSA} < \text{MSE}, \end{cases} \tag{2.27}$$

and

$$\hat{\sigma}_a^2 = \begin{cases} \left[\left(1 - \frac{1}{m}\right)\text{MSA} - \text{MSE}\right] \Big/ n & \text{if } \left(1 - \frac{1}{m}\right)\text{MSA} \geq \text{MSE}, \\ 0 & \text{if } \left(1 - \frac{1}{m}\right)\text{MSA} < \text{MSE}. \end{cases} \tag{2.28}$$

Although this is certainly the correct way of stating the MLEs, we also state them in a manner that may well be more useful for data analysts. This is because we state the data conditions first:

$$\text{if } \left(1 - \frac{1}{m}\right)\text{MSA} \geq \text{MSE} \text{ then } \hat{\sigma}_a^2 = \left[\left(1 - \frac{1}{m}\right)\text{MSA} - \text{MSE}\right] \Big/ n \text{ and } \hat{\sigma}^2 = \text{MSE}, \tag{2.29}$$

$$\text{if } \left(1 - \frac{1}{m}\right)\text{MSA} < \text{MSE} \text{ then } \hat{\sigma}_a^2 = 0 \text{ and } \hat{\sigma}^2 = \frac{\text{SST}}{mn}, \tag{2.30}$$

where $\text{SST} = \Sigma_i \Sigma_j (y_{ij} - \bar{y}_{..})^2$ is the total sum of squares corrected for the mean.

– iv. *Expected values and bias*

The expected value of $\hat{\mu} = \bar{y}_{..}$ is easy:

$$\text{E}[\hat{\mu}] = \text{E}[\bar{y}_{..}] = \mu,$$

i.e., the ML estimator $\hat{\mu}$ is unbiased. And expected values of the ML *solutions* (not estimators) are easily defined: from (2.26)

$$\text{E}[\dot{\sigma}^2] = \text{E}[\text{MSE}] = \sigma^2$$

and

$$\begin{aligned} \mathrm{E}[\dot{\sigma}_a^2] &= \frac{(1-1/m)\mathrm{E}[\mathrm{MSA}] - \mathrm{E}[\mathrm{MSE}]}{n} \\ &= \frac{(1-1/m)(\sigma^2 + n\sigma_a^2) - \sigma^2}{n} \\ &= (1-1/m)\sigma_a^2 - \sigma^2/mn. \end{aligned}$$

But this direct derivation of expected values does not carry over to ML estimators. The reason is that, as seen in (2.27) and (2.28), each of those estimators takes two different forms: e.g., $\hat{\sigma}^2$ is MSE if $(1-1/m)\mathrm{MSA} \geq \mathrm{MSE}$, but $\hat{\sigma}^2$ is SST/mn if $(1-1/m)\mathrm{MSA} < \mathrm{MSE}$. Therefore, for

$$p = P\{(1-1/m)\mathrm{MSA} < \mathrm{MSE}\} \tag{2.31}$$

$$\mathrm{E}[\hat{\sigma}^2] = (1-p)\mathrm{E}[\mathrm{MSE}|\dot{\sigma}_a^2 \geq 0] + p\mathrm{E}[\mathrm{SST}/m|\dot{\sigma}_a^2 < 0]. \tag{2.32}$$

Similarly, and because $\hat{\sigma}_a^2 \not< 0$,

$$\mathrm{E}[\hat{\sigma}_a^2] = (1-p)\mathrm{E}[\dot{\sigma}_a^2|\dot{\sigma}_a^2 \geq 0]. \tag{2.33}$$

These expectations have no closed form. Not only does p depend on the values σ^2 and σ_a^2, because p can be expressed as

$$p = P\left\{\mathcal{F}_{m-1}^{m(n-1)} \geq (1-1/m)(1+n\sigma_a^2/\sigma^2)\right\},$$

but also, the expectations in (2.32) and (2.33) are conditional expectations over just parts of the real line, the non-negative part and, separately, the negative part.

Finally, with the ML estimators having expected values with no closed form, that is also the case for bias.

– v. *Asymptotic sampling variances*

With $\hat{\mu} = \bar{y}..$ it is easily shown that

$$\mathrm{var}(\hat{\mu}) = \frac{\sigma^2 + n\sigma_a^2}{mn}.$$

There is a very general result in ML estimation that the large-sample asymptotic dispersion matrix of ML estimators is the inverse of the negative of the expected value of the matrix of second derivatives of the log likelihood, i.e., of l of (2.21). This general result includes the fact for LMs that covariances between MLs of fixed effects and variance

components are zero. See, for example, Searle et al. [1992, p. 239, eq. (39)]. Thus with (2.25) leading to

$$l_{\sigma^2} = \frac{-m(n-1)}{2\sigma^2} + \frac{\text{SSE}}{2\sigma^4}$$

and

$$l_{\lambda} = \frac{-m}{2\lambda} + \frac{\text{SSA}}{2\lambda^2} + \frac{mn(\bar{y}_{\cdot\cdot} - \mu)^2}{2\lambda^2},$$

we get the second derivatives

$$l_{\sigma^2\sigma^2} = \frac{m(n-1)}{2\sigma^4} - \frac{2\text{SSE}}{2\sigma^6},$$

$$l_{\sigma^2\lambda} = 0$$

and

$$l_{\lambda\lambda} = \frac{m}{2\lambda^2} - \frac{2\text{SSA}}{2\lambda^3} - \frac{2mn(\bar{y}_{\cdot\cdot} - \mu)^2}{2\lambda^3}.$$

Thus

$$-\text{E}[l_{\sigma^2\sigma^2}] = \frac{-m(n-1)}{2\sigma^4} + \frac{2m(n-1)\sigma^2}{2\sigma^6} = \frac{m(n-1)}{2\sigma^4}$$

$$-\text{E}[l_{\lambda\lambda}] = \frac{-m}{2\lambda^2} + \frac{(m-1)\lambda}{\lambda^3} + \frac{mnm(n\sigma^2 + n^2\sigma_a^2)}{\lambda^3 m^2 n^2} = \frac{m}{2\lambda^2}.$$

Therefore, with $l_{\sigma^2\lambda} = 0$

$$\text{var}\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\lambda} \end{bmatrix} \rightarrow \left(-\text{E}\begin{bmatrix} l_{\sigma^2\sigma^2} & l_{\sigma^2\lambda} \\ l_{\sigma^2\lambda} & l_{\lambda\lambda} \end{bmatrix}\right)^{-1} = \begin{bmatrix} \frac{2\sigma^4}{m(n-1)} & 0 \\ 0 & \frac{2\lambda^2}{m} \end{bmatrix}. \tag{2.34}$$

Then, with $\hat{\sigma}_a^2 = (\hat{\lambda} - \hat{\sigma}^2)/n$, the large-sample dispersion matrix for the MLEs of the variance components is

$$\text{var}\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\sigma}_a^2 \end{bmatrix} = 2\sigma^4 \begin{bmatrix} \frac{1}{m(n-1)} & \frac{-1}{mn(n-1)} \\ \frac{-1}{mn(n-1)} & \frac{1}{n^2}\left(\frac{\lambda^2/\sigma^4}{m} + \frac{1}{m(n-1)}\right) \end{bmatrix}. \tag{2.35}$$

Note that in (2.34) the large-sample variance of the ML estimator $\hat{\sigma}^2$ is $2\sigma^4/m(n-1)$. This is the same as $\text{var}(\dot{\sigma}^2) = \text{var}(\text{MSE})$. But $\dot{\sigma}^2 = \text{MSE}$ is not the same as $\hat{\sigma}^2$. As in (2.27), $\hat{\sigma}^2$ is MSE when $\dot{\sigma}_a^2 \geq 0$, but $\dot{\sigma}_a^2 < 0$ leads to $\hat{\sigma}^2 = \text{SST}/mn$. However, (2.34) is an asymptotic result, in which $\hat{\sigma}^2$, by virtue of being an ML estimator, is consistent, and so cannot be negative. Hence, in that asymptotic situation, $\hat{\sigma}_a^2 < 0$ never

occurs and so $\hat{\sigma}^2$ is never SST/mn. It is always MSE, with variance $2\sigma^4/m(n-1)$ as in (2.34).

In contrast, the exact variance of $\hat{\sigma}^2$ is, using p of (2.31),

$$\text{var}(\hat{\sigma}^2) = (1-p)\text{E}[(\text{MSE})^2|\dot{\sigma}_a^2 \geq 0]+p\text{E}[(\text{SST})^2|\dot{\sigma}_a^2 < 0]/m^2n^2-(\text{E}[\hat{\sigma}^2])^2.$$

Again intractability is apparent, and numerical evaluation has to be used for each particular case. See Yu et al. (1994).

– vi. *REML estimation*

In contrast to the ML solutions of (2.26) the REML solutions are $\dot{\sigma}^2 = \text{MSE}$ (the same as ML) and $\dot{\sigma}^2 = (\text{MSA} - \text{MSE})/n$, as can be derived from the Chapter 6 general REML equation.

c. Unbalanced data

– i. *Likelihood*

Following $\lambda = \sigma^2 + n\sigma_a^2$ in (2.23) we now define

$$\lambda_i = \sigma^2 + n_i\sigma_a^2. \tag{2.36}$$

Then the likelihood of (2.21), after writing $y_{ij} - \mu$ as $y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \mu$ and simplifying, becomes

$$l = -\tfrac{1}{2}N\log 2\pi - \tfrac{1}{2}\Sigma_i \log\lambda_i - \tfrac{1}{2}(N-m)\log\sigma^2 - \frac{\text{SSE}}{2\sigma^2} - \Sigma_i \tfrac{n_i(\bar{y}_{i\cdot}-\mu)^2}{2\lambda_i}. \tag{2.37}$$

– ii. *ML equations and their solutions*

With $\partial\lambda_i/\partial\sigma^2 = 1$ and $\partial\lambda_i/\partial\sigma_a^2 = n_i$ we differentiate l of (2.37) to get (using $l_\theta \equiv \partial \log L/\partial\theta$)

$$l_\mu = \Sigma_i \frac{n_i(\bar{y}_{i\cdot}-\mu)}{\lambda_i}, \tag{2.38}$$

$$l_{\sigma^2} = \frac{-(N-m)}{2\sigma^2} - \tfrac{1}{2}\Sigma_i \tfrac{1}{\lambda_i} + \frac{\text{SSE}}{2\sigma^4} + \Sigma_i \tfrac{n_i(\bar{y}_{i\cdot}-\mu)^2}{2\lambda_i^2}, \tag{2.39}$$

and

$$l_{\sigma_a^2} = -\tfrac{1}{2}\Sigma_i \tfrac{n_i}{\lambda_i} + \Sigma_i \tfrac{n_i^2(\bar{y}_{i\cdot}-\mu)^2}{\lambda_i^2}. \tag{2.40}$$

The ML equations are obtained by equating the above expressions to zero using $\dot{\mu}$, $\dot{\sigma}^2$ and $\lambda_i = \dot{\sigma}_e^2 + n_i\dot{\sigma}_a^2$ as the solutions. Carrying out this procedure with l_μ of (2.38) gives

$$\dot{\mu} = \sum_i \frac{n_i\bar{y}_{i\cdot}}{\dot{\lambda}_i} \Big/ \sum_i \frac{n_i}{\dot{\lambda}_i} = \frac{\sum_i \frac{n_i\bar{y}_{i\cdot}}{\dot{\sigma}^2 + n_i\dot{\sigma}_a^2}}{\sum_i \frac{n_i}{\dot{\sigma}^2 + n_i\dot{\sigma}_a^2}} = \frac{\sum_i \bar{y}_{i\cdot}/\dot{\text{var}}(\bar{y}_{i\cdot})}{\sum_i [1/\dot{\text{var}}(\bar{y}_{i\cdot})]}. \tag{2.41}$$

We see that $\dot{\mu}$ is a weighted average of the $\bar{y}_{i\cdot}$ weighted inversely by an estimate of $\text{var}(\bar{y}_{i\cdot}) = \sigma_a^2 + \sigma^2/n_i$.

Derivation of $\dot{\sigma}_a^2$ and $\dot{\sigma}^2$ comes from equating the right-hand sides of (2.39) and (2.40) to zero, so giving

$$\frac{\text{SSE}}{\dot{\sigma}^4} - \frac{N-m}{\dot{\sigma}^2} + \sum_i \frac{n_i(\bar{y}_{i\cdot} - \dot{\mu})^2}{\dot{\lambda}_i^2} - \sum_i \frac{1}{\dot{\lambda}_i} = 0 \tag{2.42}$$

and

$$\sum_i \frac{n_i^2(\bar{y}_{i\cdot} - \dot{\mu})^2}{\dot{\lambda}_i^2} = \sum_i \frac{n_i}{\dot{\lambda}_i}. \tag{2.43}$$

With $\dot{\lambda} = \dot{\sigma}_e^2 + n_i\dot{\sigma}_a^2$ occurring in the denominators of the terms being summed (over i) in these equations, there is clearly no analytic solution for the estimators, but there is when the data are balanced (i.e., $n_i = n$ and $\lambda_i = \lambda \; \forall \; i$).

– iii. *ML estimators*

As with balanced data, solutions $\dot{\mu}$, $\dot{\sigma}^2$ and $\dot{\sigma}_a^2$ are ML estimators only if the triplet $(\dot{\mu},\ \dot{\sigma}^2,\ \dot{\sigma}_a^2)$ is in the 3-space of $(\mu,\ \sigma^2,\ \sigma_a^2)$. And in ensuring that this is achieved, the negativity problem raises its head again. For each data set, equations (2.41), (2.42) and (2.43) have to be solved numerically, using some iterative method suited to the numerical solution of non-linear equations. After doing this, the ML estimators are as follows:

when $\dot{\sigma}_a^2 \geq 0$,

$$\hat{\sigma}^2 = \dot{\sigma}^2, \quad \hat{\sigma}_a^2 = \dot{\sigma}_a^2 \quad \text{and} \quad \hat{\mu} = \dot{\mu}; \tag{2.44}$$

when $\dot{\sigma}_a^2 < 0$,

$$\hat{\sigma}^2 = \text{SST}/N, \quad \hat{\sigma}_a^2 = 0 \quad \text{and} \quad \hat{\mu} = \bar{y}_{\cdot\cdot}\,. \tag{2.45}$$

In the latter case, when $\dot{\sigma}_a^2 < 0$, the argument for having $\hat{\sigma}_a^2 = 0$ is essentially the same as with balanced data, whereupon it is left to the reader to show that log L reduces to being such that on equating its derivatives to zero one obtains $\hat{\sigma}^2 = \text{SST}/N$, as in (2.45) for balanced data and $\hat{\mu} = \bar{y}_{..}$ (see E 2.5). Having been derived by the method of maximum likelihood, the estimators in (2.44) and (2.45) are, as is well known, asymptotically normally distributed.

The question might well be raised as to what to do if the numerical solution of (2.42) and (2.43) yields a negative value for $\dot{\sigma}^2$. Fortunately, it can be shown that $L = e^l \to 0$ as σ^2 tends to zero or to infinity, and so L must have a maximum at a positive value of σ^2 (see E 2.6).

d. Bias

With balanced data we were able to specify p, the probability of the solution for $\dot{\sigma}^2$ to the ML equations being negative – in (2.31). But with unbalanced data $F = \text{MSA}/\text{MSE}$ does not have a distribution that is proportional to an $\mathcal{F}$, so this probability cannot be easily specified. Moreover, although we know that $\hat{\sigma}^2 = \text{SST}/N$ with probability p, and the expected value of SST is readily derived, the expected value of $\hat{\sigma}^2$ when $\dot{\sigma}_a^2 < 0$ cannot easily be derived. Thus, in general, the bias in the solutions obtained to (2.42) cannot be derived analytically.

e. Sampling variances

Large-sample variances come from a matrix similar to (2.34), namely the inverse of the negative of the expected value of the Hessian (matrix of second derivatives) of log L with respect to μ, σ^2 and σ_a^2. Keeping in mind that, by definition, $\sigma_a^2 > 0$ (because if $\sigma_a^2 = 0$ the model and L change), we differentiate the three first differentials of (2.38), (2.39) and (2.40) and take expected values of the resulting second differentials. [Details are shown in Searle et al. (1992) in Sec. 3.7b.] Arraying these expected values in a matrix gives, after inverting that matrix, the matrix of large-sample variance-covariance matrix:

$$\text{var}\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \\ \hat{\sigma}_a^2 \end{bmatrix} \simeq \begin{bmatrix} \Sigma_i \frac{n_i}{\lambda_i} & 0 & 0 \\ 0 & \frac{N-m}{2\sigma^4} + \frac{1}{2}\Sigma_i \frac{1}{\lambda_i^2} & \frac{1}{2}\Sigma_i \frac{n_i}{\lambda_i^2} \\ 0 & \frac{1}{2}\Sigma_i \frac{n_i}{\lambda_i^2} & \frac{1}{2}\Sigma_i \frac{n_i^2}{\lambda_i^2} \end{bmatrix}^{-1} . \qquad (2.46)$$

Therefore

$$\text{var}(\hat{\mu}) \simeq \left(\sum_i \frac{n_i}{\lambda_i}\right)^{-1} = \left(\sum_i \frac{n_i}{\sigma^2 + n_i\sigma_a^2}\right)^{-1}$$

and

$$\text{var}\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\sigma}_a^2 \end{bmatrix} \simeq \frac{2}{D}\begin{bmatrix} \sum_i \frac{n_i^2}{\lambda_i^2} & -\sum_i \frac{n_i}{\lambda_i^2} \\ -\sum_i \frac{n_i}{\lambda_i^2} & \frac{N-m}{\sigma^4} + \sum_i \frac{1}{\lambda_i^2} \end{bmatrix} \tag{2.47}$$

where

$$D = \frac{N-m}{\sigma^4}\sum_i \frac{n_i^2}{\lambda_i^2} + \sum_i \frac{1}{\lambda_i^2}\sum_i \frac{n_i^2}{\lambda_i^2} - \left(\sum_i \frac{n_i}{\lambda_i^2}\right)^2.$$

2.3 NORMALITY, RANDOM EFFECTS AND REML

a. Balanced data

– i. *Likelihood*

For the 1-way classification with model equation $\mathrm{E}[y_{ij}] = \mu + a_i$ the part of the likelihood of y not involving fixed effects is simply that part not involving μ. And for balanced data that is easily derived. From (2.24) we reconstruct L, the likelihood of $\mathbf{y}$ and write it as

$$L(\mu, \sigma^2, \sigma_a^2|\mathbf{y}) = \frac{\exp\left\{-\frac{1}{2}\left[\frac{\text{SSE}}{\sigma^2} + \frac{\text{SSA}}{\lambda} + \frac{(\bar{y}_{..} - \mu)^2}{\lambda/mn}\right]\right\}}{(2\pi)^{\frac{1}{2}mn}\sigma^{2[\frac{1}{2}m(n-1)]}\lambda^{\frac{1}{2}m}}.$$

Since $\bar{y}_{..}$ is independent of both SSE and SSA, the preceding expression can be factored as

$$L(\mu, \sigma^2\, \sigma_a^2|\mathbf{y}) = L(\mu|\bar{y}_{..})L(\sigma^2, \sigma_a^2|\text{SSA}, \text{SSE}),$$

where $L(\mu|\bar{y}_{..})$ is the likelihood of μ given $\bar{y}_{..}$, namely

$$L(\mu|\bar{y}_{..}) = \frac{\exp\left[\frac{-(\bar{y}_{..} - \mu)^2}{2\lambda/mn}\right]}{(2\pi)^{\frac{1}{2}}(\lambda/mn)^{\frac{1}{2}}}, \tag{2.48}$$

and

$$L(\sigma^2, \sigma_a^2|\text{SSE}, \text{SSA}) = \frac{\exp\left[-\frac{1}{2}\left(\frac{\text{SSE}}{\sigma^2} + \frac{\text{SSA}}{\lambda}\right)\right]}{(2\pi)^{\frac{1}{2}(mn-1)}\sigma^{2[\frac{1}{2}m(n-1)]}\lambda^{\frac{1}{2}(m-1)}(mn)^{\frac{1}{2}}} \quad (2.49)$$

is the likelihood function of σ^2 and σ_a^2 given SSA and SSE. Thus, because μ is not involved in (2.49) that is the likelihood for REML.

– ii. *REML equations and their solutions*

The REML equations come from maximizing the logarithm of (2.49). Denoting this by l_R we find

$$\begin{aligned} l_R &= -\tfrac{1}{2}(mn-1)\log 2\pi - \tfrac{1}{2}\log mn - \tfrac{1}{2}m(n-1)\log\sigma^2 \\ &\quad -\tfrac{1}{2}(m-1)\log\lambda - \frac{\text{SSE}}{2\sigma^2} - \frac{\text{SSA}}{2\lambda}. \end{aligned} \quad (2.50)$$

Equating to zero the derivative of l_R with respect to σ^2 and λ gives solutions $\dot{\sigma}_R^2$ and $\dot{\lambda}_R$ as $\dot{\lambda}_R = \text{SSA}/(a-1) = \text{MSA}$ and

$$\dot{\sigma}_R^2 = \frac{\text{SSE}}{m(n-1)} = \text{MSE}; \quad \text{and} \quad \text{thus } \dot{\sigma}_{a,R}^2 = \frac{1}{n}(\text{MSA} - \text{MSE}). \quad (2.51)$$

These are the REML solutions.

– iii. *REML estimators*

Similar to the situation with ML, the preceding REML solutions are REML estimators only when both are non-negative. $\dot{\sigma}_R^2$ can never be negative, but $\dot{\sigma}_{a,R}^2$ can be, whereupon we have to maximize l_R subject to $\dot{\sigma}_{a,R}^2 = 0$, which leads to $\dot{\sigma}_R^2$ then being $\text{SST}/(mn-1)$. Thus the REML estimators are

when $\dot{\sigma}_{a,R}^2 > 0$,

$$\hat{\sigma}_R^2 = \text{MSE} \quad \text{and} \quad \hat{\sigma}_{a,R}^2 = \frac{1}{n}(\text{MSA} - \text{MSE});$$

when $\dot{\sigma}_{a,R}^2 \leq 0$,

$$\hat{\sigma}_R^2 = \frac{\text{SST}}{mn-1} \quad \text{and} \quad \hat{\sigma}_{a,R}^2 = 0. \quad (2.52)$$

– iv. *Comparison with ML*

Comparing the REML estimators of (2.52) with the ML estimators of (2.27) and (2.28), we see that the condition for a negative solution for σ_a^2 is not quite the same in the two cases. In REML it is MSA < MSE whereas in ML it is $(1 - 1/m)$MSA < MSE; and the positive estimator is similarly slightly different: (MSA – MSE)/n in REML but $[(1-1/m)$MSA - MSE$]/n$ in ML. Also, when there is a negative solution for σ_a^2, the resulting estimator of σ^2 is not the same in the two cases: SST$/(mn-1)$ in REML but SST$/mn$ in ML. Each of these differences has a common feature: that with REML we see SSA being divided by $m-1$ where it is divided by m in ML; and in REML the divisor of SST is $mn-1$ whereas it is mn in ML. In both instances the REML divisor is one less than the ML divisor. In this way REML is taking account of the degree of freedom that gets utilized in estimating μ—even though REML does not explicitly involve the estimation of μ. Nevertheless, it is a general feature of REML estimation of variance components from balanced data that degrees of freedom for fixed effects get taken into account. The simplest example is that of estimating σ^2 from a simple sample of n independent observations $x_1, x_2, \cdots, x_n$, from $\mathcal{N}(\mu, \sigma^2)$. The ML estimator is $\Sigma_i(x_i - \bar{x})^2/n$ whereas the REML estimator is $\Sigma_i(x_i - \bar{x})^2/(n-1)$.

– v. *Bias*

What has just been said about REML might lead one to surmise that REML estimators are unbiased. They are not. The same need for non-negative estimates arises as with ML estimation. Similar to (2.31) we define, for balanced data

$$\begin{aligned} p_R &= \mathrm{P}\{\dot{\sigma}^2_{a,R} < 0\} = \mathrm{P}\{\mathrm{MSA} < \mathrm{MSE}\} \\ &= \mathrm{P}\{\mathcal{F}_{m-1}^{m(n-1)} > 1 + n\sigma_a^2/\sigma_e^2\}. \end{aligned} \tag{2.53}$$

Then, based on (2.52), the expected value of $\hat{\sigma}^2_R$ is

$$\mathrm{E}[\hat{\sigma}^2_R] = (1 - p_R)\mathrm{E}[\mathrm{MSE}|\dot{\sigma}^2_{a,R} \geq 0] + p_R\mathrm{E}[\mathrm{SST}|\dot{\sigma}^2_{a,R} < 0]/(mn-1). \tag{2.54}$$

– vi. *Sampling variances*

Based on (2.50), we can easily find the large-sample dispersion matrix,

$$\text{var}\begin{bmatrix} \hat{\sigma}^2_{a,R} \\ \hat{\lambda}_R \end{bmatrix} \simeq \begin{bmatrix} -\text{E}[l_{R,\sigma^2,\sigma^2_a}] & -\text{E}[l_{R,\sigma^2,\lambda}] \\ -\text{E}[l_{R,\lambda,\sigma^2}] & -\text{E}[l_{R,\lambda,\lambda}] \end{bmatrix}^{-1},$$

which leads to exactly the same results as in (2.35) except that in the lower right-hand element the term $(\lambda^2/\sigma^4)/m$ is $(\lambda^2/\sigma^4)/(m-1)$.

b. Unbalanced data

In keeping with (2.20), the likelihood function for unbalanced data is

$$L(\mu, \sigma^2, \sigma^2_a|\mathbf{y}) = \frac{\exp\left\{-\left[\dfrac{\Sigma_i\Sigma_j(y_{ij}-\mu)^2}{2\sigma^2} - \Sigma_i\dfrac{n_i^2\sigma^2_a(\bar{y}_{i\cdot}-\mu)^2}{2\sigma^2(\sigma^2+n_i\sigma^2_a)}\right]\right\}}{(2\pi)^{\frac{1}{2}N}\sigma^{2\frac{1}{2}(N-m)]}\prod_{i=1}^{a}(\sigma^2+n_i\sigma^2_a)^{\frac{1}{2}}}.$$

There is no straightforward factoring of this likelihood that permits separating a function of μ in the manner of (2.48) for balanced data. Nevertheless, equations for REML estimators can be established—as a special case of the equations for the general case. This is left until Chapter 6.

2.4 MORE ON RANDOM EFFECTS AND NORMALITY

a. Tests and confidence intervals

– i. *For the overall mean, μ*

With balanced data ($n_i \equiv n$) we can show (E 2.8) that

$$\frac{\bar{y}_{\cdot\cdot} - \mu}{\sqrt{\text{MSA}/mn}} \sim \mathcal{T}_{m-1},$$

the t-distribution on $m-1$ degrees of freedom. A test of $H_0: \mu = \mu_0$ is then to reject H_0 when

$$\left|\frac{\bar{y}_{\cdot\cdot} - \mu_0}{\sqrt{\text{MSA}/mn}}\right| > t_{m-1,\alpha/2}$$

and a corresponding confidence interval for μ is

$$\bar{y}_{\cdot\cdot} \pm t_{m-1,\alpha/2}\sqrt{\text{MSA}/mn}.$$

– ii. ***For*** σ^2

Tests and confidence intervals for σ^2 are based on the result that

$$\frac{(N-m)s^2}{\sigma^2} \sim \chi^2_{N-m},$$

for both balanced and unbalanced data. A somewhat unusual application of this would be to form a test of a specified value of σ^2 or, more commonly, to form a confidence interval:

$$\left(\frac{\text{SSE}}{\chi^2_{N-m,1-\alpha/2}}, \frac{\text{SSE}}{\chi^2_{N-m,\alpha/2}}\right).$$

– iii. ***For*** σ_a^2

For balanced or unbalanced data a likelihood ratio test (see E 2.10) of $H_0\colon \sigma_a^2 = 0$ is to reject H_0 when $F = \text{MSA/MSE} > \mathcal{F}^{m-1}_{N-m,1-\alpha}$. For balanced data ($n_i \equiv n$) a confidence interval for σ_a^2/σ^2 is given by

$$\left(\frac{F/\mathcal{F}^{m-1}_{N-m,1-\alpha/2} - 1}{n}, \frac{F/\mathcal{F}^{m-1}_{N-m,\alpha/2} - 1}{n}\right).$$

For unbalanced data no exact intervals exist for σ_a^2/σ^2 in closed form. Approximate intervals are described in Searle et al. (1992) in Sections 3.6d–vi. Exact intervals for other functions of the variances are discussed in Khuri et al. (1998).

b. Predicting random effects

– i. ***A basic result***

Our model assumes that $a_i \sim$ i.i.d. $\mathcal{N}(0, \sigma_a^2)$ where the a_i are unknown. If we wanted to guess a value for a_i in the absence of any data or information, we could do no better than to guess the mean value of a_i, namely zero. However, suppose we have data known as having a correlation of 0.99 with a_i: then we could use that information to get a prediction of a_i better than its zero mean. The information would adjust our prediction away from that mean of zero.

The basic result for doing this is quite general, as follows. Suppose we have two random variables one of which, Y, cannot be observed but which, in particular cases, we wish to predict; and the other, X, can be

observed and which is to be used for predicting Y. Then the predictor we use is the minimum mean squared error predictor of Y, based on X; it is $\mathrm{E}[Y|X]$, the conditional mean of Y given X (see E 2.9).

Motivation for this result

$$\text{Best Predictor} = \mathrm{E}[Y|X] \tag{2.55}$$

is dealt with extensively in Chapter 8, including its derivation, its properties and applications. Here we just list results for the 1-way classification [see Searle, 1971, Sec. 2.4f–v].

– ii. *In a 1-way classification*

Returning to the linear model, we wish to predict a_i given the data. The only portion of the data relevant to a_i is the sample mean for class i, $\bar{y}_{i\cdot}$. Using the general result in Section S.1 of Appendix S for a conditional mean of a normal variate,

$$\begin{aligned}
\mathrm{E}[a_i|\mathbf{y}] &= \mathrm{E}[a_i|\bar{y}_{i\cdot}] = \mathrm{E}[a_i] + \mathrm{cov}(a_i, \bar{y}_{i\cdot})[\mathrm{var}(\bar{y}_{i\cdot})]^{-1}(\bar{y}_{i\cdot} - \mathrm{E}[\bar{y}_{i\cdot}]) \\
&= 0 + \sigma_a^2 \frac{1}{\sigma_a^2 + \sigma^2/n_i}(\bar{y}_{i\cdot} - \mu) \\
&= \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2/n_i}(\bar{y}_{i\cdot} - \mu).
\end{aligned} \tag{2.56}$$

This is the best predictor of a_i, which we denote by $\mathrm{BP}(a_i)$.

Immediately a serious problem confronts us concerning the use of (2.56). It depends on the parameters μ, σ_a^2, and σ^2 whose values are unknown. The usual solution is to replace them with estimates and get an estimated best predicted value, which we denote as $\tilde{a}_i$, i.e.,

$$\tilde{a}_i = \widehat{\mathrm{BP}}(a_i) = \hat{\mathrm{E}}[a_i|\bar{y}_{i\cdot}] = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}^2/n_i}(\bar{y}_{i\cdot} - \hat{\mu}).$$

It is instructive to compare the estimated best predictor with the estimator of α_i under the fixed effects model (2.2) using the constraint $\sum_i n_i\alpha_i = 0$. For prediction in the random model, with balanced data, we have

$$\tilde{a}_i = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}^2/n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}).$$

For estimation in the fixed model we have

$$\hat{\alpha}_i = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}).$$

Both are based on $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ but $\tilde{a}_i$ is "shrunk" compared to $\hat{\alpha}_i$. It is always smaller than $\hat{\alpha}_i$, the degree to which it is smaller depending on the relative size of $\hat{\sigma}_a^2$ and $\hat{\sigma}^2/n$. If σ_a^2 is estimated to be large with respect to the estimate of σ^2/n (either $\hat{\sigma}_a^2$ is large compared to $\hat{\sigma}^2$ and/or the sample size per class is large) then the two values for the class effect are similar. This corresponds to the situation where there is a lot of variation (relatively speaking) between classes and not much is to be gained by assuming that the effects are selected from a common distribution. On the other hand, when σ_a^2 is estimated to be small with respect to σ^2/n, the shrinkage can be extensive and the two values can differ greatly.

2.5 BERNOULLI DATA: FIXED EFFECTS

We return to the ideas of Section 2.1 wherein y_{ij} is distributed independently, and $\mathrm{E}[y_{ij}] = \mu_i$, but use a distribution to accommodate binary data. Hence the only possible distribution is the Bernoulli.

a. Model equation

As previously, we consider m classes indexed by i, with the ith class having n_i observations. A model for the responses y_{ij} which are coded as 1 or 0 would be

$$\begin{aligned} \mathrm{E}[y_{ij}] &= \pi_i, \\ y_{ij} &\sim \text{indep. Bernoulli}(\pi_i). \end{aligned} \tag{2.57}$$

The more usual notation for the mean of a Bernoulli distribution is p_i which we reserve for Section 2.6 where p_i is random; and here we use π_i for the fixed effects case – all this being in accord with our convention of Greek letters for fixed effects and Roman letters for random effects.

b. Likelihood

The likelihood for the data is

$$\begin{aligned} L &= \prod_i \prod_j \pi_i^{y_{ij}} (1-\pi_i)^{1-y_{ij}} = \prod_i \pi_i^{y_{i\cdot}} (1-\pi_i)^{n_i - y_{i\cdot}} \\ &= \prod_i \left(\frac{\pi_i}{1-\pi_i}\right)^{y_{i\cdot}} (1-\pi_i)^{n_i}. \end{aligned} \tag{2.58}$$

Therefore

$$l = \log L = \sum_i \{y_{i\cdot} \log[\pi_i/(1-\pi_i)] + n_i \log[1-\pi_i]\}. \tag{2.59}$$

c. ML equations and their solutions

The ML equations come from differentiating l of (2.59) with respect to the π_i to obtain

$$\begin{aligned} \frac{\partial l}{\partial \pi_k} &= y_{k\cdot}\left(\frac{1}{\pi_k} + \frac{1}{1-\pi_k}\right) - \frac{n_k}{1-\pi_k} \\ &= \frac{y_{k\cdot}}{\pi_k}\left(\frac{1}{1-\pi_k}\right) - \frac{n_k}{1-\pi_k}. \end{aligned} \tag{2.60}$$

Setting this equal to zero gives

$$\hat{\pi}_k = \bar{y}_{k\cdot} = \text{sample proportion of 1s in class } k.$$

d. Likelihood ratio test

The hypothesis H_0: μ_i *all equal* is tested using the likelihood ratio statistic

$$\Lambda = \frac{\prod_i (\bar{y}_{\cdot\cdot})^{y_{i\cdot}} (1-\bar{y}_{\cdot\cdot})^{n_i - y_{i\cdot}}}{\prod_i (\bar{y}_{i\cdot})^{y_{i\cdot}} (1-\bar{y}_{i\cdot})^{n_i - y_{i\cdot}}}$$

giving

$$\begin{aligned} \log \Lambda &= \sum_i \left[y_{i\cdot} \log\left(\frac{\bar{y}_{\cdot\cdot}}{\bar{y}_{i\cdot}}\right) + (n_i - y_{i\cdot}) \log\left(\frac{1-\bar{y}_{\cdot\cdot}}{1-\bar{y}_{i\cdot}}\right)\right] \\ &= \sum_i \left[n_i \hat{\pi}_i \log\left(\frac{\hat{\pi}}{\hat{\pi}_i}\right) + n_i(1-\hat{\pi}_i) \log\left(\frac{1-\hat{\pi}}{1-\hat{\pi}_i}\right)\right], \end{aligned} \tag{2.61}$$

where $\hat{\pi} = \bar{y}_{\cdot\cdot}$ = overall sample proportions of 1s.

The large–sample test is given by

$$\text{Reject } H_0 \text{ if } -2\log\Lambda \geq \chi^2_{m-1,1-\alpha}. \tag{2.62}$$

e. The usual chi-square test

A test used more commonly in practice than (2.62) is the chi-square test of independence, or (equivalently) the chi-square test for equality of binomial proportions (Snedecor and Cochran, 1989, Sec. 11.7). It is

Table 2.1: Successes and Failures in a 1-Way Classification

	Classification Level						
Outcome	1	2	$\cdots$	i	$\cdots$	m	Total
Success	$y_{1\cdot}$	$y_{2\cdot}$	$\cdots$	$y_{i\cdot}$	$\cdots$	$y_{m\cdot}$	$y_{\cdot\cdot}$
Failure	$n_1 - y_{1\cdot}$	$n_2 - y_{2\cdot}$	$\cdots$	$n_i - y_{i\cdot}$	$\cdots$	$n_m - y_{m\cdot}$	$N - y_{\cdot\cdot}$
Total	n_1	n_2	$\cdots$	n_i	$\cdots$	n_m	N

best described by starting with Table 2.1. The usual chi-square test is to reject H_0 when

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{m-1,1-\alpha},$$

where O_{ij} and E_{ij} are, respectively, observed and expected values. For Table 2.1, there are but two values for j, 1 and 2, and

$$O_{i1} = y_{i\cdot} \quad \text{and} \quad O_{i2} = n_i - y_{i\cdot}$$
$$E_{i1} = n_i\bar{y}_{\cdot\cdot} \quad \text{and} \quad E_{i2} = n_i(1 - \bar{y}_{\cdot\cdot}),$$

and hence

$$\begin{aligned} \chi^2 &= \sum_i \left[\frac{(y_{i\cdot} - n_i\bar{y}_{\cdot\cdot})^2}{n_i\bar{y}_{\cdot\cdot}} + \frac{(n_i - y_{i\cdot} - (n_i - n_i\bar{y}_{\cdot\cdot}))^2}{n_i - n_i\bar{y}_{\cdot\cdot}} \right] \\ &= \sum_i \left[\frac{n_i(\hat{\pi}_i - \hat{\pi})^2}{\hat{\pi}} + \frac{n_i(-\hat{\pi}_i + \hat{\pi})^2}{1 - \hat{\pi}} \right] \qquad (2.63) \\ &= \sum_i \frac{n_i(\hat{\pi}_i - \hat{\pi})^2}{\hat{\pi}(1 - \hat{\pi})}. \end{aligned}$$

An interesting question is: How does χ^2 of (2.63) compare to $-2\log\Lambda$ of (2.62)? To answer this we use a Taylor series expansion of $f(x) = x\log(c/x)$ about c:

$$f(x) \doteq f(c) + f'(c)(x - c) + \tfrac{1}{2}f''(c)(x - c)^2,$$

where

$$f'(c) = \frac{\partial}{\partial x} f(x)\bigg|_{x=c} \quad \text{and} \quad f''(c) = \frac{\partial^2}{\partial x^2} f(x)\bigg|_{x=c}.$$

This gives

$$f(x) \doteq -(x-c) - \tfrac{1}{2}(x-c)^2/c.$$

Applying this to $\log \Lambda$, wherein each term is of the form $x \log(c/x)$, gives

$$\log \Lambda \doteq -\sum_i \left[n_i(\hat{\pi}_i - \hat{\pi}) + \frac{n_i}{2} \frac{(\hat{\pi}_i - \hat{\pi})^2}{\hat{\pi}} + n_i[1 - \hat{\pi}_i - (1-\hat{\pi})] + \frac{n_i}{2} \frac{[1 - \hat{\pi}_i - (1-\hat{\pi})]^2}{1-\hat{\pi}} \right] \tag{2.64}$$

or

$$\begin{aligned} -2 \log \Lambda &\doteq \sum_i \left[\frac{n_i(\hat{\pi}_i - \hat{\pi})^2}{\hat{\pi}} + \frac{n_i(-\hat{\pi}_i + \hat{\pi})^2}{1-\hat{\pi}} \right] \\ &= \sum_i \frac{n_i(\hat{\pi}_i - \hat{\pi})^2}{\hat{\pi}(1-\hat{\pi})} \\ &= \chi^2 \end{aligned}$$

of (2.63). Thus, when $\hat{\pi}_i$ is not too far from $\hat{\pi}$, the two statistics give similar results.

f. Large-sample tests and intervals

Large-sample tests for testing $H_0 : \pi_i = \pi_{i0}$, where π_{i0} is a specified value, can be based on

$$z = \frac{\hat{\pi}_i - \pi_{i0}}{\sqrt{\frac{\pi_{i0}(1-\pi_{i0})}{n_i}}} \sim \mathcal{AN}(0,1), \tag{2.65}$$

where $\mathcal{AN}$ means asymptotically normally distributed.

The test statistic for the hypothesis $H_0 : \pi_i - \pi_k = \pi_{i0} - \pi_{k0}$ would be

$$z = \frac{\hat{\pi}_i - \hat{\pi}_k - (\pi_{i0} - \pi_{k0})}{\sqrt{\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i} + \frac{\hat{\pi}_k(1-\hat{\pi}_k)}{n_k}}} \sim \mathcal{AN}(0,1),$$

where $\pi_{i0} - \pi_{k0}$ is the hypothesized difference under H_0. For example, to test $H_0 : \pi_i = \pi_k$ versus the alternative $H_1 : \pi_i \neq \pi_k$, we set $\pi_{i0} - \pi_{k0}$ equal to zero and reject H_0 if

$$\left| \frac{\hat{\pi}_i - \hat{\pi}_k}{\sqrt{\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i} + \frac{\hat{\pi}_k(1-\hat{\pi}_k)}{n_k}}} \right| > z_{\alpha/2}, \tag{2.66}$$

where z_α is the $100\alpha\%$ percentile of the standard normal distribution, i.e., if $Z \sim \mathcal{N}(0,1)$, then $P\{Z > z_\alpha\} = \alpha$. Alternatively, we could perform a likelihood ratio test or a χ^2 test using only the data in columns i and k of Table 2.1. These two tests are quite similar (see E 2.7). Note that in (2.66) we use the estimated values of π_i and π_k rather than the values under H_0 as we did in (2.65). This is because hypothesizing a difference between π_i and π_k under H_0 does not tell us the actual values of π_i and π_k under H_0.

Large-sample confidence intervals for the π_i are given by

$$\hat{\pi}_i \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}},$$

again using (2.65). The corresponding confidence intervals for $\pi_i - \pi_k$ are given by

$$\hat{\pi}_i - \hat{\pi}_k \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i} + \frac{\hat{\pi}_k(1-\hat{\pi}_k)}{n_k}}.$$

g. Exact tests and confidence intervals

The likelihood ratio and χ^2 tests of $H_0: \pi_1 = \pi_2 = \cdots = \pi_m$ and the normality-based confidence intervals and tests of the preceding section are based on large-sample distributional approximations which can be inaccurate for small samples. The usual rule of thumb (Snedecor and Cochran, 1989, p. 127) is that the approximation is accurate when most of the "expected values" of Table 2.1 i.e., $n_i\bar{y}_{\cdot\cdot}$ and $n_i(1-\bar{y}_{\cdot\cdot})$ are greater than five and can give inaccurate results when expected values are less than one.

Exact tests can be based on the conditional distribution of the table entries given the marginal totals. Under $H_0: \pi_i$ *all equal* the conditional probability of a sample is given by

$$\prod_{i=1}^{m} \binom{n_i}{y_{i\cdot}} \Big/ \binom{N}{y_{\cdot\cdot}}. \tag{2.67}$$

A p-value can be calculated via the usual definition: the probability, under H_0, of a result as extreme or more extreme than that observed. This would be done by summing (2.67) over all the possible data configurations (as in Table 2.1) which are "more extreme" than the observed table.

For two populations ($m = 2$) we have a 2×2 table and it is straightforward to designate *more extreme* tables. That is, once a single entry in the 2×2 table is known, and conditional on the margins, all the remaining entries are fixed. We can thus enumerate the more extreme tables by varying this single entry from the observed table in a direction "away" from H_0. This test is known as *Fisher's exact test*.

For $m > 2$ populations we must choose a definition of "more extreme." For example, a common choice is whether a table of possible data gives a larger value of the χ^2 statistic than that given by the observed table. The problem is then a computational one since the number of possible tables with given margins gets unmanageably large. Special software (e.g., Mehta and Patel, 1992) is usually required.

Exact confidence intervals for π_i can be calculated (Mood et al., 1974, p. 393) as (π_{iL}, π_{iU}) where the π_{iL} and π_{iU} solve

$$\sum_{k=y_{i\cdot}}^{n_i} \binom{n_i}{k} \pi_{iL}^k (1-\pi_{iL})^{n_i-k} = \alpha/2 \quad \text{and}$$

$$\sum_{k=1}^{y_{i\cdot}} \binom{n_i}{k} \pi_{iU}^k (1-\pi_{iU})^{n_i-k} = \alpha/2.$$

These are known as the Clopper–Pearson intervals and can be somewhat conservative. Blyth and Still (1983) give less conservative intervals in tabular form and accurate approximate intervals. Santner and Snell (1980) give intervals for $\pi_i - \pi_k$ and π_i/π_k.

h. Example: Snake strike data

An experiment was conducted at Cornell University to find factors that determine whether a snake would strike at a target or fail to do so. Snakes were placed in a cage with a target that looked something like an artificial mouse and a binary response was recorded as to whether the snake struck at the target within five minutes. A concern was that some snakes would always strike at targets, whereas others would not strike at all, obscuring any effect due to target differences. Table 2.2 show data for six snakes.

We are interested in testing homogeneity across the snakes. The likelihood ratio test of (2.62) gives a statistic of 9.10 to be compared to a χ^2 distribution with five degrees of freedom. The asymptotic p-value of this test is approximately 0.10.

Table 2.2: Number of Occurrences of Strike or No Strike for Each Snake

	Snake						
Outcome	1	2	3	4	5	6	Total
Strike	2	2	3	0	1	1	9
No strike	2	2	0	2	0	0	6
Total	4	4	3	2	1	1	15

The chi-square statistic of Section 2.5e is equal to 6.67, again to be compared to a χ^2 distribution with five degrees of freedom, giving an asymptotic p-value of 0.25. Exact calculations using the conditional distribution (2.67) give a p-value for the χ^2 statistic of 0.27 and 0.24 for the likelihood ratio statistic.

2.6 BERNOULLI DATA: RANDOM EFFECTS

a. Model equation

The analog of the random effects model for normally distributed data, (2.13), would be

$$\begin{aligned} \mathrm{E}[y_{ij}] &= p_i, \\ y_{ij} &\sim \text{indep. Bernoulli}(p_i), \\ p_i &\sim \text{i.i.d. } G, \end{aligned} \tag{2.68}$$

where G is a distribution for the p_i and we maintain our convention of using Roman letters for random effects. Normality cannot be assumed for the p_i since they are probabilities and are restricted to the interval (0,1).

b. Beta-binomial model

A logical choice for G is the beta distribution since it is a flexible distribution on (0,1) and leads to mathematically tractable results. If p_i from (2.68) follows a beta distribution with parameters α and β, its density is given by

$$p_i^{\alpha-1}(1-p_i)^{\beta-1}/B(\alpha,\beta), \tag{2.69}$$

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx \tag{2.70}$$

is the beta function. It then follows that

$$\begin{aligned} \mathrm{E}[p_i] &= \frac{\alpha}{\alpha+\beta} \quad \text{and} \\ \mathrm{var}(p_i) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned} \tag{2.71}$$

Model (2.68) along with (2.69) we call the *beta–binomial model*:

$$\begin{aligned} \mathrm{E}[y_{ij}|p_i] &= p_i \\ y_{ij}|p_i &\sim \text{indep. Bernoulli}(p_i) \\ p_i &\sim \text{i.i.d. beta}(\alpha,\beta). \end{aligned} \tag{2.72}$$

– i. *Means, variances, and covariances*

It is straightforward to calculate moments of the y_{ij} under model (2.72). We have

$$\mathrm{E}[y_{ij}] = \mathrm{E}\left[\mathrm{E}[y_{ij}|p_i]\right] = \mathrm{E}[p_i] = \frac{\alpha}{\alpha+\beta} \tag{2.73}$$

$$\begin{aligned} \mathrm{var}(y_{ij}) &= \mathrm{E}[\mathrm{var}(y_{ij})|p_i] + \mathrm{var}(\mathrm{E}[y_{ij}|p_i]) \\ &= \mathrm{E}[p_i(1-p_i)] + \mathrm{var}(p_i) \\ &= \mathrm{E}[p_i] - \mathrm{E}[p_i^2] + \mathrm{E}[p_i^2] - (\mathrm{E}[p_i])^2 \\ &= \mathrm{E}[p_i](1-\mathrm{E}[p_i]) \\ &= \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta} = \frac{\alpha\beta}{(\alpha+\beta)^2}. \end{aligned} \tag{2.74}$$

The result (2.74) also reflects the fact that being binary forces y_{ij} to have a marginal Bernoulli distribution with variance equal to the mean times 1 minus the mean.

We can calculate a covariance similarly:

$$\begin{aligned} \mathrm{cov}(y_{ij}, y_{il}) &= \mathrm{cov}\left(\mathrm{E}[y_{ij}|p_i], \mathrm{E}[y_{il}|p_i]\right) + \mathrm{E}\left[\mathrm{cov}(y_{ij}, y_{il}|p_i)\right] \\ &= \mathrm{cov}(p_i, p_i) + 0 \\ &= \mathrm{var}(p_i) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad \text{for } j \neq l. \end{aligned} \tag{2.75}$$

The covariance between y_{ij} and y_{kl} is zero for $i \neq k$. Thus we have an intraclass correlation of

$$\begin{aligned} \rho &= \text{corr}(y_{ij}, y_{il}) = \frac{\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]}{\alpha\beta/(\alpha+\beta)^2} \\ &= \frac{1}{\alpha+\beta+1} \end{aligned} \tag{2.76}$$

where $\text{corr}(\cdot,\cdot)$ denotes correlation. Some authors (e.g., Williams, 1975; Griffiths, 1973) suggest a reparameterization of (2.69) in terms of the mean $\mu = \alpha/(\alpha+\beta)$ and the intraclass correlation, ρ, or a related quantity, $\tau = 1/(\alpha+\beta)$. Reparameterized in such a way, the mean is, of course, μ, and the covariance is $\mu(1-\mu)\rho$ or $\mu(1-\mu)\tau/(\tau+1)$.

– ii. *Overdispersion*

If y_{ij} $(j = 1, 2, \ldots, n_i)$ were independent Bernoulli random variables with mean μ then $y_{i\cdot}$ would follow a binomial(n_i, μ) distribution with variance $n_i\mu(1-\mu)$. Under the beta-binomial model

$$\text{var}(y_{i\cdot}) = \sum_j \text{var}(y_{ij}) + \sum_{j\neq l} \text{cov}(y_{ij}, y_{il}) \tag{2.77}$$

$$= n_i\mu(1-\mu) + \binom{n_i}{2}\mu(1-\mu)\rho$$

$$= n_i\mu(1-\mu)\left[1 + \frac{n_i-1}{2}\rho\right]. \tag{2.78}$$

As long as $\rho > 0$, which is required by the beta-binomial model ($\alpha > 0, \beta > 0$), the variance will be larger than the binomial variance. This is often termed *overdispersion.*

Examination of the preceding expressions for $\text{var}(y_{i\cdot})$ reveals that no detail from the beta-binomial model is used. The only assumption made is that the variances and covariances are the same for all j and l. Thus, overdispersion can arise in a variety of contexts with non-independent data.

– iii. *Likelihood*

The likelihood is given by

$$\begin{aligned} L &= \prod_{i=1}^{m} \int_0^1 \prod_{j=1}^{n_i} p_i^{y_{ij}} (1-p_i)^{1-y_{ij}} g(p_i;\alpha,\beta)\, dp_i \\ &= \prod_{i=1}^{m} \frac{1}{B(\alpha,\beta)} \int_0^1 p_i^{y_{i\cdot}} (1-p_i)^{n_i-y_{i\cdot}} p_i^{\alpha-1} (1-p_i)^{\beta-1}\, dp_i \\ &= \prod_{i=1}^{m} \frac{B(\alpha+y_{i\cdot},\beta+n_i-y_{i\cdot})}{B(\alpha,\beta)}, \end{aligned} \tag{2.79}$$

the last equality coming about from the definition of the beta function (2.70).

Using the results that $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ and that

$$\Gamma(s+h)/\Gamma(s) = (s+h-1)(s+h-2)\cdots(s+1)s = \prod_{t=0}^{h-1}(s+t),$$

we have

$$\begin{aligned} l &= \log L \\ &= \sum_{i=1}^{m} \left[\log \frac{\Gamma(\alpha+y_{i\cdot})}{\Gamma(\alpha)} + \log \frac{\Gamma(\beta+n_i-y_{i\cdot})}{\Gamma(\beta)} - \log \frac{\Gamma(\alpha+\beta+n_i)}{\Gamma(\alpha+\beta)} \right] \\ &= \sum_{i=1}^{m} \left[\sum_{h=0}^{y_{i\cdot}-1} \log(\alpha+h) + \sum_{h=0}^{n_i-y_{i\cdot}-1} \log(\beta+h) - \sum_{h=0}^{n_i-1} \log(\alpha+\beta+h) \right]. \end{aligned} \tag{2.80}$$

In (2.80), and (2.81) shown below, any sum with an upper limit of -1 is interpreted as zero. Under the parameterization $\tau = 1/(\alpha+\beta)$, equation (2.80) takes the form (see E 2.13)

$$l = \sum_{i=1}^{m} \left[\sum_{h=0}^{y_{i\cdot}-1} \log(\mu+h\tau) + \sum_{h=0}^{n_i-y_{i\cdot}-1} \log(1-\mu+h\tau) - \sum_{h=0}^{n_i-1} \log(1+h\tau) \right]. \tag{2.81}$$

– iv. *ML estimation*

Closed-form maximizing values of l do not exist, so numerical maximization must be used to find maximum likelihood estimates for any given data set.

– v. *Large-sample variances*

ML estimators based on (2.80) or the reparameterized version (2.81) are asymptotically normally distributed with means equal to the true values and variances given by the inverse of the information matrix. In particular for (2.80)

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{AN}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \Sigma_{\hat{\alpha},\hat{\beta}} = \begin{pmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{21} & \sigma^{22} \end{pmatrix}^{-1} \right), \tag{2.82}$$

where

$$\begin{aligned} \sigma^{11} &= -\mathrm{E}\left[\frac{\partial^2 l}{\partial \alpha^2}\right] \\ &= \sum_{i=1}^{m}\left[\sum_{k=0}^{n_i} \mathrm{P}\{y_{i\cdot} = k\}\left(\sum_{h=0}^{k-1}\frac{1}{(\alpha+h)^2}\right) + \sum_{h=0}^{n_i-1}\frac{1}{(\alpha+\beta+h)^2}\right] \\ \sigma^{12} &= -\mathrm{E}\left[\frac{\partial^2 l}{\partial \alpha\, \partial \beta}\right] = \sigma^{21} \\ &= \sum_{i=1}^{m}\sum_{h=0}^{n_i-1}\frac{1}{(\alpha+\beta+h)^2} \\ \sigma^{22} &= -\mathrm{E}\left[\frac{\partial^2 l}{\partial \beta^2}\right] \\ &= \sum_{i=1}^{m}\left[\sum_{k=0}^{n_i} \mathrm{P}\{y_{i\cdot} = k\}\left(\sum_{h=0}^{n_i-k-1}\frac{1}{(\beta+h)^2}\right) + \sum_{h=0}^{n_i-1}\frac{1}{(\alpha+\beta+h)^2}\right], \end{aligned}$$

and $P\{y_{i\cdot} = k\} = \binom{n_i}{k} B(\alpha+k, n_i+\beta-k)/B(\alpha,\beta)$. As noted below (2.80) any sum with an upper limit of -1 is set equal to zero.

For the μ and τ parameterization,

$$\begin{pmatrix} \hat{\mu} \\ \hat{\tau} \end{pmatrix} \sim \mathcal{AN}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, \Sigma_{\hat{\mu},\hat{\tau}} = \begin{pmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{21} & \sigma^{22} \end{pmatrix}^{-1} \right), \tag{2.83}$$

where

$$\sigma^{11} = \sum_{i=1}^{m}\left[\sum_{k=0}^{n_i} \mathrm{P}\{y_{i\cdot} = k\}\left(\sum_{h=0}^{k-1}\frac{1}{(\mu+h\tau)^2} + \sum_{h=0}^{n_i-k-1}\frac{1}{(1-\mu+h\tau)^2}\right)\right]$$

$$\sigma^{12} = \sum_{i=1}^{m}\left[\sum_{k=0}^{n_i} P\{y_{i\cdot}=k\}\left(\sum_{h=0}^{k-1}\frac{h}{(\mu+h\tau)^2}+\sum_{h=0}^{n_i-k-1}\frac{h}{(1-\mu+h\tau)^2}\right)\right]$$

$$\sigma^{21} = \sigma^{12}$$

$$\sigma^{22} = \sum_{i=1}^{m}\left[\sum_{k=0}^{n_i} P\{y_{i\cdot}=k\}\left(\sum_{h=0}^{k-1}\frac{h^2}{(\mu+h\tau)^2}+\sum_{h=0}^{n_i-k-1}\frac{h^2}{(1-\mu+h\tau)^2}\right),\right.$$
$$\left.+\sum_{h=0}^{n_i-1}\frac{h^2}{(1+h\tau)^2}\right]$$

and

$$P\{y_{i\cdot}=k\} = \binom{n_i}{k} B(\frac{\mu}{\tau}+k, n_i+\frac{1-\mu}{\tau}-k)/B\left(\frac{\mu}{\tau},\frac{1-\mu}{\tau}\right).$$

– vi. *Large-sample tests and intervals*

Large-sample inferences concerning μ can be based on the asymptotic distribution (2.82) or (2.83). Let $\text{var}(\hat{\mu})$ denote the (1,1) entry of $\Sigma_{\hat{\mu},\hat{\tau}}$. Then a large-sample confidence interval would be

$$\hat{\mu} \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\mu})}, \tag{2.84}$$

where $\widehat{\text{var}}(\cdot)$ indicates that the estimated values of μ and τ have been substituted in the variance formula. Large-sample tests could be based on

$$z = \frac{\hat{\mu}-\mu}{\sqrt{\widehat{\text{var}}(\hat{\mu})}} \sim \mathcal{AN}(0,1). \tag{2.85}$$

Alternatively, tests and confidence regions can be based on the likelihood ratio statistic. To test $H_0: \mu = \mu_0$, we calculate $\hat{\tau}(\mu_0)$ (i.e., the value of τ which maximizes the likelihood when μ is fixed at μ_0). The large-sample test is then to reject H_0 if

$$-2\{l[\mu_0,\hat{\tau}(\mu_0)] - l[\hat{\mu},\hat{\tau}]\} > \chi^2_{1,1-\alpha}, \tag{2.86}$$

where $l(\mu,\tau) = \log L(\mu,\tau)$ is a function of μ and τ.

A confidence interval for μ which corresponds to the test (2.86) is the set of values μ^* given by

$$\left\{\mu^*: -2\{l[\mu^*,\hat{\tau}(\mu^*)] - l[\hat{\mu},\hat{\tau}]\} < \chi^2_{1,1-\alpha}\right\} \tag{2.87}$$

i.e., the set of values of μ^* for which we can accept the $H_0: \mu = \mu^*$. This set must be calculated numerically.

Large-sample inferences for τ must be handled carefully. The usual hypothesis of interest is $H_0: \tau = 0$, which, as long as μ is not zero or one, is equivalent to no correlation, or equivalently, no variation in the p_i across the one-way classification. When $\tau = 0$ and for large samples, the maximum likelihood estimator is exactly zero half the time (for a similar situation see E 2.18). It thus does *not* have a large-sample normal distribution and the usual large-sample theory for $-2 \log \Lambda$ fails.

In this simple case an easy modification is available: Calculate the usual likelihood ratio statistic and make a simple adjustment to the critical value. Under $H_0: \tau = 0$ the maximum likelihood estimator of μ is $\bar{y}_{\cdot\cdot}$. The test is thus to reject H_0 when

$$-2\left[l(\bar{y}_{\cdot\cdot}, 0) - l(\hat{\mu}, \hat{\tau})\right] > \chi^2_{1,1-2\alpha}. \tag{2.88}$$

Roughly speaking, this can be thought of as adjusting a test statistic appropriate for a two-sided test (the likelihood ratio test) in order to test a one-sided hypothesis ($H_0: \tau = 0$ versus $H_1: \tau > 0$).

In the less usual case when a specified value of $\tau > 0$ is of interest, the large-sample distribution of $\hat{\tau}$ *is* asymptotically normal. Tests and confidence intervals can then be based on the standard normal distribution just as with μ in (2.85) or on the likelihood ratio test with the usual critical point, $\chi^2_{1,1-\alpha}$.

– vii. *Prediction*

As before, we wish to calculate the best predicted values as given by

$$\mathrm{BP}(p_i) = \mathrm{E}[p_i|y] = \mathrm{E}[p_i|y_{i\cdot}]. \tag{2.89}$$

Under the beta-binomial model, (2.72), it is straightforward to show (E 2.14) that the conditional distribution of p_i given $y_{i\cdot}$ is beta with parameters $\alpha + y_{i\cdot}$ and $\beta + n_i - y_{i\cdot}$. The conditional mean is therefore given by

$$\mathrm{E}[p_i|y_{i\cdot}] = \frac{\alpha + y_{i\cdot}}{\alpha + \beta + n_i} = \frac{\mu/\tau + y_{i\cdot}}{1/\tau + n_i} \tag{2.90}$$

and the estimated best predictor, $\tilde{p}_i$, is given by

$$\tilde{p}_i = \frac{\hat{\mu}/\hat{\tau} + y_{i\cdot}}{1/\hat{\tau} + n_i}$$

$$= \hat{\mu}\left(\frac{1/\hat{\tau}}{1/\hat{\tau}+n_i}\right) + \bar{y}_{i\cdot}\left(\frac{n_i}{1/\hat{\tau}+n_i}\right)$$

$$= \hat{\mu}\left(\frac{1/\hat{\tau}}{1/\hat{\tau}+n_i}\right) + \hat{\pi}_i\left(\frac{n_i}{1/\hat{\tau}+n_i}\right). \tag{2.91}$$

As with the normal random effects model, the estimated best predictor, $\tilde{p}_i$, is a weighted average of an overall estimate, $\hat{\mu}$, and the fixed effects estimate, $\hat{\pi}_i$. It is also a shrinkage estimator, with the individual predicted values, $\tilde{p}_i$, being closer to the overall estimate, $\hat{\mu}$, than are the $\hat{\pi}_i$.

c. Logit-normal model

The reason a normal distribution cannot be assumed for p_i in (2.68) is that it is restricted to the interval (0,1). An alternative approach is to transform p_i using $\text{logit}(p_i) \equiv \log[p_i/(1-p_i)]$. The range for $\text{logit}(p_i)$ is $(-\infty, \infty)$ as p_i ranges from zero to one and a normal distribution can be assumed for $\text{logit}(p_i)$. This gives the model

$$\begin{aligned} \mathrm{E}[y_{ij}|p_i] &= p_i \\ y_{ij}|p_i &\sim \text{indep. Bernoulli}(p_i) \\ l_i = \text{logit}(p_i) &\sim \text{i.i.d. } \mathcal{N}(\mu, \sigma^2). \end{aligned} \tag{2.92}$$

– i. *Likelihood*

The likelihood for this model, similar to that derived in (2.79), is

$$L = \prod_{i=1}^{m} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} p_i^{y_{ij}}(1-p_i)^{1-y_{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(l_i-\mu)^2} dl_i. \tag{2.93}$$

This can be written in a slightly simpler fashion as

$$L = \prod_{i=1}^{m} \int_{-\infty}^{\infty} \left(\frac{p_i}{1-p_i}\right)^{y_{i\cdot}} (1-p_i)^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(l_i-\mu)^2} dl_i. \tag{2.94}$$

With a change of variables of $z_i = (l_i - \mu)/\sigma$ this becomes

$$L = \prod_{i=1}^{m} \int_{-\infty}^{\infty} \frac{e^{(\mu+\sigma z_i)y_{i\cdot}}}{(1+e^{\mu+\sigma z_i})^{n_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz_i. \tag{2.95}$$

Unfortunately, neither L nor $l = \log L$ can be appreciably simplified and both calculation and maximization of l must be performed using numerical methods.

– ii. *Calculation of the likelihood*

Changing variables again, using $\frac{1}{2}z_i^2 = v_i^2$ in (2.95) allows the log likelihood to be written as

$$l = \sum_{i=1}^{m} \log \int_{-\infty}^{\infty} \frac{e^{(\mu+\sqrt{2}\sigma v_i)y_{i\cdot}}}{\left(1+e^{\mu+\sqrt{2}\sigma v_i}\right)} \frac{1}{\sqrt{\pi}} e^{-v_i^2}\, dv_i. \tag{2.96}$$

In this form, each integral in l can be evaluated using Gauss–Hermite quadrature wherein

$$\int_{-\infty}^{\infty} g(x)dx \doteq \sum_{k=-r}^{r} w_k g(x_k), \tag{2.97}$$

where r is the order of integration, the w_k are weights, and the x_k are evaluation points. Generally, using large values of r increases the computation time and increases accuracy. Values of w_k and x_k are given in references on numerical integration, e.g., (Abramowitz and Stegun, 1964, Table 25.10). Using this approximation,

$$l \doteq \sum_{i=1}^{m} \log \left(\sum_{k=-r}^{r} w_k \frac{e^{(\mu+\sqrt{2}\sigma x_k)y_{i\cdot}}}{\sqrt{\pi}(1+e^{\mu+\sqrt{2}\sigma x_k})} \right). \tag{2.98}$$

For speed of computation, values of r as small as 2 or 3 have been recommended in practice (Goldstein, 1986; Hedeker and Gibbons, 1994), but this can lead to inaccurate results. If μ is not near zero, the integral can be difficult to evaluate accurately (Liu and Pierce, 1994). Values of r of 10 or greater often give good accuracy.

– iii. *Means, variances, and covariances*

Under model (2.92) we calculate the mean by writing $p_i = 1/(1+e^{-l_i})$ with $l_i \sim \mathcal{N}(\mu, \sigma^2)$. Therefore,

$$\begin{aligned} \mathrm{E}[y_{ij}] &= \mathrm{E}\left[\mathrm{E}[y_{ij}|p_i]\right] = \mathrm{E}[p_i] \\ &= \mathrm{E}\left[\frac{1}{1+e^{-l_i}}\right] = \int_{-\infty}^{\infty} \frac{1}{1+e^{-l_i}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(l_i-\mu)^2}\, dl_i \\ &= \int_{-\infty}^{\infty} \frac{1}{1+e^{-(\mu+\sigma z_i)}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2}\, dz_i. \end{aligned} \tag{2.99}$$

Again, this cannot be evaluated in closed form, but can be approximated as before using Gauss–Hermite quadrature.

Since y_{ij} is binary, it has a marginal Bernoulli distribution with mean, $\mathrm{E}[y_{ij}]$, given by (2.99). Its variance is therefore $\mathrm{E}[y_{ij}]\,(1 - \mathrm{E}[y_{ij}])$.

From first principles, the covariance of two observations in the same level of the one-way classification is $\mathrm{cov}(y_{ij}, y_{il}) = \mathrm{E}[y_{ij}y_{il}] - \mathrm{E}[y_{ij}]\mathrm{E}[y_{il}]$. The second part of this can be evaluated using (2.99) and the first part calculated as

$$\mathrm{E}[y_{ij}y_{il}] = \int_{-\infty}^{\infty} \left(\frac{1}{1 + e^{-(\mu+\sigma z_i)}}\right)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2}\, dz_i. \tag{2.100}$$

How does σ relate to the correlation between observations in the same level of the classification? Table 2.3 gives values of the correlation for several values of μ and σ.

Table 2.3: Correlations for the Logit-normal Model

	μ						
σ	-2	-1	-0.5	0	0.5	1	2
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.118	0.158	0.169	0.174	0.169	0.158	0.118
3	0.521	0.536	0.539	0.541	0.539	0.536	0.521
5	0.694	0.698	0.699	0.699	0.699	0.698	0.694

– iv. *Large-sample tests and intervals*

As in Section 2.5f, large-sample inferences concerning μ can be based on the large-sample normal distribution of $\hat{\mu}$ or the asymptotic chi-square distribution of $-2\log\Lambda$. Derivatives for calculating the observed information matrix must be calculated numerically (Hedeker and Gibbons, 1994), which makes dealing with $-2\log\Lambda$ more attractive.

A large-sample test of $H_0\colon \sigma^2 = 0$ is made by rejecting H_0 if

$$-2\log\Lambda = -2\left\{l[\hat{\mu}(\sigma^2 = 0), 0] - l[\hat{\mu}, \hat{\sigma}^2]\right\} > \chi^2_{1,1-2\alpha}. \tag{2.101}$$

A disadvantage of this approach is that the MLEs must be calculated under the random effects model, which is computationally difficult. An alternative is to consider score tests as given by, for example, Commenges et al. (1994). Scores tests use as their statistic the derivative of the log likelihood evaluated under the null hypothesis (Cox and Hinkley, 1974, p. 315), which does not require estimation under the model with random effects. For the simple case of model (2.92), the test reduces to the usual Pearson chi-square test of Section 2.5e (see E 2.16).

– v. *Prediction*

For prediction we want $\mathrm{E}[p_i|y_{ij}]$ for which the conditional distribution of p_i given $y_{i\cdot}$ is required. Again it is more convenient to consider $p_i = 1/(1+e^{-(\mu+\sigma z_i)})$ where $z_i \sim \mathcal{N}(0,1)$. We thus need

$$f_{z_i|y_{i\cdot}}(z|y) = \frac{f_{y_{i\cdot}|z_i}(y|z)f_{z_i}(z)}{\int f_{y_{i\cdot}|z_i}(y|z)f_{z_i}(z)dz}. \tag{2.102}$$

The numerator is given by

$$f_{y_{i\cdot}|z_i}(y|z)f_{z_i}(z) = \left(\frac{1}{1+e^{-(\mu+\sigma z)}}\right)^{y}\left(1-\frac{1}{1+e^{-(\mu+\sigma z)}}\right)^{n_i-y}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2},$$

so the estimated predicted value can now be calculated as

$$\begin{aligned}\tilde{p}_i &= \hat{E}[p_i|y_{i\cdot}] \qquad (2.103)\\ &= \frac{\int_{-\infty}^{\infty}\left(1+e^{-(\hat{\mu}+\hat{\sigma}z)}\right)^{-(y_{i\cdot}+1)}\left(1+e^{\hat{\mu}+\hat{\sigma}z}\right)^{-(n_i-y_{i\cdot})}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_i^2}dz}{\int_{-\infty}^{\infty}\left(1+e^{-(\hat{\mu}+\hat{\sigma}z)}\right)^{-(y_{i\cdot})}\left(1+e^{\hat{\mu}+\hat{\sigma}z}\right)^{-(n_i-y_{i\cdot})}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_i^2}dz}\end{aligned}$$

As in Section 2.6b numerical evaluation must be used; one possible method is Gauss–Hermite quadrature.

d. Probit-normal model

A model similar to the logit-normal model is the probit-normal model, which is obtained by replacing the logit function in (2.92) by Φ^{-1}, where Φ is the standard normal cumulative distribution function (c.d.f.):

$$\Phi(t) = \int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx. \tag{2.104}$$

This does not appreciably simplify the calculations, except $\mathrm{E}[y_{ij}]$ slightly, which is given by

$$\mathrm{E}[y_{ij}] = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \tag{2.105}$$

(see E 2.19).

Otherwise derivations and formulae closely follow those for the logit-normal in Section 2.7. For example, the analog of (2.100) is

$$\mathrm{E}[y_{ij}y_{il}] = \int_{-\infty}^{\infty}\Phi(\mu+\sigma z_i)^2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_i^2}dz_i\,. \tag{2.106}$$

2.7 COMPUTING

Even the most mathematically tractable of the models for binary data with random effects, the beta-binomial model, poses numerical difficulties. The more easily generalizable logit- and probit-normal models of the preceding section raise further problems. Though we have indicated a possible approach using Gauss–Hermite quadrature, this quickly becomes intractable, even for problems of moderate size. More is said about these issues in Chapter 10.

2.8 EXERCISES

E 2.1 Show that $N \log\left(1 + \frac{m-1}{N-m}\mathcal{F}^{m-1}_{N-m,1-\alpha}\right)$ tends to $\chi^2_{m-1,1-\alpha}$ for large N.

E 2.2 For the F-test of Section 2.1c, with $m = 5$ and $N = 20$, 50, and 100, calculate the significance level achieved if the asymptotic critical value is used instead of the exact critical value.

E 2.3 Derive (2.11).

E 2.4 Show for Section 2.2b that the ML solutions maximize the likelihood.

(a) First derivatives of l are shown at (2.25). Use them to find the three second derivatives with respect to μ.

(b) For the results in (a) and for the second derivatives shown in Section 2.2b–v, replace parameters by solutions of the ML equations.

(c) Put results from (b) in a matrix and explain why that matrix is negative definite, and hence the solutions maximize l.

E 2.5 When $\hat{\sigma}^2_a < 0$, and hence $\hat{\sigma}^2_a = 0$:

(a) Use (2.41) and (2.42) to show that for unbalanced data, $\hat{\mu} = \bar{y}_{\cdot\cdot}$ and $\hat{\sigma}^2 = \mathrm{SST}/N$.

(b) Why is (2.43) not used?

E 2.6 With

$$\begin{aligned} \log L &= -\tfrac{1}{2}N\log 2\pi - \tfrac{1}{2}(N-m)\log\sigma^2 - \tfrac{1}{2}\textstyle\sum_i \log\lambda_i \\ &\quad - \mathrm{SSE}/2\sigma^2 - \textstyle\sum[n_i(\bar{y}_{i\cdot} - \mu)^2/2\lambda_i] \end{aligned} \tag{2.107}$$

show that $L \to -\infty$ as $\sigma^2 \to 0$ and as $\sigma^2 \to \infty$, so that L must have a maximum for a positive value of σ^2.

E 2.7 For the hypothesis $H_0: \pi_i = \pi_j$ compare the χ^2 test, which uses only columns i and j of Table 2.1, with the test given by (2.66).

E 2.8 For the balanced data situation ($n_i \equiv n$) for model (2.15) show that

$$\frac{\bar{y}_{\cdot\cdot} - \mu}{\sqrt{\text{MSA}/mn}} \sim \mathcal{T}_{m-1},$$

E 2.9 Show that $\text{E}[Y|X]$ is the minimum mean square error predictor of Y. That is, show that $g(X) = \text{E}[Y|X]$ minimizes $\text{E}\left[(Y - g(X))^2\right]$ among all functions $g(\cdot)$ of X.

E 2.10 Suppose that X and Y are bivariate normal with correlation ρ. Show that if X is k standard deviations above its mean, then the minimum mean square error predictor is that Y will be ρk standard deviations above its mean.

E 2.11 (a) Derive $F = \text{MSA/MSE}$ as the LRT statistic for $H: \sigma_a^2 = 0$ in the one-way classification, random, normal model, balanced data. The derivation is lengthy. The following steps help.

(i.) Denote (2.24) as $l(\mu, \sigma^2, \sigma)$.

(ii.) Find $l(\dot{\mu}, \dot{\sigma}^2, \dot{\sigma})$.

(iii.) Find $l(\dot{\mu}_0, \dot{\sigma}_0^2)$ under H.

(iv.) Define $q = (m-1)F/m(n-1)$.

(v.) Show that $\partial(-2\log\Lambda)/\partial q > 0$ if $\dot{\sigma}_a^2 > 0$.

(vi.) Explain how this leads to F being a test statistic.

(b) For unbalanced data explain why (ii) cannot be obtained analytically, and so neither can $\log\Lambda$. But find (iii).

(c) Despite (b), show that F is a test statistic for $H: \sigma_a^2 = 0$. See Searle et al. (1992, p. 76).

E 2.12 From (2.60) find $-\text{E}\left[\partial^2 l/\partial\pi_k^2\right]$ and from that the sampling variance of $\hat{\pi}_k$.

E 2.13 For the beta-binomial model given by (2.72) with the parameterization $\mu = \alpha/(\alpha+\beta)$ and $\tau = 1/(\alpha+\beta)$, derive (2.81) from (2.80).

E 2.14 Prove the penultimate sentence before (2.90).

E 2.15 For the beta-binomial model given by (2.72) show that the conditional distribution of p_i given $y_{i\cdot}$ is again beta, but with parameters $\alpha + y_{i\cdot}$ and $\beta + n_i - y_{i\cdot}$.

E 2.16 Prove (2.90).

E 2.17 Show that the derivative of the log of L from (2.95) evaluated at $\sigma = 0$ is a function of

$$\sum_i n_i(\hat{\pi}_i - \pi)^2,$$

as in (2.63). The score test of $H_0 : \sigma = 0$ is based on this statistic. Hence show that when properly standardized and with the MLEs substituted for unknown parameters, this is the same as the Pearson chi-square statistic of Section 2.5e for testing homogeneity in the fixed effects model. *Hint*: To calculate the derivative you will need to use L'Hospital's rule.

E 2.18 For large samples from model (2.92), show that the maximum likelihood estimator of σ is equal to zero with probability 1/2. Hence show that $-2\log\Lambda$ for testing $H_0 : \sigma = 0$ is zero with probability 1/2.

E 2.19 For the probit-normal model show that $\mathrm{E}[y_{ij}] = \Phi(\mu/\sqrt{1+\sigma^2})$.

Chapter 3

SINGLE-PREDICTOR REGRESSION

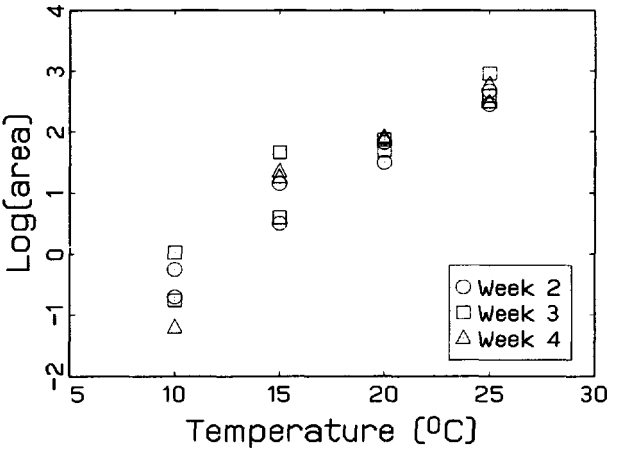
3.1 INTRODUCTION

Chapter 2 deals with the one-way classification for data which are either normally or Bernoulli distributed. For each of these distributions this chapter covers simple regression, i.e., regression with a single predictor. For the one-way classification in Chapter 2 we described the class means by using different parameters for each class. As such, we made no assumptions about the form of the mean of y as a function of the classification variable. In contrast, for simple linear regression we make the restrictive assumption that y is a linear function of a predictor, x. For example, we will consider a case study in which the mean of y, log radial growth of colonies of *Phytophthora infestans sporangia* inoculated onto potato leaflets, is modeled as a linear function of the temperature at which the colonies were allowed to grow.

In practice, there are many situations where the linearity assumption is not met, in which case we must regard it either as a crude approximation or merely the first step in a more in-depth analysis. For some cases it is adequate to assume that the mean of y is a linear function of x, but over only a short interval of x. For example, a plot of the radial growth data (Figure 3.1) for weeks 2, 3 and 4 for temperatures 15°C through 25°C shows an approximately linear relationship. However, over the entire range of the experiment (down to 10°C), the relationship appears nonlinear.

In other cases we must transform y and/or x before the linearity as-

Figure 3.1: Log(area) versus temperature for the lesion data.



sumption is met even approximately. Alternatively or additionally, we might try more complicated models with multiple predictors designed to encompass more flexible functional forms. We describe such models in subsequent chapters.

A different approach is to model some known function of the mean of y, call it μ, as linear in x. This is called a *generalized linear model*, examples of which are found in Sections 3.7 and 3.8. There we argue that in many instances of Bernoulli-distributed data, it does not make sense to assume a simple linear regression model. Instead we model $\log[\mu/(1-\mu)]$ as linear in the predictor.

3.2 NORMALITY: SIMPLE LINEAR REGRESSION

a. Model

We begin with normally distributed data, for which the well-known and frequently used simple linear regression model is given by

$$\begin{aligned} \mathrm{E}[y_i] &= \mu(x_i) = \alpha + \beta x_i \\ y_i &\sim \text{indep. } \mathcal{N}[\mu(x_i), \sigma^2]; \qquad i = 1, 2, \ldots, N, \end{aligned} \tag{3.1}$$

where the notation $\mu(x)$ is used to indicate that μ is a function of x. The x_i are assumed to be known constants, either fixed as part of the data collection process or regarded as fixed by considering the

conditional distribution given the x's. Note that (3.1) encompasses four assumptions:

1. The y_i follow a normal distribution.
2. The y_i are independent.
3. The y_i all have the same variance, σ^2.
4. The mean of y is a linear function of the predictor, x.

Any or all of these may be regarded simply as adequate approximations or as the beginning of a more serious analysis.

b. Likelihood

The log likelihood is easily derived:

$$l = \log L = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i (y_i - \alpha - \beta x_i)^2. \quad (3.2)$$

c. Maximum likelihood estimators

The maximum likelihood estimators can be found by equating the derivatives of the log likelihood to zero. Those derivatives are

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) \quad (3.3)$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) \quad (3.4)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2. \quad (3.5)$$

In equating these to zero, we can replace parameters by MLEs (e.g., replace α by $\hat{\alpha}$) because the solutions to the resulting equations are indeed the MLEs. The distinction between solutions and estimators (discussed in Section 1.7a–iii) is not problematic here because α and β can be any real numbers and therefore so can $\hat{\alpha}$ and $\hat{\beta}$, as evident in

(3.6) and (3.7). And, from equation (3.8), $\hat{\sigma}^2$ is non-negative, in accord with the definition of σ^2. We therefore have

$$\begin{aligned} \sum_i (y_i - \mathrm{E}[y_i]) &= 0 \quad \text{or} \\ \sum_i y_i &= n\hat{\alpha} + \hat{\beta} \sum_i x_i \end{aligned} \tag{3.6}$$

from (3.3);

$$\begin{aligned} \sum_i x_i (y_i - \mathrm{E}[y_i]) &= 0 \quad \text{or} \\ \sum_i x_i y_i &= \hat{\alpha} \sum_i x_i + \hat{\beta} \sum_i x_i^2 \end{aligned} \tag{3.7}$$

from (3.4); and

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \tag{3.8}$$

from (3.5). We can straightforwardly solve these equations:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)/N}{\sum_i x_i^2 - (\sum_i x_i)^2/N} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\sigma}^2 &= \frac{1}{N} \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2. \end{aligned} \tag{3.9}$$

Again, the Chapter 6 general REML equation yields the REML estimator of σ^2. It is exactly $\hat{\sigma}^2$ of (3.9) except for the important replacement of N by $N-2$ to account for the two fixed effects α and β (see Section 1.7a–ii).

d. Distributions of MLEs

Standard derivations (e.g., Weisberg, 1980, p. 44) give the distributions of the MLEs. Defining $S_{xx} = \sum_i (x_i - \bar{x})^2$, the MLEs of α and β are bivariate normal:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \frac{\sigma^2}{S_{xx}} \begin{pmatrix} \sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \right]. \tag{3.10}$$

They are independent of $\hat{\sigma}^2$, which is distributed as a multiple of a chi-square distribution:

$$\frac{N\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-2}. \tag{3.11}$$

e. Tests and confidence intervals

Tests and confidence intervals can be derived utilizing the t-distribution. For example, a confidence interval for β is given by

$$\hat{\beta} \pm t_{N-2,\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta})}$$

which is

$$\hat{\beta} \pm t_{N-2,\alpha/2}\sqrt{\frac{\frac{N}{N-2}\hat{\sigma}^2}{S_{xx}}}. \tag{3.12}$$

Similarly, a test of $H_0: \beta \leq 0$ versus $H_A: \beta > 0$ would be to reject the H_0 if

$$\hat{\beta} \Big/ \sqrt{\frac{\frac{N}{N-2}\hat{\sigma}^2}{S_{xx}}} > t_{N-2,\alpha}. \tag{3.13}$$

Tests and confidence intervals for α can be derived in the same manner.

A confidence interval for σ^2 can be calculated using (3.11) as

$$\left(\frac{N\hat{\sigma}^2}{\chi^2_{N-2,1-\alpha/2}}, \frac{N\hat{\sigma}^2}{\chi^2_{N-2,\alpha/2}}\right). \tag{3.14}$$

f. Illustration

For the *Phytophthora* data of Figure 3.1 (using weeks 2 through 4, temperatures 15°C through 25°C only) we consider a linear regression model for the average of the two measurements on a leaflet

$$\begin{aligned} \text{E}[\text{ALD}_j] &= \mu_j = \alpha + \beta\text{TEMP}_j \\ \text{ALD}_j &\sim \text{indep. } \mathcal{N}(\mu_j, \sigma^2), \end{aligned}$$

where $\text{ALD}_j = y_j$ is the average log diameter of the lesions on the jth leaflet, and $\text{TEMP}_j = x_j$ is the temperature for the jth leaflet. For these data, the maximum likelihood estimators are $\hat{\alpha} = -1.296, \hat{\beta} =$

0.157, and $\hat{\sigma}^2 = 0.0755$. From these we calculate a 95% confidence interval for β as

$$\begin{aligned} 0.157 \pm t_{N-2,0.025}\sqrt{\frac{\frac{N}{N-2}0.0755}{S_{xx}}} &= 0.157 \pm t_{16,0.025}\sqrt{\frac{\frac{18}{16}(0.0755)}{300}} \\ &= 0.157 \pm 2.120\sqrt{0.000283} \\ &= 0.157 \pm 0.0357 \\ &= (0.121, 0.193). \end{aligned}$$

We are thus 95% confident that the average log lesion diameter increases by between 0.121 and 0.193 with each increase in temperature of 1° C over the range of temperatures from 15°C through 25°C.

3.3 NORMALITY: A NONLINEAR MODEL

a. Model

A variation on (3.1) is to assume that the model is nonlinear in its parameters. A simple example is

$$\begin{aligned} \mathrm{E}[y_i] &= \mu(x_i) = e^{\alpha+\beta x_i} \\ y_i &\sim \text{indep. } \mathcal{N}[\mu(x_i), \sigma^2]; \qquad i = 1, 2, \ldots, N. \end{aligned} \tag{3.15}$$

Note that we have changed only the mean of y_i, not assumptions about its distribution.

b. Likelihood

The log likelihood is basically the same as (3.2):

$$l = \log L = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i [y_i - \mu(x_i)]^2 \tag{3.16}$$

except that $\mu(x_i)$ in (3.16) is $e^{\alpha+\beta x_i}$ instead of $\alpha + \beta x_i$ of (3.1).

c. Maximum likelihood estimators

From (3.16) we can see that to maximize the likelihood with respect to α and β we minimize the residual sum of squares: $\sum[y_i - \mu(x_i)]^2$. So

maximum likelihood for homoscedastic, normal distribution models is equivalent to least squares, even with nonlinear models. More formally, we can differentiate l of (3.16) to try to maximize it. The derivatives are

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - e^{\alpha+\beta x_i})e^{\alpha+\beta x_i} \tag{3.17}$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - e^{\alpha+\beta x_i})e^{\alpha+\beta x_i}x_i \tag{3.18}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{N}(y_i - e^{\alpha+\beta x_i})^2. \tag{3.19}$$

Setting these equal to zero, with parameters replaced by MLEs (e.g., α replaced by $\hat{\alpha}$ as in (3.9)), gives

$$\sum_i y_i e^{\hat{\alpha}+\hat{\beta}x_i} = \sum_i e^{2(\hat{\alpha}+\hat{\beta}x_i)} \tag{3.20}$$

$$\sum_i y_i e^{\hat{\alpha}+\hat{\beta}x_i}x_i = \sum_i e^{2(\hat{\alpha}+\hat{\beta}x_i)}x_i \tag{3.21}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (y_i - e^{\hat{\alpha}+\hat{\beta}x_i})^2 \tag{3.22}$$

The first two equations can be solved for $\hat{\alpha}$ and $\hat{\beta}$ which are then substituted in the third equation to find $\hat{\sigma}^2$. To solve (3.20) and (3.21) for $\hat{\alpha}$ and $\hat{\beta}$ reduce the equations to

$$\sum_i y_i e^{\hat{\beta}x_i} = e^{\hat{\alpha}}\sum_i e^{2\hat{\beta}x_i} \tag{3.23}$$

and

$$\sum_i y_i e^{\hat{\beta}x_i}x_i = e^{\hat{\alpha}}\sum_i e^{2\hat{\beta}x_i}x_i \tag{3.24}$$

or

$$\sum_i y_i e^{\hat{\beta}x_i} / \sum_i y_i e^{\hat{\beta}x_i}x_i = \sum_i e^{2\hat{\beta}x_i} / \sum_i e^{2\hat{\beta}x_i}x_i. \tag{3.25}$$

This equation must be solved numerically to find $\hat{\beta}$. We can then obtain

$$\hat{\alpha} = \log \frac{\sum_i y_i e^{\hat{\beta}x_i}}{\sum_i e^{2\hat{\beta}x_i}}$$

$$= \log \frac{\sum_i y_i e^{\hat{\beta} x_i} x_i}{\sum_i e^{2\hat{\beta} x_i} x_i} \tag{3.26}$$

and

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i (y_i - e^{\hat{\alpha} + \hat{\beta} x_i})^2. \tag{3.27}$$

What about restricted maximum likelihood for this model? The usual basis for REML is linear combinations of the data chosen to be free of the fixed effects. With a nonlinear model such as this one, that is not possible.

d. Distributions of MLEs

The MLEs are nonlinear functions of the data and do not having closed-form expressions and this precludes our working out their exact, small-sample distributions. Simulations and calculations for small sample sizes show that the estimators for the parameters α, β, and σ^2 are biased.

The large-sample distributions of the MLEs (see Section S.4d of Appendix S) are, as usual, tractable:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\sigma}^2 \end{pmatrix} \sim \mathcal{AN} \left[\begin{pmatrix} \alpha \\ \beta \\ \sigma^2 \end{pmatrix}, \frac{\sigma^2}{\Delta} \begin{pmatrix} \sum \mu_i^2 x_i^2 & -\sum \mu_i^2 x_i & 0 \\ -\sum \mu_i^2 x_i & \sum \mu_i^2 & 0 \\ 0 & 0 & \frac{2\sigma^2 \Delta}{N} \end{pmatrix} \right], \tag{3.28}$$

where $\Delta = \sum \mu_i^2 x_i^2 \sum \mu_i^2 - (\sum \mu_i^2 x_i)^2$, and $\mu_i = \exp(\alpha + \beta x_i)$.

This points out that the nice properties of ML estimators under the linear model (3.1) are quite delicate. A small change to the model [from (3.1) to (3.15)] causes the estimators not to have closed-form expressions, makes it impossible to work out their small-sample distribution and causes the estimator to be biased (in small samples).

3.4 TRANSFORMING VERSUS LINKING

a. Transforming

A temptation to resolve the difficulties inherent in (3.15) is to work with the log transform of the data and assume the model

$$\mathrm{E}[\log y_i] = \alpha^* + \beta^* x_i$$

$$\log y_i \sim \text{indep. } \mathcal{N}(\alpha^* + \beta^* x_i, \sigma^{*2}); \quad i = 1, 2, \ldots, N, \quad (3.29)$$

where we have used superscript asterisks to indicate that the parameters are different from those of (3.15). Since the model for $\log y_i$ is now linear with homoscedastic, normal distributions, we regain the nice properties of ML estimators.

b. Linking

In (3.15) we assume that a function of the mean is linear in the parameters. We will call this the link function. In that case we are assuming that $\log \mathrm{E}[y_i] = \alpha + \beta x_i$, so we are using a log link. However, this is not the same as in (3.29). First, using (3.15), $\mathrm{E}[\log y_i]$ does not exist (since negative values of y_i are possible). More practically, even if y_i had a distribution such that it was positive with probability 1, by Jensen's inequality (Casella and Berger, 1990, p. 182)

$$\mathrm{E}[\log y_i] < \log(\mathrm{E}[y_i]) = \log[e^{\alpha+\beta x_i}] = \alpha + \beta x_i. \quad (3.30)$$

Second, (3.15) has y_i homoscedastic on the original scale while (3.29) has y_i homoscedastic after taking the log transformation. Under (3.29) it is easy to show (E 3.3) that the standard deviation of y_i increases proportionally with the mean.

c. Comparisons

Thus a log transformation of the data is not the same as using the log link, namely that $\log(\mathrm{E}[y_i])$ follows a linear model. Choosing between (3.15) and (3.29) would ordinarily be done by checking whether the variance is constant on the original scale or increases with the mean (Ruppert et al., 1989).

3.5 RANDOM INTERCEPTS: BALANCED DATA

From the example of observing lesion growth on potato leaflets described at the start of this chapter, let us concentrate on y_{ij} being the log lesion area observed in week i at temperature x_{ij}. Then, as an extension of traditional single-predictor regression, $\mathrm{E}[y_{ij}] = \alpha + \beta x_{ij}$, we now consider

$$\mathrm{E}[y_{ij}|a_i] = \mu + a_i + \beta x_{ij} \quad (3.31)$$

for $i = 1, 2, \ldots, m$ and (initially for balanced data) for $j = 1, 2, \ldots, n$. The important feature attributed to (3.31) is that the intercepts $\mu + a_i$ are taken as being random; that is, we treat the a_i as random effects, normally distributed with zero mean:

$$a_i \sim \text{ i.i.d. } \mathcal{N}(0, \sigma_a^2). \tag{3.32}$$

In every week, $i = 1, 2, \ldots, m$, the same n temperatures are used so that $x_{ij} = x_j$ for $j = 1, 2, \ldots, n$ for all i. Thus (3.31) becomes

$$\mathrm{E}[y_{ij}|a_i] = \mu + a_i + \beta x_j. \tag{3.33}$$

a. The model

Suppose that for a given i we write down the equation (3.33) for each $j = 1, 2, \ldots, n$. This gives n equations

$$\begin{aligned} \mathrm{E}[y_{i1}|a_i] &= \mu + a_i + \beta x_1, \\ \mathrm{E}[y_{i2}|a_i] &= \mu + a_i + \beta x_2, \\ &\vdots \\ \mathrm{E}[y_{in}|a_i] &= \mu + a_i + \beta x_n. \end{aligned} \tag{3.34}$$

Now define three column vectors of order n:

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in} \end{bmatrix}, \qquad \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{x}_0 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{3.35}$$

Thus $\mathbf{y}_i$ is the vector of all n observations in the ith week, $\mathbf{1}_n$ is a vector of n ones, and $\mathbf{x}_0$ is the vector of the n different x-values associated with the n observations each week. Then, with the definitions of (3.35), equations (3.34) can be written succinctly as

$$\mathrm{E}[\mathbf{y}_i|a_i] = \mu \mathbf{1}_n + a_i \mathbf{1}_n + \beta \mathbf{x}_0 \tag{3.36}$$

which we rewrite as

$$\mathrm{E}[\mathbf{y}_i|a_i] = \mathbf{X}_0 \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \mathbf{Z}_0 a_i \tag{3.37}$$

with

$$\mathbf{X}_0 = [\mathbf{1}_n \ \mathbf{x}_0] \quad \text{and} \quad \mathbf{Z}_0 = \mathbf{1}_n. \tag{3.38}$$

We now assume normality for $y_{ij}|a_i$ in the form

$$\mathbf{y}_i|a_i \sim \text{ indep. } \mathcal{N}\left(\mu\mathbf{1}_n + a_i\mathbf{1}_n + \beta\mathbf{x}_0,\ \sigma^2\mathbf{I}_n\right). \tag{3.39}$$

Then with (1.14) and (1.17) we have $\text{var}(y_{ij}) = \sigma_a^2 + \sigma^2$ and also $\text{cov}(y_{ij}, y_{ij'}) = \sigma_a^2$. Thus on defining

$$\mathbf{V}_0 = \text{var}(\mathbf{y}_i) = \sigma^2\mathbf{I}_n + \sigma_a^2\mathbf{J}_n, \tag{3.40}$$

we have

$$\mathbf{V}_0^{-1} = \frac{1}{\sigma^2}\left(\mathbf{I}_n - \frac{\sigma_a^2}{\sigma^2 + n\sigma_a^2}\mathbf{J}_n\right) \tag{3.41}$$

where, as described in Section M.1 of Appendix M, $\mathbf{I}_n$ is an identity matrix and $\mathbf{J}_n$ is a square matrix of all ones.

Notation For purposes of subsequent algebra it turns out to be useful to define

$$\tau = \frac{n\sigma_a^2}{\sigma^2 + n\sigma_a^2} = \frac{\sigma_a^2}{\sigma^2/n + \sigma_a^2} \tag{3.42}$$

with

$$1 - \tau = \frac{\sigma^2}{\sigma^2 + n\sigma_a^2} = (\sigma^2/n\sigma_a^2)\tau. \tag{3.43}$$

This gives

$$\mathbf{V}_0^{-1} = \frac{1}{\sigma^2}(\mathbf{I}_n - \tau\bar{\mathbf{J}}_n) \quad \text{for} \quad \bar{\mathbf{J}} = \frac{1}{n}\mathbf{J}_n. \tag{3.44}$$

To encompass all the data, namely $\mathbf{y}_i$ for $i = 1, 2, \ldots, m$, we define

$$\mathbf{y} = \left\{{}_c\mathbf{y}_i\right\}_{i=1}^m \quad \text{and} \quad \mathbf{a} = \left\{{}_c a_i\right\}_{i=1}^m. \tag{3.45}$$

Then from (3.37) we get

$$\mathbf{E}[\mathbf{y}|\mathbf{a}] = \mathbf{X}\begin{bmatrix} \mu \\ \beta \end{bmatrix} + \mathbf{Z}\mathbf{a}$$

and

$$\mathrm{E}[\mathbf{y}] \;=\; \mathbf{X}\begin{bmatrix}\mu\\ \beta\end{bmatrix} \tag{3.46}$$

with

$$\mathbf{X} = \mathbf{1}_m \otimes \mathbf{X}_0 \qquad \text{and} \qquad \mathbf{Z} = \mathbf{I}_m \otimes \mathbf{Z}_0 \tag{3.47}$$

where $\otimes$ represents the direct (or Kronecker) product operation as described in Section M.2 of Appendix M. Similarly, with the $\mathbf{y}_i$–vectors being independent (being data from different weeks) and with every $\mathbf{y}_i$ having the same variance–covariance matrix, namely $\mathbf{V}_0$ of (3.40),

$$\mathbf{V} = \mathrm{var}(\mathbf{y}) = \mathbf{I}_m \otimes \mathbf{V}_0, \tag{3.48}$$

a block diagonal matrix of m matrices $\mathbf{V}_0$ on the diagonal. And

$$\mathbf{V}^{-1} = \left(\mathbf{I}_m \otimes \mathbf{V}_0^{-1}\right). \tag{3.49}$$

Thus

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\begin{bmatrix}\mu\\ \beta\end{bmatrix},\ \mathbf{V}\right). \tag{3.50}$$

b. Estimating μ and β

ML estimators (under normality, and assuming $\mathbf{V}$ known) of μ and β come from the general expression given in Section S.6a of Appendix S.

$$\begin{bmatrix}\hat{\mu}\\ \hat{\beta}\end{bmatrix} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{3.51}$$

– i. *Estimation*

On substituting for $\mathbf{X}$ and $\mathbf{V}^{-1}$ from (3.47) and (3.49) we find (3.51) reduces to

$$\begin{bmatrix}\hat{\mu}\\ \hat{\beta}\end{bmatrix} = \left[\frac{1}{m} \otimes (\mathbf{X}_0'\mathbf{V}_0^{-1}\mathbf{X}_0)^{-1}\right]\left(\mathbf{1}_m' \otimes \mathbf{X}_0'\mathbf{V}_0^{-1}\right)\mathbf{y}. \tag{3.52}$$

To simplify this expression note that

$$\mathbf{X}_0'\mathbf{V}_0^{-1} \;=\; \begin{bmatrix}\mathbf{1}_n'\\ \mathbf{x}_0'\end{bmatrix}\frac{1}{\sigma^2}\left(\mathbf{I}_n - \tau\bar{\mathbf{J}}_n\right) = \begin{bmatrix}(\tau/n\sigma_a^2)\mathbf{1}_n'\\ (\mathbf{x}_0' - \tau\bar{x}.\mathbf{1}_n')/\sigma^2\end{bmatrix} \tag{3.53}$$

and

$$\mathbf{X}_0'\mathbf{V}_0^{-1}\mathbf{X}_0 = \mathbf{X}_0'\mathbf{V}_0^{-1}[\mathbf{1}_n \quad \mathbf{x}_0] = \left(\tau/\sigma_a^2\right)\begin{bmatrix} 1 & \bar{x}. \\ \bar{x}. & S_{xx}\sigma_a^2/\tau\sigma^2 + \bar{x}.^2 \end{bmatrix} \tag{3.54}$$

where

$$S_{xx} = \sum_{j=1}^{n}(x_j - \bar{x}.)^2 \quad \text{with} \quad \bar{x}. = \sum_{j=1}^{n} x_j/n. \tag{3.55}$$

Also for (3.52), using (3.53) leads to

$$\left(\mathbf{1}_m' \otimes \mathbf{X}_0'\mathbf{V}_0^{-1}\right)\mathbf{y} = \left(m\tau/\sigma_a^2\right)\begin{bmatrix} \bar{y}.. \\ S_{xy}\sigma_a^2/\tau\sigma^2 + \bar{x}.\bar{y}.. \end{bmatrix} \tag{3.56}$$

where

$$S_{xy} = \sum_{j}(x_j - \bar{x}.)(\bar{y}._j - \bar{y}..).$$

The inverse of (3.54) is easily derived, being

$$\left(\mathbf{X}_0'\mathbf{V}_0^{-1}\mathbf{X}_0\right)^{-1} = \frac{\sigma^2}{S_{xx}}\begin{bmatrix} S_{xx}\sigma_a^2/\tau\sigma^2 + \bar{x}.^2 & -\bar{x}. \\ -\bar{x}. & 1 \end{bmatrix}, \tag{3.57}$$

and post-multiplying this by (3.56) leads (see E 3.2), from (3.52), to

$$\hat{\mu} = \bar{y}.. - \hat{\beta}\bar{x}.. \tag{3.58}$$

and

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{j=1}^{n} x_j\bar{y}._j - n\bar{x}.\bar{y}..}{\sum_{j=1}^{n} x_j^2 - n\bar{x}.^2}. \tag{3.59}$$

An immediately noticeable feature of this result is that $\hat{\mu}$ and $\hat{\beta}$ do not depend on the unknown variances σ_a^2 and σ^2. Also, these estimators are exactly the same (for balanced data) as when the a_i-effects are fixed rather than random, or indeed if there are no a_i-effects. Moreover, (3.58) and (3.59) are precisely the results that occur in traditional analysis of covariance as in, for example, equations (6) and (7) of Searle (1987, Chapter 6).

– ii. ***Unbiasedness***

On using $\mathrm{E}[a_i] = 0$ of (3.32) the expected values of $\hat{\mu}$ and $\hat{\beta}$ are

$$\mathrm{E}[\hat{\beta}] = \frac{\sum_{j=1}^{n} x_j(\mu + \mathrm{E}[\bar{a}.] + \beta x_j) - n\bar{x}.(\mu + \mathrm{E}[\bar{a}.] + \beta\bar{x}.)}{\sum_{j=1}^{n} x_j^2 - n\bar{x}.^2} = \beta$$

and

$$\mathrm{E}[\hat{\mu}] = (\mu + \mathrm{E}[\bar{a}.] + \beta\bar{x}.) - \beta\bar{x}. = \mu\,.$$

Thus the estimators $\hat{\mu}$ and $\hat{\beta}$ are unbiased.

– iii. ***Sampling distributions***

From (3.58) and (3.59) we see that $\hat{\mu}$ and $\hat{\beta}$ are linear combinations of the normally distributed y_{ij}—see (3.39), and so the estimators themselves are bivariate normally distributed. Their means are μ and β, and using the well-known result that the variance-covariance matrix of estimators $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ of (3.51) is $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, we see that this is $(1/m)(\mathbf{X}_0'\mathbf{V}_0^{-1}\mathbf{X}_0)^{-1}$ as occurs in (3.52). And so from (3.57)

$$\mathrm{var}(\hat{\mu}) = \frac{\sigma_a^2}{m\tau} + \frac{\sigma^2\bar{x}.^2}{mS_{xx}} = \frac{\sigma_a^2 + \sigma^2/n}{m} + \frac{\sigma^2\bar{x}.^2}{mS_{xx}}, \tag{3.60}$$

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{mS_{xx}}, \tag{3.61}$$

and

$$\mathrm{cov}(\hat{\mu}, \hat{\beta}) = -\bar{x}.\mathrm{var}(\hat{\beta}). \tag{3.62}$$

Except for the occurrence of σ_a^2 in (3.60), these results are essentially the same as with standard, simple regression of Section 3.1. One apparent difference is the presence of m in the denominators, arising from the fact that S_{xx} is defined, in (3.55), as $S_{xx} = \Sigma_j(x_j - \bar{x}.)^2 = \Sigma_j x_j^2 - n\bar{x}.^2$ and not as

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_j^2 - mn\bar{x}.^2 = m\left(\sum_{j=1}^{n} x_j^2 - n\bar{x}.^2\right) = mS_{xx}.$$

c. Estimating variances

In Section 3.5b the ML estimation of μ and β was dealt with on the assumption of (normality and) knowing $\mathbf{V}$. Interestingly, the resulting estimators, $\hat{\mu}$ and $\hat{\beta}$ of (3.58) and (3.59), do not depend on $\mathbf{V}$; i.e., they do not depend on σ_a^2 and σ^2. However, these variances are often unknown, in which case we will want to estimate them, not only for their own sake but also to use them in, for example, the variances of the estimators in Section 3.5b–iii. And if we are going to estimate σ_a^2 and σ^2 from the same data set as will be used for estimating μ and β (as is often done) it will be advisable to use ML for estimating all four parameters, μ, β, σ_a^2 and σ^2, simultaneously. In doing this, the equations for estimating μ and β will be exactly the same as those already considered except that the ML estimators $\hat{\sigma}_a^2$ and $\hat{\sigma}^2$ will replace σ_a^2 and σ^2 in those equations. However, because σ_a^2 and σ^2 do not occur in those equations the estimators $\hat{\mu}$ and $\hat{\beta}$ when estimating all four parameters will be exactly the same as already derived in (3.58) and (3.59). This means that to estimate σ_a^2 and σ^2 we can maximize, with respect to σ_a^2 and σ^2, the likelihood with μ and β replaced by $\hat{\mu}$ and $\hat{\beta}$. Thus if we write the log likelihood as $l(\mu, \beta, \sigma^2, \sigma_a^2)$ we maximize

$$l^* = \log L(\hat{\mu}, \hat{\beta}, \sigma^2, \sigma_a^2).$$

– i. *When ML solutions are estimators*

Suppose we write $\boldsymbol{\theta} = \mathrm{E}[\mathbf{y}] = \mathbf{X}\begin{bmatrix} \mu \\ \beta \end{bmatrix}$ of (3.50), in which the normality assumed therein gives the likelihood as

$$L = \frac{\exp\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\theta})'\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\theta})\}}{(2\pi)^{\frac{mn}{2}}|\mathbf{V}|^{\frac{1}{2}}}. \tag{3.63}$$

Then, using $\mathbf{V}^{-1}$ of (3.49) and

$$\begin{aligned} |\mathbf{V}|^{\frac{1}{2}} &= |\mathbf{V}_0|^{\frac{1}{2}m} = |\sigma^2\mathbf{I}_n + \sigma^2\mathbf{J}_n|^{\frac{1}{2}m} = \left[\sigma^{2(n-1)}(\sigma^2 + n\sigma_a^2)\right]^{\frac{1}{2}m} \\ &= \sigma^{2\left[\frac{m(n-1)}{2}\right]}(\sigma^2 + n\sigma_a^2)^{\frac{1}{2}m}, \end{aligned} \tag{3.64}$$

it can be shown that

$$l^* = -\frac{1}{2}\left[mn\log 2\pi + m(n-1)\log\sigma^2 + m\log(\sigma^2 + n\sigma_a^2)\right.$$

$$+ \frac{S_1}{\sigma^2} + \frac{n\sigma_a^2 S_2}{\sigma^2(\sigma^2 + n\sigma_a^2)}\Bigg] \qquad (3.65)$$

where

$$S_1 = \sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\beta}x_j)^2 \quad \text{and} \quad S_2 = \sum_i n(\bar{y}_{i\cdot} - \hat{\mu} - \hat{\beta}\bar{x}_\cdot)^2. \qquad (3.66)$$

After substituting for $\hat{\mu}$ and $\hat{\beta}$ from (3.58) and (3.59)

$$S_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}[y_{ij} - \bar{y}_{\cdot\cdot} - \hat{\beta}(x_j - \bar{x}_\cdot)]^2 = \mathrm{SST} - \hat{\beta}^2 m S_{xx} \qquad (3.67)$$

and

$$S_2 = \sum_{i=1}^{m} n\left[\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} - \hat{\beta}(\bar{x}_\cdot - \bar{x}_\cdot)\right]^2 = \sum_{i=1}^{m} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \mathrm{SSA} \qquad (3.68)$$

where the familiar notation for sums of squares in analysis of variance is introduced:

$$\mathrm{SSA} = \textstyle\sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2, \quad \text{with} \quad \mathrm{MSA} = \mathrm{SSA}/(m-1) \qquad (3.69)$$

$$\mathrm{SSE} = \textstyle\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2, \quad \text{with} \quad \mathrm{MSE} = \mathrm{SSE}/[m(n-1)] \qquad (3.70)$$

$$\mathrm{SST} = \textstyle\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \mathrm{SSA} + \mathrm{SSE}. \qquad (3.71)$$

We also have occasion later to use notation familiar to analysis of covariance:

$$\mathrm{SSC} = \hat{\beta}^2 m S_{xx} = \frac{\left[\sum_i \sum_j (x_j - \bar{x}_\cdot)(y_{ij} - \bar{y}_{i\cdot})\right]^2}{\sum_i \sum_j (x_j - \bar{x}_\cdot)^2}, \qquad (3.72)$$

the sum of squares due to covariance, and

$$\mathrm{SSR} = \mathrm{SSE} - \mathrm{SSC} = \sum_i \sum_j \left[y_{ij} - \bar{y}_{\cdot\cdot} - \hat{\beta}(x_j - \bar{x}_\cdot)\right]^2, \qquad (3.73)$$

the sum of squares for residual. This gives (3.67) as

$$S_1 = \mathrm{SSA} + \mathrm{SSR}. \qquad (3.74)$$

Now, differentiating l^* with respect to σ_a^2 and σ^2 and equating the results to zero yields ML solutions $\dot{\sigma}_a^2$ and $\dot{\sigma}^2$ (see Section 1.7a–iii, for

the overhead dot notation). First, $\partial l^*/\partial \sigma_a^2 = 0$ yields

$$\frac{-mn}{\dot{\sigma}^2 + n\dot{\sigma}_a^2} + \frac{n\, S_2}{(\dot{\sigma}^2 + n\dot{\sigma}_a^2)^2} = 0$$

so giving

$$\dot{\sigma}^2 + n\dot{\sigma}_a^2 = \text{SSA}/m. \tag{3.75}$$

Next, after some considerable algebra, and using (3.75) we find that $\partial l^*/\partial \sigma^2$ ultimately yields

$$\hat{\sigma}^2 = \frac{\text{SSE} - m\hat{\beta}^2 S_{xx}}{m(n-1)} = \frac{\text{SSR}}{m(n-1)}. \tag{3.76}$$

This is exactly the same result, for balanced data, as when the a_i-effects are fixed, not random. Indeed it is the standard analysis of covariance result as in, for example, equations (23) and (25) of Chapter 6 of Searle (1987). Then from (3.75)

$$\dot{\sigma}_a^2 = \frac{1}{n}\left(\frac{\text{SSA}}{m} - \hat{\sigma}^2\right) = \frac{1}{n}\left[\left(1 - \frac{1}{m}\right)\text{MSA} - \text{MSE}\right] + \frac{\hat{\beta}^2 S_{xx}}{n(n-1)}. \tag{3.77}$$

Providing $\dot{\sigma}_a^2$ is not negative it is the ML estimator $\hat{\sigma}_a^2 = \dot{\sigma}_a^2$. In passing we note that (3.77) is the same as $\dot{\sigma}_a^2$ in (2.25) except for the addition of the term in $\hat{\beta}^2$.

– ii. *When an ML solution is negative*

But it is possible for $\dot{\sigma}_a^2$ to be negative, in which case it is not the ML estimator. And then neither is $\hat{\sigma}^2$ of (3.76) the ML estimator of σ^2. This is so because the general method of maximum likelihood estimation demands that ML estimators be within the range of their corresponding parameters; and negative $\dot{\sigma}_a^2$ is not within the non-negative range of σ_a^2. To overcome this difficulty we must adopt the ML methods for this situation (see Searle et al., 1992, Section 3.7a–iii) which lead to taking $\hat{\sigma}_a^2 = 0$. This being so we then have to maximize

$$l^*(\hat{\mu}, \hat{\beta}, \sigma^2, \hat{\sigma}_a^2 = 0) = -\frac{1}{2}\left[mn \log 2\pi + mn \log \sigma^2 + \frac{S_1}{\sigma^2}\right], \tag{3.78}$$

obtained from (3.65) by replacing σ_a^2 with zero. Equating to zero the differential of (3.78) with respect to σ^2 gives what will be denoted as $\hat{\sigma}_0^2$, which is $\hat{\sigma}_0^2 = S_1/mn$. And so using (3.74) for S_1 we have

$$\hat{\sigma}_0^2 = \frac{S_1}{mn} = \frac{\text{SSA} + \text{SSR}}{mn}. \tag{3.79}$$

d. Tests of hypotheses – using LRT

The likelihood ratio technique (LRT) of hypothesis testing is described in general terms in Section 1.7b–i. Suppose $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximizes a likelihood function $L(\boldsymbol{\theta})$ involving parameters $\boldsymbol{\theta}$. And denote by $\hat{\boldsymbol{\theta}}_0$ the value of $\boldsymbol{\theta}$ which maximizes the likelihood when the parameters are limited (restricted or defined) by a null hypothesis H_0 pertaining to some of the elements of $\boldsymbol{\theta}$. Then the likelihood ratio is

$$\Lambda = \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}. \tag{3.80}$$

It leads to a test statistic for the hypothesis H. We illustrate this for two hypotheses that are often of interest, namely $H_0: \beta = 0$, and $H_0: \sigma_a^2 = 0$.

Rather than using (3.80) in practical application, it is often easier to use the negative of twice its logarithm:

$$\begin{aligned} -2\log\Lambda &= -2\log L(\hat{\boldsymbol{\theta}}_0) + 2\log L(\hat{\boldsymbol{\theta}}) \\ &= -2l^*(\hat{\boldsymbol{\theta}}_0) + 2l^*(\hat{\boldsymbol{\theta}}). \end{aligned} \tag{3.81}$$

– i. *Using the maximized log likelihood* $l^*(\hat{\boldsymbol{\theta}})$

The maximized log likelihood is l^* of (3.65) with σ^2 replaced by $\hat{\sigma}^2$ of (3.76) and, on assuming $\dot{\sigma}_a^2$ is positive, with σ_a^2 replaced by $\hat{\sigma}_a^2 = \dot{\sigma}_a^2$ of (3.77). Then, with $\hat{\mu}$ and $\hat{\beta}$ used in S_1 and S_2 as in (3.67) and (3.68) we have

$$\begin{aligned} -2l^*(\hat{\boldsymbol{\theta}}) = mn\log 2\pi + m(n-1)\log\left(\frac{\text{SSR}}{m(n-1)}\right) + m\log\left(\frac{\text{SSA}}{m}\right) \\ + \frac{\text{SSA}+\text{SSR}}{\text{SSR}/[m(n-1)]} - \left[\frac{(n-1)\text{SSA}}{\text{SSR}} - 1\right]\frac{\text{SSA}}{\text{SSA}/m}, \end{aligned} \tag{3.82}$$

after some simplification of $n\hat{\sigma}_a^2/\hat{\sigma}^2$ for the last term. Then, after collecting all the terms in (3.82) that do not involve SSA and SSR into what we will call $f_1(m,n)$, we get

$$-2l^*(\hat{\boldsymbol{\theta}}) = f_1(m,n) + m(n-1)\log\text{SSR} + m\log\text{SSA}. \tag{3.83}$$

It is to be noticed that the LRT is defined in terms of maximum likelihood estimators. Yet in going from (3.65) to (3.83) we did, in fact, use $\hat{\sigma}_a^2 = \dot{\sigma}_a^2$, as if $\dot{\sigma}_a^2 > 0$. But we know that when $\dot{\sigma}_a^2 < 0$ we

take $\hat{\sigma}_a^2 = 0$; and no account of this has been used in deriving $\mathrm{L}(\hat{\boldsymbol{\theta}})$ of (3.83). The reason for this is that (see E 3.4), for $H_0: \sigma_a^2 = 0$, having $\hat{\sigma}_a^2 = 0$ leads to LRT = 1, for which value one would never reject H_0.

– ii. *Testing the hypothesis* $H_0: \sigma_a^2 = 0$

To derive $-2\log\Lambda$ for $H_0: \sigma_a^2 = 0$ we need first to estimate μ, β and σ^2 from the likelihood adapted by using 0 for σ_a^2. But with $\hat{\mu}$ and $\hat{\beta}$ of (3.58) and (3.59) not involving σ^2 or σ_a^2, we know they will be the same for $\hat{\boldsymbol{\theta}}_0$. And the estimator of σ^2 for $\hat{\boldsymbol{\theta}}_0$ will be $\hat{\sigma}_0^2$ obtained in (3.79). Therefore, $-2l^*(\hat{\boldsymbol{\theta}}_0)$ can be found by replacing σ^2 by $\hat{\sigma}_0^2$ of (3.79) and σ_a^2 by 0 in (3.65). This gives

$$\begin{aligned} -2l^*(\hat{\boldsymbol{\theta}}_0) &= mn\log 2\pi + m(n-1)\log\hat{\sigma}_0^2 + m\log\hat{\sigma}_0^2 + S_1/\hat{\sigma}_0^2 \\ &= mn\log 2\pi + mn\log\left(\frac{\mathrm{SSA}+\mathrm{SSR}}{mn}\right) + \frac{\mathrm{SSA}+\mathrm{SSR}}{(\mathrm{SSA}+\mathrm{SSR})/m} \\ &= f_2(m,n) + mn\log(\mathrm{SSA}+\mathrm{SSR}). \end{aligned} \tag{3.84}$$

Therefore, on ignoring $f_1(m,n)$ and $f_2(m,n)$, we get from (3.81)

$$\begin{aligned} -2\log\Lambda &= -2l^*(\hat{\boldsymbol{\theta}}_0) + 2l^*(\hat{\boldsymbol{\theta}}) \\ &= mn\log(\mathrm{SSA}+\mathrm{SSR}) \\ &\qquad -m(n-1)\log\mathrm{SSR} - m\log\mathrm{SSA} \\ &= mn\log\{\mathrm{SSR}(\mathrm{SSA}/\mathrm{SSR}+1)\} \\ &\qquad -m(n-1)\log\mathrm{SSR} - m\log\mathrm{SSA}. \end{aligned} \tag{3.85}$$

We now write

$$q = \frac{\mathrm{SSA}}{\mathrm{SSR}} = \frac{(m-1)\mathrm{MSA}}{[m(n-1)-1]\mathrm{MSR}} = \frac{m-1}{m(n-1)-1}F$$

where F is the F-statistic in an analysis of covariance for testing $H_0: a_i$ *all equal*; and if the a_i are all equal then, with probability 1.0, we have $\sigma_a^2 = 0$. And, on using q

$$-2\log\Lambda = m[n\log(q+1) - \log q].$$

To investigate the monotonicity of $-2\log\Lambda$ with respect to q we consider

$$\frac{\partial}{\partial q}[-2\log\Lambda] = m\left[\frac{n}{q+1} - \frac{1}{q}\right] = \frac{m}{q(q+1)}[(n-1)q - 1]$$

$$> 0 \quad \text{if} \quad q > \frac{1}{n-1} \tag{3.86}$$

$$\Rightarrow \quad \text{SSA} > \frac{\text{SSR}}{n-1}$$

$$\Rightarrow \quad \dot{\sigma}_a^2 > 0 \qquad \text{from (3.77);} \tag{3.87}$$

and this is the condition for $\hat{\sigma}_a^2 = \dot{\sigma}^2$. And, in (3.86) we see that $-2\log\Lambda$ is monotonic increasing with q, i.e., $\log\Lambda$ is monotonic decreasing with increasing q, which is just what we want. This algebra shows that the usual F-test in an analysis of covariance is the LRT for $H_0: \sigma_a^2 = 0$.

– iii. ***Testing*** $H_0: \beta = 0$

This is easy. First, $l^*(\hat{\boldsymbol{\theta}})$ stays as is, in (3.83). Second, under $H_0: \beta = 0$ we simply put $\hat{\beta} = 0$ in the estimators $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\sigma}_a^2$. Thus $\hat{\mu}$ of (3.58) becomes $\hat{\mu}_0 = \bar{y}..$, S_1 of (3.67) becomes SST and S_2 of (3.68) stays the same, $S_2 = \text{SSA}$. Also, with no β in the model, SSC of (3.73) becomes SSR = SSE. Therefore from (3.75) and (3.76)

$$\dot{\sigma}_0^2 + n\dot{\sigma}_{a,0}^2 = \text{SSA}/m \quad \text{and} \quad \hat{\sigma}_0^2 = \dot{\sigma}_0^2 = \text{SSE}/[m(n-1)].$$

The only effective change in all of this is that $l^*(\hat{\boldsymbol{\theta}}_0)$ will be $l^*(\hat{\boldsymbol{\theta}})$with SSR replaced by SSE. Doing this in (3.83) gives

$$-2l^*(\hat{\boldsymbol{\theta}}_0) \;=\; f_3(m,n) + m(n-1)\log\text{SSE} + m\log\text{SSA}. \tag{3.88}$$

and so, on subtracting (3.83) from (3.88) and ignoring f_1 and f_3,

$$\begin{aligned} -2\log\Lambda &= -m(n-1)\log\text{SSR} + m(n-1)\log\text{SSE} \\ &= -m(n-1)\log\frac{\text{SSR}}{\text{SSE}} \\ &= -m(n-1)\log\left(1 - \frac{\text{SSC}}{\text{SSE}}\right), \end{aligned} \tag{3.89}$$

from (3.73) where, in (3.72) SSC is $\hat{\beta}^2 m S_{xx}$, which is what is usually called SS(Regression). Thus (3.89) suggests what we will denote by q_β as the test statistic:

$$q_\beta = \frac{\text{SS(Regression)}}{\text{SSE}} = \frac{\hat{\beta}^2 S_{xx}}{(n-1)\text{MSE}}.$$

In point of fact the usual analysis of variance statistic is

$$\begin{aligned} F &= \frac{\text{SS(Regression)}/1}{\text{SS(Residual after fitting } \mu + a_i + \beta x_{ij})/[m(n-1)-1]} \\ &= \frac{[m(n-1)-1]\text{SS(Regression)}}{\text{SSE - SS(Regression)}} \\ &= \frac{kq_\beta}{1-q_\beta} \end{aligned} \tag{3.90}$$

where SS(Residual) has $k = m(n-1)-1$ degrees of freedom. And from (3.90)

$$q_\beta = \frac{F}{k+F}.$$

As in the previous Section (3.5d-ii) we see that, with balanced data, the usual analysis of variance F–test is the LRT of $H_0 : \beta = 0$.

e. Illustration

We return to the *Phytophthora* data of Figure 3.1 using all six weeks of data but only temperatures 15°C through 25°C. Since conditions vary from week to week, possibly causing lesion sizes to change, and since our goal would probably be to draw conclusions about a hypothetical population of experiments replicated over time, we treat the week-specific intercepts as random. Accordingly, we model the average log diameter (ALD) per leaflet as

$$\begin{aligned} \text{E}[\text{ALD}_{ij}|a_i] &= \mu_{ij} = \mu + a_i + \beta \text{TEMP}_{ij} \\ \text{ALD}_{ij}|a_i &\sim \text{indep. } \mathcal{N}(\mu_{ij}, \sigma^2), \\ a_i &\sim \text{i.i.d. } \mathcal{N}(0, \sigma_a^2), \end{aligned} \tag{3.91}$$

where $\text{ALD}_{ij} = y_{ij}$ is the average log diameter of the lesion, and $\text{TEMP}_{ij} = x_{ij}$ is the temperature, both being defined for the jth leaflet during the ith week.

Using SAS PROC MIXED (SAS Institute, 1998) the restricted maximum likelihood estimators are $\hat{\mu} = -1.818$, $\hat{\beta} = 0.170$, $\hat{\sigma}^2 = 0.219$ and $\hat{\sigma}_a^2 = 0.250$. So, for example, the estimate of β tells us that log diameter increases about 0.170 with each increase in temperature of one degree.

The value of $\hat{\sigma}_a^2$ indicates the magnitude of the variation of the weekly intercepts: They have a standard deviation of about 0.5. We can also use it to provide an estimate of the correlation of the observations taken in the same week. Using (2.18) the ML estimate of the correlation is

$$\begin{aligned} \widehat{\text{corr}}(y_{ij}, y_{ij'}) &= \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}^2} \\ &= \frac{0.250}{0.250 + 0.219} = 0.65, \end{aligned}$$

which is appreciably high. Is the correlation statistically significantly different from zero? We can test it by testing $H_0: \sigma_a^2 = 0$ versus $H_A: \sigma_a^2 > 0$. As in Snedecor and Cochran (1989) we form the F-statistic, $F = \text{MSA/MSR}$, which is equal to 7.85. The critical value is $\mathcal{F}^5_{30,0.95} = 2.54$, so the test easily rejects H_0 with a p-value, $\text{P}\{\mathcal{F}^5_{30} \geq 7.85\}$, which is less than 0.001.

f. Predicting the random intercepts

As a predictor of a we use, as in Section 2.4b, the best predictor, BP:

$$\text{BP}(\mathbf{a}) = \text{E}[\mathbf{a}|\mathbf{y}].$$

This is valid for all forms of probability distributions of $\mathbf{a}$ and $\mathbf{y}$ (see Section 9.2). In the model being considered here the individual vectors $\mathbf{y}_i$ are independent and the only information about a_i is that contained in $\bar{y}_{i\cdot}$. Therefore we consider just

$$\text{BP}(a_i) = \text{E}[a_i|\bar{y}_{i\cdot}]$$

and under the normality conditions of (3.32) and (3.34) this is

$$\begin{aligned} \text{BP}(a_i) &= \text{E}[a_i] + \text{cov}(a_i, \bar{y}_{i\cdot})[\text{var}(\bar{y}_{i\cdot})]^{-1}(\bar{y}_{i\cdot} - \text{E}[\bar{y}_{i\cdot}]) \\ &= \sigma_a^2\left(\sigma_a^2 + \sigma^2/n\right)^{-1}[\bar{y}_{i\cdot} - (\mu + \beta\bar{x}_{\cdot})] \\ &= \frac{n\sigma_a^2}{\sigma^2 + n\sigma_a^2}(\bar{y}_{i\cdot} - \mu - \beta\bar{x}_{\cdot}). \end{aligned} \tag{3.92}$$

This is very similar to $\text{BP}(a_i)$ of Section 2.4b–ii.

We now face a problem: how to convert the algebraic expression of (3.92) to a numerical value that can be of practical use? In other words, how can we estimate $\mathrm{BP}(a_i)$? Because it is a ratio of variances multiplying a linear function of μ and β, the derivation of an optimum estimator is undoubtedly difficult. Several alternative possibilities do exist. The easy part is to use $\hat{\mu}$ and $\hat{\beta}$ in place of μ and β. At least for balanced data, $\hat{\mu}$ and $\hat{\beta}$ do not involve σ^2 and σ_a^2, so no matter what we do about those variances, using $\hat{\mu}$ of (3.58) and $\hat{\beta}$ of (3.59) seems appropriate. Thus if σ^2 and σ_a^2 are known we could use, as an estimator of $\mathrm{BP}(a_i)$,

$$\begin{aligned} \mathrm{BP}^0(a_i) &= \frac{n\sigma_a^2}{\sigma^2 + n\sigma_a^2}\left(\bar{y}_{i\cdot} - \hat{\mu} - \hat{\beta}\bar{x}_{\cdot}\right) \\ &= \frac{n\sigma_a^2}{\sigma^2 + n\sigma_a^2}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \end{aligned} \tag{3.93}$$

after substituting for $\hat{\mu}$ and $\hat{\beta}$. The sampling variance would be

$$\mathrm{var}[\mathrm{BP}^0(a_i)] = \frac{(1 - 1/m)\sigma_a^4}{\sigma_a^2 + \sigma^2/n}. \tag{3.94}$$

If we do not know σ^2 and σ_a^2 we may be prepared to assume that some prior estimates are true values, in which case we could use them in (3.93) and (3.94).

But lacking true (or satisfactory prior) values for σ^2 and σ_a^2 we need to estimate them. Suppose we do this, using ML. Then, bearing in mind that under large-sample theory the ML estimator of a function of parameters represented by $\boldsymbol{\theta}$, say $f(\boldsymbol{\theta})$, is $f(\hat{\boldsymbol{\theta}})$ for $\hat{\boldsymbol{\theta}}$ being the ML estimator of $\boldsymbol{\theta}$, we calculate the ML estimate of $\mathrm{BP}(a_i)$ as

$$\widehat{\mathrm{BP}}(a_i) = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}^2/n}\left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right). \tag{3.95}$$

But if we use $\widehat{\mathrm{BP}}(a_i)$ of (3.95), goodness knows how we could derive its variance, especially if, as is often the case, the estimates of all four parameters have been obtained from the same data set. Moreover, the very form of $\widehat{\mathrm{BP}}(a_i)$ creates complications for ascertaining variances. For example, what is the variance of a ratio of estimated variance components, let alone of that ratio multiplied by a mean?

A practical way out of this predicament is to assume the variance components are known, leading to BP^0 of (3.93); derive its variance and

in that variance replace σ^2 and σ_a^2 by estimates (or assumed values) thereof. Thus use as the variance (3.94) with σ^2 and $\hat{\sigma}_a^2$ replaced by their estimates:

$$\widehat{\text{var}}\left[\text{BP}^0(a_i)\right] = \frac{(1-1/m)(\hat{\sigma}_a^2)^2}{\hat{\sigma}_a^2 + \hat{\sigma}^2/n}.$$

Properties of this are unknown—but at least it is a practical procedure. Of course, if $\hat{\sigma}_a^2 = 0$ every $\widehat{\text{BP}}(a_i)$ is the same, namely zero, consistent with $\hat{\sigma}_a^2 = 0$.

3.6 RANDOM INTERCEPTS: UNBALANCED DATA

The preceding section deals with balanced data, by which we mean that every data vector $\mathbf{y}_i$ for $i = 1, 2, \ldots, m$ contains n observations y_{ij} for $j = 1, 2, \ldots, n$. And corresponding to every $\mathbf{y}_i$ is the same vector $\mathbf{x}_0$ of the n x-values, x_j for $j = 1, 2, \ldots, n$. Now we deal with unbalanced data, which is the situation when for each i (i.e., each week of the example) there may be some y-values missing (i.e., at some temperatures data are missing). Thus $\mathbf{y}_i$ may contain fewer than n observations: we denote the number of observations in $\mathbf{y}_i$ by n_i. Likewise, corresponding to $\mathbf{y}_i$ there will be only n_i x-values. They will, of course, be n_i values from the $\mathbf{x}_0$ vector, but not necessarily the same n_i values even for two values of n_i that are the same. So now, instead of the balanced data case of every $\mathbf{y}_i$ of order n being associated with $\mathbf{x}_0$ of order n, each $\mathbf{y}_i$ of order n_i is associated with its own $\mathbf{x}_i$ of order n_i, its elements being n_i values occurring in $\mathbf{x}_0$. Table 3.1 shows some illustrative examples.

Notation

Elements of $\mathbf{y}_i$ are y_{ij} for $j = 1, 2, \ldots, n_i$. The n_i-values associated with elements of $\mathbf{y}_i$ are denoted x_{ij} for $j = 1, 2, \ldots, n_i$. But this use of j is for y_{ij} and x_{ij} being numbered consecutively from $j = 1$ through $j = n_i$ without regard for the value of j when it is used for x_j in $\mathbf{x}_0$ of (3.35) for the balanced data case. For example, in Table 3.1, the entry 6 in $\mathbf{x}_0$ is x_j for $j = 5$. But that same 6 in $\mathbf{x}_1$ is x_{13}, in $\mathbf{x}_2$ it is x_{24} and in $\mathbf{x}_4$ it is $\mathbf{x}_{43}$. And, of course, as these second subscripts indicate, the elements of $\mathbf{x}_i$ are written one after the other in the usual way, so that $\mathbf{x}_i$ is $n_i \times 1$; it is not $n \times 1$ with gaps or zeros as might be suggested by Table 3.1.

This particularly affects one's understanding of average x-values. With balanced data the average x-value was the same for every i, the average of all n elements of $\mathbf{x}_0$: it was denoted as $\bar{x}_{i\cdot}$. Now we have $\bar{x}_{i\cdot}$,

Table 3.1: Illustrative Examples of $\mathbf{x}_i$

Balanced Data	Unbalanced Data			
$\mathbf{x}_0$	$\mathbf{x}_1$ $n_1 = 5$	$\mathbf{x}_2$ $n_2 = 6$	$\mathbf{x}_3$ $n_3 = 4$	$\mathbf{x}_4$ $n_4 = 5$
		Elements of $\mathbf{x}_i$		
2	$x_{11} = 2$	$x_{21} = 2$		
3	$x_{12} = 3$		$x_{31} = 3$	$x_{41} = 3$
4		$x_{22} = 4$		
5		$x_{23} = 5$	$x_{32} = 5$	$x_{42} = 5$
6	$x_{13} = 6$	$x_{24} = 6$		$x_{43} = 6$
7	$x_{14} = 7$	$x_{25} = 7$	$x_{33} = 7$	
8	$x_{15} = 8$		$x_{34} = 8$	$x_{44} = 8$
9		$x_{26} = 9$		$x_{45} = 9$

the average of the n_i x-values in $\mathbf{x}_i$, namely $\bar{x}_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}/n_i$. And special care in this regard is needed in reducing results for unbalanced data to those for balanced data. For then, not only does $n_i = n$ and $\mathbf{x}_i = \mathbf{x}_0$ but also $\bar{x}_{i\cdot}$ becomes $\bar{x}_{\cdot\cdot}$.

a. The model

With $\mathbf{y}_i$ and $\mathbf{x}_i$ having order n_i we have, comparable to (3.36)

$$\begin{aligned} \mathrm{E}[\mathbf{y}_i|a_i] &= \mu\mathbf{1}_{n_i} + a_i\mathbf{1}_{n_i} + \beta\mathbf{x}_i \\ &= \mathbf{X}_i \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \mathbf{1}_{n_i} a_i, \end{aligned}$$

for

$$\mathbf{X}_i = [\mathbf{1}_{n_i} \quad \mathbf{x}_i].$$

Through steps similar to those leading up to (3.46) we now have

$$\mathrm{E}[\mathbf{y}|\mathbf{a}] = \mathbf{X} \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \mathbf{Z}\mathbf{a}$$

for

$$\mathbf{X} = \left\{ {}_c\,[\mathbf{1}_{n_i} \quad \mathbf{x}_i] \right\}_{i=1}^{m} \quad \text{and} \quad \mathbf{Z} = \left\{ {}_d\,\mathbf{1}_{n_i} \right\}_{i=1}^{m}. \tag{3.96}$$

For variance specifications we maintain the homoscedasticity of a_i in (3.32) and of $\mathbf{y}_i|a_i$ in (3.39) (with n and $\mathbf{x}_0$ replaced by n_i and $\mathbf{x}_i$) and so, akin to (3.40), have

$$\mathbf{V}_i = \text{var}(\mathbf{y}_i) = \sigma^2\mathbf{I}_{n_i} + \sigma_a^2\mathbf{J}_{n_i}.$$

Notation

For notational simplification and clarity in this section we make the changes:

$$\mathbf{I}_i \equiv \mathbf{I}_{n_i} \qquad \text{and} \qquad \mathbf{J}_i \equiv \mathbf{J}_{n_i},$$

$$\tau_i = \frac{n_i\sigma_a^2}{\sigma^2 + n_i\sigma_a^2} \qquad \text{and} \qquad 1 - \tau_i = \frac{\sigma^2}{n_i\sigma_a^2}\tau_i,$$

and, for example

$$\left\{_d \quad \right\} \qquad \text{for} \qquad \left\{_d \quad \right\}_{i=1}^{m}.$$

Thus we write

$$\mathbf{V}_i = \sigma^2\mathbf{I}_i + \sigma_a^2\mathbf{J}_i$$

with

$$\mathbf{V}_i^{-1} = \frac{1}{\sigma^2}\left(\mathbf{I}_i - \frac{\sigma_a^2}{\sigma^2 + n_i\sigma_a^2}\mathbf{J}_i\right) = \frac{1}{\sigma^2}\left(\mathbf{I}_i - \tau_i\bar{\mathbf{J}}_i\right), \qquad (3.97)$$

similar to (3.44). Then

$$\mathbf{V} = \text{var}(\mathbf{y}) = \text{var}\left(\left\{_c\, \mathbf{y}_i\right\}\right) = \left\{_d\, \mathbf{V}_i\right\}. \qquad (3.98)$$

b. Estimating μ and β when variances are known

– i. *ML estimators*

As in (3.51) we need $\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ for estimating μ and β. Thus from (3.96), (3.97) and (3.98) we get

$$\mathbf{X}'\mathbf{V}^{-1} = \left\{_r \left[\begin{array}{c} \mathbf{1}_i' \\ \mathbf{x}_i' \end{array}\right]\right\} \frac{1}{\sigma^2}\left\{_d\, \mathbf{I}_i - \tau_i\bar{\mathbf{J}}_i\right\}$$

$$= \left\{ \begin{bmatrix} (\tau_i/n_i\sigma_a^2)\mathbf{1}_i' \\ (\mathbf{x}_i' - \tau_i\bar{x}_{i\cdot}\mathbf{1}_i')/\sigma^2 \end{bmatrix} \right\}_r .$$

Also

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{X}'\mathbf{V}^{-1}\left\{ [\mathbf{1}_i \quad \mathbf{x}_i] \right\}_c$$

$$= \frac{1}{\sigma_a^2}\begin{bmatrix} \Sigma\tau_i & \Sigma\tau_i\bar{x}_{i\cdot} \\ \Sigma\tau_i\bar{x}_{i\cdot} & \frac{\sigma_a^2}{\sigma^2}\mathrm{SSE}_{xx} + \Sigma\tau_i\bar{x}_{i\cdot}^2 \end{bmatrix} \tag{3.99}$$

after adding and subtracting $\sum_i n_i\bar{x}_{i\cdot}^2/\sigma^2$ in the (2,2) element to get

$$\mathrm{SSE}_{xx} = \sum_i \left(\sum_j x_{ij}^2 - n_i\bar{x}_{i\cdot}^2 \right) = \sum_{i=1}^{m}\sum_{j=1}^{n} (x_{ij} - \bar{x}_{i\cdot})^2 .$$

And, with algebra similar to that used for deriving $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$,

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \frac{1}{\sigma_a^2}\begin{bmatrix} \Sigma\tau_i\bar{y}_{i\cdot} \\ \frac{\sigma_a^2}{\sigma^2}\mathrm{SSE}_{xy} + \Sigma_i\tau_i\bar{x}_{i\cdot}\bar{y}_{i\cdot} \end{bmatrix} . \tag{3.100}$$

For the moment, write

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \frac{1}{\sigma_a^2}\mathbf{B} \quad \text{for} \quad \mathbf{B} = \begin{bmatrix} p & q \\ q & r \end{bmatrix} = \sigma_a^2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} . \tag{3.101}$$

Then

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma_a^2\mathbf{B}^{-1} = \frac{\sigma_a^2}{|\mathbf{B}|}\begin{bmatrix} r & -q \\ -q & p \end{bmatrix}$$

with, from comparing (3.99) and (3.101),

$$|\mathbf{B}| = pr - q^2 = \sum_i \tau_i \left(\frac{\sigma_a^2}{\sigma^2}\mathrm{SSE}_{xx} + \sum_i \tau_i\bar{x}_{i\cdot}^2 - \frac{(\Sigma\tau_i\bar{x}_{i\cdot})^2}{\Sigma\tau_i} \right) .$$

Thus

$$\begin{bmatrix} \hat{\mu} \\ \hat{\beta} \end{bmatrix} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

$$= \frac{\sigma_a^2}{|\mathbf{B}|} \begin{bmatrix} \frac{\sigma_a^2}{\sigma^2}\mathrm{SSE}_{xx} + \Sigma\tau_i\bar{x}_{i\cdot}^2 & -\Sigma\tau_i\bar{x}_{i\cdot} \\ -\Sigma\tau_i\bar{x}_{i\cdot} & \Sigma\tau_i \end{bmatrix}$$

$$\times \begin{bmatrix} \Sigma\tau_i\bar{y}_{i\cdot}/\sigma_a^2 \\ \mathrm{SSE}_{xy}/\sigma^2 + \Sigma\tau_i\bar{x}_{i\cdot}\bar{y}_{i\cdot}/\sigma_a^2 \end{bmatrix}. \quad (3.102)$$

This gives

$$\hat{\beta} = \frac{\sigma_a^2}{|\mathbf{B}|}\left[\frac{-\sum\tau_i\bar{x}_{i\cdot}\sum\tau_i\bar{y}_{i\cdot}}{\sigma_a^2} + \frac{\sum\tau_i\mathrm{SSE}_{xy}}{\sigma^2} + \frac{\sum\tau_i\bar{x}_{i\cdot}\bar{y}_{i\cdot}}{\sigma_a^2}\right] \quad (3.103)$$

$$= \frac{\sigma_a^2}{|\mathbf{B}|}\sum\tau_i\left[\frac{\mathrm{SSE}_{xy}}{\sigma^2} + \frac{1}{\sigma_a^2}\left(\sum\tau_i\bar{x}_{i\cdot}\bar{y}_{i\cdot} - \frac{\Sigma\tau_i\bar{x}_{i\cdot}\Sigma\tau_i\bar{y}_{i\cdot}}{\Sigma\tau_i}\right)\right].$$

$$= \frac{(\sigma_a^2/\sigma^2)\mathrm{SSE}_{xy} + \mathrm{WSSA}_{xy}}{(\sigma_a^2/\sigma^2)\mathrm{SSE}_{xx} + \mathrm{WSSA}_{xx}}, \quad (3.104)$$

where WSSA is for *weighted sum of squares* in the sense of

$$\mathrm{WSSA}_{xx} = \sum\tau_i\bar{x}_{i\cdot}^2 - \frac{(\Sigma\tau_i\bar{x}_{i\cdot})^2}{\Sigma\tau_i} = \sum\tau_i(\bar{x}_{i\cdot} - \tilde{x}_{\cdot\cdot})^2$$

for $\tilde{x}_{\cdot\cdot}$ of (3.105). WSSA_{xy} is defined similarly. $\hat{\mu}$ can also be derived from (3.102). Tedious algebra (see E 3.8) which includes the use of weighted means

$$\tilde{y}_{\cdot\cdot} = \frac{\Sigma_i\tau_i\bar{y}_{i\cdot}}{\Sigma\tau_i} \quad \text{and} \quad \tilde{x}_{\cdot\cdot} = \frac{\Sigma_i\tau_i\bar{x}_{i\cdot}}{\Sigma_i\tau_i}, \quad (3.105)$$

ultimately reduces to

$$\hat{\mu} = \tilde{y}_{\cdot\cdot} - \hat{\beta}\tilde{x}_{\cdot\cdot}. \quad (3.106)$$

Notationally this is similar to previous results but with, of course, using the weighted means and the not-so-simple expression for $\hat{\beta}$ in (3.104).

These expressions for $\hat{\mu}$ and $\hat{\beta}$ are ML estimators provided that σ^2 and σ_a^2 are known. And using those known values together with the x- and y-values of the data gives $\hat{\mu}$ and $\hat{\beta}$ as ML estimates.

– ii. *Unbiasedness*

Using the same procedure as in Section 3.5b–i, it is not difficult to show for $\mathrm{E}[\hat{\beta}]$ that

$$\mathrm{E}[\mathrm{SSE}_{xy}] = \beta(\mathrm{SSE}_{xx}) \qquad \text{and} \qquad \mathrm{E}[\mathrm{WSSA}_{xy}] = \beta(\mathrm{WSSA}_{xx})$$

and so $\hat{\beta}$ is unbiased. And then the unbiasedness of $\hat{\mu}$ is easily established (see E 3.9).

– iii. *Sampling variances*

Some solid algebra (see E 3.11), assuming σ^2 and σ_a^2 are known, yields

$$\mathrm{var}(\hat{\beta}) = \frac{1}{\dfrac{\mathrm{SSE}_{xx}}{\sigma^2} + \dfrac{\mathrm{WSSA}_{xx}}{\sigma_a^2}} \tag{3.107}$$

and

$$\begin{aligned} \mathrm{var}(\hat{\mu}) &= \frac{\sigma_a^2}{\Sigma\tau_i} + \tilde{x}^2\mathrm{var}(\hat{\beta}) \\ &= \sum_i(\sigma^2/n_i + \sigma_a^2) + \tilde{x}_{\cdot\cdot}^2\mathrm{var}(\hat{\beta}). \end{aligned} \tag{3.108}$$

– iv. *Predicting a_i*

The only change from Section 3.5e is that everywhere n occurs it is replaced by n_i. Thus

$$\mathrm{BP}(a_i) = \frac{n_i\sigma_a^2}{\sigma^2 + n_i\sigma_a^2}\,(\bar{y}_{i\cdot} - \mu - \beta\bar{x}_{i\cdot})\,,$$

and everything follows from this just as in Section 3.5 but with $\hat{\mu}$ and $\hat{\beta}$ of (3.106) and (3.104) in place of μ and β.

3.7 BERNOULLI - LOGISTIC REGRESSION

Consider the example of Section 3.1: We are interested in the growth of *Phytophthora infestans sporangia* lesions as a function of x, the temperature. Suppose our response, y, is the presence (coded as $y = 1$) or absence (coded as $y = 0$) of certain diameter growth. This would be much easier to judge than measuring the radius of colony growth.

What sort of model can we reasonably hypothesize for y as a function of x? Since y is binary it must follow a Bernoulli distribution. Since $\mathrm{E}[y]$ will be modeled as (and vary as) a function of x and since $\mathrm{var}(y) = \mathrm{E}[y](1 - \mathrm{E}[y])$ the variance cannot be assumed constant. Further, since $\mathrm{E}[y] = \mathrm{P}\{y = 1\}$, the mean must be bounded between zero and one. Therefore $\mathrm{E}[y]$ cannot be assumed to be linear in x unless it is modeled over only a short range of x. Otherwise it would lead to values of $\mathrm{E}[y]$ not in the interval (0,1). Alas, three of the four assumptions (all except independence) of the simple linear regression model of Section 3.2a cannot be used for binary data.

One way to deal with the range restriction inherent in $\mathrm{E}[y]$ is in the same parsimonious fashion as in Section 2.6c. Instead of modeling $\mathrm{E}[y]$ directly we instead model $\mathrm{logit}(\mathrm{E}[y]) = \log\left(\dfrac{\mathrm{E}[y]}{1 - \mathrm{E}[y]}\right)$.

a. Logistic regression model

The preceding discussion is motivation for the widely used logistic regression model:

$$\mathrm{E}[y_i] = \pi(x_i) = \frac{1}{1 + e^{-(\alpha + \beta x_i)}} \tag{3.109}$$

or equivalently

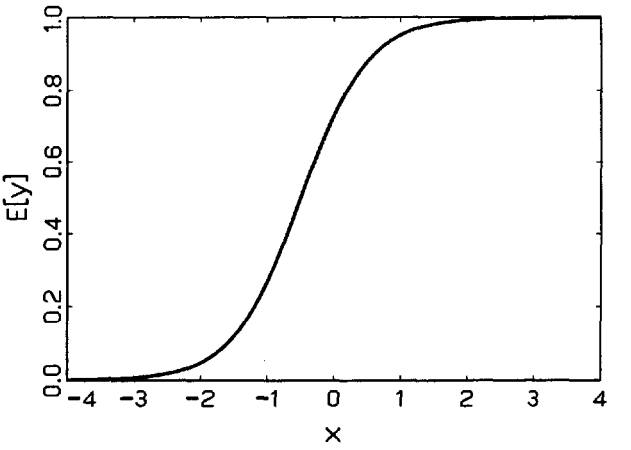
$$\mathrm{logit}(\mathrm{E}[y_i]) = \mathrm{logit}[\pi(x_i)] = \alpha + \beta x_i$$

$$y_i \sim \text{indep. Bernoulli}[\pi(x_i)].$$

In (3.109) we have again used the notation π for the mean of y since it is a probability.

How is this model different from (3.1)? Clearly we are assuming a different distribution for y_i. Also, by modeling $\mathrm{logit}(\mathrm{E}[y_i])$ as linear in x_i we are, in fact, hypothesizing a nonlinear model for $\mathrm{E}[y_i]$ as given in

Figure 3.2: Plot of E[y] for the Logistic Regression Model (3.109) with $\alpha = 1$ and $\beta = 2$.



(3.109). As an example of the form of E[y], Figure 3.2 shows a plot of E[y] as x ranges from -4 to 4 when $\alpha = 1$ and $\beta = 2$.

Several comments about the form of E[y_i] are in order to understand more fully the logistic regression model. The equation describes a regression line that is always (when viewed over a wide enough range of x and as long as β is not zero) an S-shaped curve as demonstrated in Figure 3.2. It is increasing if β is greater than zero, decreasing if β is less than zero and flat if β is equal to zero. Thus β governs how quickly the curve increases or decreases whereas α governs its horizontal location. The curve reaches its half height (of 0.5) when $\alpha + \beta x = 0$ or, equivalently, when $x = -\alpha/\beta$.

Another primary interpretation of β is related to the idea of odds. If the probability of an event is π, then the *odds* of the event is defined as $\pi/(1-\pi)$. For example, if the probability of an event is 1/3, then its odds are $\frac{1/3}{2/3} = 1/2$ or the odds are 1 to 2; if the probability is 3/4, then the odds of the event are $\frac{3/4}{1/4} = 3$. Thus, another way to state (3.109) is that the log of the odds of the event P$\{y_i = 1\}$ is $\alpha + \beta x_i$ or that the odds of a success are $e^{\alpha+\beta x_i}$.

A common way to interpret β in the linear regression model (3.1) is to consider how much E[y] changes when x is increased by a single unit. For that model we get the simple result that the difference in expected values is just $[\alpha + \beta(x+1)] - [\alpha + \beta x] = \beta$.

The equivalent calculation for (3.109) is that

$$\begin{aligned}\beta &= \text{logit}[\pi(x+1)] - \text{logit}[\pi(x)] \\ &= \log[\text{odds}(x+1)] - \log[\text{odds}(x)] \\ &= \log\left[\frac{\text{odds}(x+1)}{\text{odds}(x)}\right]\end{aligned}$$

or

$$e^{\beta} = \frac{\text{odds}(x+1)}{\text{odds}(x)},$$

where $\text{odds}(x) \equiv \pi(x)/[1-\pi(x)]$. This result is described by saying that β is the *log odds ratio* or that e^{β} is the *odds ratio.*

While the formulation (3.109) is quite standard and leads to easy interpretations of α and β in the scale of $\text{logit}[\pi(x_i)]$, for which it is linear in x, it does not give a straightforward interpretation on the π scale. For ease of interpretation $\text{E}[y_i]$ is sometimes reparameterized in terms of $x_h = -\alpha/\beta$, the halfway point on the x-axis, and γ, the slope of the curve at x_h. In this parameterization

$$\text{E}[y_i] = \frac{1}{1+e^{-4\gamma(x_i - x_h)}}. \tag{3.110}$$

b. Likelihood

Since the y_i are independent and Bernoulli distributed, the likelihood is straightforward to evaluate:

$$\begin{aligned}L &= \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1-\pi(x_i)]^{1-y_i} \\ &= \prod_{i=1}^{n} \{\pi(x_i)/[1-\pi(x_i)]\}^{y_i} [1-\pi(x_i)]. \end{aligned} \tag{3.111}$$

Using

$$\pi(x_i)/[1-\pi(x_i)] = e^{\alpha+\beta x_i}$$

and

$$1-\pi(x_i) = (1+e^{\alpha+\beta x_i})^{-1}$$

gives L as

$$L = \prod_{i=1}^{n} e^{y_i(\alpha+\beta x_i)}(1+e^{\alpha+\beta x_i})^{-1} \tag{3.112}$$

and the log likelihood as

$$l = \log L = \sum_{i=1}^{n} y_i(\alpha+\beta x_i) - \log(1+e^{\alpha+\beta x_i}). \tag{3.113}$$

We immediately see an advantage of assuming the logit of $\pi(x)$ to be linear in x—it yields an extremely simple log likelihood, in (3.113).

c. ML equations

Differentiating (3.113) with respect to α and β gives

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \sum_{i=1}^{n}\left[y_i - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right] \\
&= \sum_{i=1}^{n}\left[y_i - \frac{1}{1+e^{-(\alpha+\beta x_i)}}\right] \\
&= \sum_{i=1}^{n}[y_i - \pi(x_i)] \qquad (3.114) \\
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{n}\left[x_i y_i - \frac{x_i e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right] \\
&= \sum_{i=1}^{n} x_i\left[y_i - \pi(x_i)\right]. \qquad (3.115)
\end{aligned}$$

Noting that $\pi(x_i) = \mathrm{E}[y_i]$ we can see that setting (3.114) and (3.115) equal to zero gives exactly the same equations as (3.6) and (3.7) that were derived for the simple linear regression model of Section 3.2. Unfortunately, they are not as easy to solve.

After equating (3.114) and (3.115) to zero we need to solve the (nonlinear in $\hat{\alpha}$ and $\hat{\beta}$) equations

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{1}{1+e^{-(\hat{\alpha}+\hat{\beta}x_i)}}$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \frac{x_i}{1 + e^{-(\hat{\alpha}+\hat{\beta}x_i)}}. \tag{3.116}$$

The first equation has a straightforward interpretation: the ML solutions are chosen so that the total predicted number of successes is equal to $\sum_i y_i$, the total observed number of successes. Except in some special cases (e.g E 3.4), (3.116) does not have an explicit solution and must be solved numerically.

The second derivatives of l take a convenient form:

$$\begin{aligned} \frac{\partial^2 l}{\partial \alpha^2} &= -\sum_{i=1}^{n} \frac{\partial \pi(x_i)}{\partial \alpha} \\ &= -\sum_{i=1}^{n} \frac{e^{-(\alpha+\beta x_i)}}{\left(1 + e^{-(\alpha+\beta x_i)}\right)^2}, \\ &= -\sum_{i=1}^{n} \pi(x_i)[1 - \pi(x_i)] \end{aligned} \tag{3.117}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \alpha \, \partial \beta} &= -\sum_{i=1}^{n} \frac{x_i e^{-(\alpha+\beta x_i)}}{\left(1 + e^{-(\alpha+\beta x_i)}\right)^2} \\ &= -\sum_{i=1}^{n} x_i \pi(x_i)[1 - \pi(x_i)] \end{aligned} \tag{3.118}$$

and

$$\frac{\partial^2 l}{\partial \beta^2} = -\sum_{i=1}^{n} x_i^2 \pi(x_i)[1 - \pi(x_i)]. \tag{3.119}$$

With $\mathbf{V} = \text{var}(\mathbf{y}) = \text{diag}\{\pi(x_i)[1-\pi(x_i)]\}$ and $\mathbf{X}' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}$. then we can compactly write the information matrix as

$$-\text{E}\begin{bmatrix} l_{\alpha\alpha} & l_{\alpha\beta} \\ l_{\beta\alpha} & l_{\beta\beta} \end{bmatrix} = \text{E}\left[\mathbf{X}'\mathbf{V}\mathbf{X}\right] = \mathbf{X}'\mathbf{V}\mathbf{X} \tag{3.120}$$

which shows that the large-sample variance of $\hat{\alpha}$ and $\hat{\beta}$ is $(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$.

This also yields a convenient computing algorithm (see Chapter 10) for finding $\hat{\alpha}$ and $\hat{\beta}$. Since the Hessian is negative definite, the log likelihood is concave. Hence, except in rare cases (see E 3.6), a single local maximum exists (and is the global maximum) and the Newton–Raphson algorithm described below is guaranteed to converge to the MLEs (Santner and Duffy, 1990). The algorithm proceeds as follows, with m denoting the iteration number and superscripts indicating sequential values of the parameters:

1. Obtain starting values $\alpha^{(0)}$ and $\beta^{(0)}$. Set $m = 0$.

2. Calculate

$$\begin{pmatrix} \alpha^{(m+1)} \\ \beta^{(m+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(m)} \\ \beta^{(m)} \end{pmatrix} + (\mathbf{X}'\mathbf{V}^{(m)}\mathbf{X})^{-1}\mathbf{X}'[\mathbf{y} - \boldsymbol{\pi}^{(m)}(\mathbf{x})].$$

3. Check for convergence of $\begin{pmatrix} \alpha^{(m+1)} \\ \beta^{(m+1)} \end{pmatrix}$. If it has converged, stop; otherwise set $m = m + 1$ and return to step 2.

In this algorithm $\boldsymbol{\pi}^{(m)}(\mathbf{x})$ is the notation we use for the vector $\left\{{}_c\, 1/(1 + e^{-(\alpha^{(m)}+\beta^{(m)}x_i)})\right\}_{i=1}^{n}$, and $\mathbf{V}^{(m)} = \operatorname{diag}\left\{\boldsymbol{\pi}^{(m)}(\mathbf{x})[1 - \boldsymbol{\pi}^{(m)}(\mathbf{x})]\right\}$.

d. Large-sample tests and intervals

As in Section 2.6b–v, large-sample tests and confidence intervals can be based on the asymptotic normality ($\mathcal{AN}$) of $\hat{\alpha}$ and $\hat{\beta}$,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{AN}\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\right],$$

where $\mathbf{X}'\mathbf{V}\mathbf{X}$ is given in (3.117) through (3.120). For example, to test $H_0: \beta \leq 0$ versus $H_A: \beta > 0$ we would reject H_0 if

$$\frac{\hat{\beta}}{\sqrt{\widehat{\operatorname{var}}(\hat{\beta})}} > z_\alpha, \tag{3.121}$$

where $\widehat{\operatorname{var}}(\hat{\beta})$ comes from inserting the MLEs into the lower-right-hand entry of $(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$. A large-sample confidence interval for β would be calculated as

$$\hat{\beta} \pm z_{\alpha/2}\sqrt{\widehat{\operatorname{var}}(\hat{\beta})}. \tag{3.122}$$

Similarly, the large-sample confidence interval for the odds ratio, e^{β}, would be

$$\left(e^{\hat{\beta}-z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta})}}, e^{\hat{\beta}+z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta})}}\right). \tag{3.123}$$

Alternatively, we can use the likelihood ratio test to test the two-sided hypothesis $H_0: \beta = 0$ versus $H_A: \beta \neq 0$. Under H_0 the likelihood becomes

$$L = \prod_{i=1}^{n} e^{y_i \alpha}(1 + e^{\alpha}) \tag{3.124}$$

with maximum $\hat{\alpha}_0 = \log[\bar{y}/(1-\bar{y})]$. Hence the maximized value of $l = \log L$ under H_0 is $\sum y_i \log \bar{y} + \sum(1-y_i)\log(1-\bar{y})$. The likelihood ratio statistic is then

$$\begin{aligned} -2\log\Lambda &= -2\Big[\sum y_i \log \bar{y} + \sum(1-y_i)\log(1-\bar{y}) \\ &\quad - \sum y_i(\hat{\alpha} + \hat{\beta}x_i) + \sum \log(1 + e^{\hat{\alpha}+\hat{\beta}x_i})\Big] \end{aligned} \tag{3.125}$$

and the test is to reject H_0 whenever $-2\log\Lambda$ exceeds $\chi^2_{1,1-\alpha}$.

3.8 BERNOULLI - LOGISTIC WITH RANDOM INTERCEPTS

Now consider our example from Section 3.7, but recall that the experiment has been repeated at six different times. We want to analyze all the data together but suspect that the probability of a lesion for a fixed inoculation level varied from time to time. We might hypothesize a model with a common "slope" parameter β but "intercepts" which varied from time to time. Since our goal would probably be to draw conclusions about all the experiments that could be replicated over time, we would want the intercepts to be a random effect, as in Section 3.6e.

a. Model

A reasonable model for a success for observation j in experiment i would therefore be the following:

$$\text{E}[y_{ij}|a_i] = \pi(x_{ij}) = \frac{1}{1 + e^{-(\alpha + a_i + \beta x_{ij})}} \tag{3.126}$$

or, equivalently,

$$\text{logit}[\pi(x_{ij})] = \alpha + a_i + \beta x_{ij}$$

and

$$y_{ij}|a_i \sim \text{indep. Bernoulli}[\pi(x_{ij})]$$

$$a_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma_a^2).$$

Conditional on a_i the y_{ij} follow a logistic regression model with intercepts that vary with the index i (with time in our example). Thus, conditional on a_i, β has the same interpretation as in the usual logistic regression model.

Since the a_i are random effects, there are two important differences between (3.126) and (3.109). First, the y_{ij} do not follow a logistic model marginally. This is because

$$\begin{aligned} \mathrm{E}[y_{ij}] &= \mathrm{E}\left[\mathrm{E}[y_{ij}|a_i]\right] \\ &= \mathrm{E}[\pi(x_{ij})] \\ &= \mathrm{E}\left[\frac{1}{1+e^{-(\alpha+a_i+\beta x_{ij})}}\right] \\ &= \int_{-\infty}^{\infty} \frac{1}{1+e^{-(\alpha+a_i+\beta x_{ij})}} \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2\sigma_a^2}a_i^2} da_i \end{aligned} \quad (3.127)$$

cannot be evaluated in closed form, and, in particular, is not of the logistic form, i.e., $1/(1+e^{-(\alpha^*+\beta^* x_{ij})})$, for some choice of α^* and β^*. However, it can be well approximated by a marginal logistic model:

$$\text{logit}\left(\mathrm{E}[y_{ij}]\right) \approx \alpha^* + \beta^* x_{ij},$$

where

$$\alpha^* = \frac{\alpha}{\sqrt{1+\lambda\sigma_a^2}} \quad \text{and} \quad \beta^* = \frac{\beta}{\sqrt{1+\lambda\sigma_a^2}}, \quad (3.128)$$

for $\lambda = 256/75\pi$ (see E 3.8).

Second, the y_{ij} and y_{ik} are correlated since they both involve the same random effect a_i. Using (1.16) we obtain

$$\text{cov}(y_{ij}, y_{ik}) = \text{cov}(\mathrm{E}[y_{ij}|a_i], \mathrm{E}[y_{ik}|a_i]) + \mathrm{E}\left[\text{cov}(y_{ij}, y_{ik}|a_i)\right]$$

$$= \operatorname{cov}\left(\frac{1}{1+e^{-(\alpha+a_i+\beta x_{ij})}}, \frac{1}{1+e^{-(\alpha+a_i+\beta x_{ik})}}\right) + 0$$

$$= \int_{-\infty}^{\infty} \frac{1}{1+e^{-(\alpha+a_i+\beta x_{ij})}} \frac{1}{1+e^{-(\alpha+a_i+\beta x_{ik})}} \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2\sigma_a^2}a_i^2} da_i$$

$$> 0, \text{ as long as } \sigma_a^2 > 0. \tag{3.129}$$

b. Likelihood

The likelihood can be calculated in the usual manner by writing the density conditional on the random effects as in (3.112) and then integrating them out:

$$L = \prod_{i=1}^{m} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} e^{(\alpha+a_i+\beta x_{ij})y_{ij}} (1+e^{\alpha+a_i+\beta x_{ij}})^{-1} \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2\sigma_a^2}a_i^2} da_i$$

$$= e^{\alpha y_{\cdot\cdot} + \beta \sum_{i,j} x_{ij} y_{ij}} \times \prod_{i=1}^{m} \int_{-\infty}^{\infty} e^{a_i y_{i\cdot}} \prod_{j=1}^{n_i} (1+e^{\alpha+a_i+\beta x_{ij}})^{-1} \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2\sigma_a^2}a_i^2} da_i$$

giving a log likelihood of

$$l = \alpha y_{\cdot\cdot} + \beta \sum_{i,j} x_{ij} y_{ij} + \sum_{i=1}^{m} \log \int_{-\infty}^{\infty} e^{a_i y_{i\cdot}} \prod_{j=1}^{n_i} (1+e^{\alpha+a_i+\beta x_{ij}})^{-1} \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2\sigma_a^2}a_i^2} da_i. \tag{3.130}$$

As in Section 2.6c–ii this must be evaluated and maximized numerically. Gauss–Hermite quadrature is again a logical method.

c. Large-sample tests and intervals

Given the intractibility of the log likelihood in (3.130), calculation of the information matrix for tests or confidence intervals is difficult and numerical at best.

Likelihood ratio tests can be performed by numerically maximizing (3.130) and log likelihoods of reduced models and calculating the negative of twice the difference between their maximized values.

d. Prediction

Conceptually, the estimated best predictor is given by

$$\tilde{a}_i = \hat{\mathrm{E}}[a_i|\bar{y}_{i\cdot}]. \tag{3.131}$$

However, as in the preceding section, closed-form expressions do not exist and are numerically difficult to compute.

e. Conditional Inference

If interest focused solely on β of (3.126) as opposed to α, σ_a^2, or the a_i, then another approach is available for inference. Suppose we rewrite model (3.126) as

$$\mathrm{E}[y_{ij}] = \pi(x_{ij}) = \frac{1}{1+e^{-(\alpha_i+\beta x_{ij})}}, \tag{3.132}$$

(using $\alpha_i = \alpha + a_i$) and make no assumptions about the α_i. Writing the log likelihood gives

$$\begin{aligned} l &= \sum_{i,j} y_{ij}(\alpha_i + \beta x_{ij}) - \log(1 + e^{\alpha_i+\beta x_{ij}}) \\ &= \sum_i \alpha_i y_{i\cdot} + \beta \sum_{i,j} y_{ij}x_{ij} - \sum_{i,j} \log(1 + e^{\alpha_i+\beta x_{ij}}), \end{aligned} \tag{3.133}$$

from which the sufficient statistics are $y_{1\cdot}, y_{2\cdot}, \ldots, y_{m\cdot}$, and $\sum_{i,j} y_{ij}x_{ij}$.

One reason for using random effects models is that it is known (Neyman and Scott, 1948) that if we let the sample size increase by letting $m \to \infty$ and we try to estimate $\alpha_1, \alpha_2, \ldots, \alpha_m$, and β by maximum likelihood then inconsistent MLEs can result.

Standard theory (Lehmann, 1986, Sec. 4.4) is to consider the conditional distribution of the sufficient statistic "associated" with the parameter of interest conditional on all the others. Conditioning on the sufficient statistic removes dependence on the remaining "nuisance" parameters. Such a methodology leads to tests which, for the marginal problem, are uniformly most powerful unbiased.

Applied in our situation, where we are assuming β is the parameter of interest, leads to consideration of the conditional distribution of $T = \sum_{i,j} y_{ij}x_{ij}$ given $S_1 = y_{1\cdot}, S_2 = y_{2\cdot}, \ldots, S_m = y_{m\cdot}$. We start with their joint distribution.

Since the y_{ij} are discrete, to calculate the probability that $S_1 = s_1$, $S_2 = s_2, \ldots, S_m = s_m$ and $T = t$, we merely sum over the y_{ij} that give $y_{1\cdot} = s_1$, $y_{2\cdot} = s_2, \ldots, y_{m\cdot} = s_m$, and $\sum_{i,j} y_{ij}x_{ij} = t$:

$$\mathrm{P}\{S_1 = s_1, S_2 = s_2, \ldots, S_m = s_m, T = t\} = \sum_R \frac{e^{\sum \alpha_i y_{i\cdot} + \beta \sum y_{ij}x_{ij}}}{\prod_1^m \prod_1^n (1 + e^{\alpha_i + \beta x_{ij}})}, \tag{3.134}$$

where $R = \{y_{ij} : y_{1\cdot} = s_1, \ldots, y_{m\cdot} = s_m, \sum y_{ij}x_{ij} = t\}$. This is equal to

$$C(s_1, \ldots, s_m, t) \frac{e^{\sum \alpha_i s_i + \beta t}}{\prod_1^m \prod_1^n (1 + e^{\alpha_i + \beta x_{ij}})}, \tag{3.135}$$

where $C(s_1, \ldots, s_m, t)$ is the number of combinations of the y_{ij} that are in R. To find the marginal distribution of $S_1, S_2, \ldots, S_m$ we sum out T:

$$f_{\mathbf{S}}(\mathbf{s}) = \sum_z C(s_1, \ldots, s_m, z) \frac{e^{\sum \alpha_i s_i + \beta z}}{\prod_1^m \prod_1^n (1 + e^{\alpha_i + \beta x_{ij}})}. \tag{3.136}$$

Then

$$\begin{aligned} f_{T|\mathbf{S}}(t|\mathbf{s}) &= \frac{C(s_1, \ldots, s_m, t) e^{\sum \alpha_i s_i + \beta t}}{\sum_z C(s_1, \ldots, s_m, z) e^{\sum \alpha_i s_i + \beta z}} \\ &= \frac{C(s_1, \ldots, s_m, t) e^{\beta t}}{\sum_z C(s_1, \ldots, s_m, z) e^{\beta z}}, \end{aligned} \tag{3.137}$$

which is independent of the α_i, as promised by sufficiency. This can be used to form tests or calculate estimates.

For example, to test $H_0 : \beta \leq 0$ versus $H_A : \beta > 0$, we use the null hypothesis distribution of $T|\mathbf{S}$, namely

$$P_{H_0}\{T = t|\mathbf{S} = \mathbf{s}\} = \frac{C(s_1, \ldots, s_m, t)}{\sum_z C(s_1, \ldots, s_m, z)}, \tag{3.138}$$

which depends only on the combinatorial coefficients and on no unknown parameters. The p-value corresponding to an observed value, $t_0 = \sum y_{ij}x_{ij}$, is

$$p = P_{H_0}\{T \geq t_0|\mathbf{S} = \mathbf{s}\} = \frac{\sum_{t \geq t_0} C(s_1, \ldots, s_m, t)}{\sum_z C(s_1, \ldots, s_m, z)}. \tag{3.139}$$

Clearly we need to know the values of $C(s_1, \ldots, s_m, t)$ to perform the calculation. Conceptually this is a simple counting task, but from a practical point of view the task can get tedious or, for larger problems, insurmountable. Specialized software (e.g., Mehta and Patel, 1992) has been written to efficiently compute these quantities for small and moderate-sized problems.

A conditional MLE for β can be calculated by maximizing (3.137).

3.9 EXERCISES

E 3.1 Under model (3.29) show that the standard deviation of y is proportional to its mean.

E 3.2 Showing all intermediate steps, derive from details given in Section 3.5b–i, the $\hat{\mu}$ and $\hat{\beta}$ of (3.58) and (3.59).

E 3.3 Derive (3.65), and from that derive (3.75) and (3.76).

E 3.4 Explain why the LR statistic for $H: \sigma_a^2 = 0$ is 1.0 when $\hat{\sigma}_a^2 = 0$.

E 3.5 Derive (3.94) and develop expressions for the covariance of $\mathrm{BP}^0(a_i)$ and $\mathrm{BP}^0(a_k)$; and for the variance of $\mathrm{BP}^0(a_i) - a_i$.

E 3.6 Show that the covariances of $\hat{\mu}$ with $\mathrm{BP}^0(a_i)$, of $\hat{\beta}$ with $\bar{y}_{i\cdot}$, and of $\hat{\beta}$ with $\bar{y}_{\cdot\cdot}$ are all zero.

E 3.7 Derive (3.99) and (3.100).

E 3.8 From (3.102) derive $\hat{\mu}$ of (3.106). *Note*: This is quite lengthy.

E 3.9 Show that $\hat{\beta}$ and $\hat{\mu}$ of (3.104) and (3.106) are unbiased.

E 3.10 Simplify $\hat{\beta}$ and $\hat{\mu}$ of (3.104) and (3.106) for $n_i = n \; \forall \, i$.

E 3.11 Derive (3.107) and (3.108).

E 3.12 For x_i taking on only the values 0 and 1, find the MLEs of α and β from (3.116).

E 3.13 Suppose there is a value d such that $y_i = 1$ for all observations with $x_i > d$ and $y_i = 0$ for all observations with $x_i < d$. Show that the log likelihood (3.113) is an increasing function of β for an appropriately chosen value of α and hence that a finite MLE does not exist.

E 3.14 Under $\beta = 0$ for model (3.109), show that $\hat{\alpha}_0$, the estimate of α, is $\log[\bar{y}/(1-\bar{y})]$.

E 3.15 Using the fact that $\dfrac{1}{1+e^{-t}} \approx \Phi\left(t\dfrac{16\sqrt{3}}{15\pi}\right)$ (Johnson and Kotz, 1970, p. 6) and the results of E 3.16, derive α^* and β^* of (3.128).

Chapter 4

LINEAR MODELS (LMs)

This chapter provides a thumbnail discussion of linear models (LMs), one of the most widely treated branches of statistics, both in theory and in practice, embracing, as it does, regression, analysis of variance, and analysis of covariance. These topics are dealt with in varying degrees of detail in a myriad of books and papers, so it is not our intention to have this book or this chapter replicate them to any great extent. We simply assume at this point that the reader is familiar with the basic ideas dealt with in Chapters 1 through 3.

The prime object of the chapter is to describe the general ideas of LMs and the analysis of data based thereon. In doing this we establish notation and concepts for use in the succeeding chapters on linear mixed models (LMMs), generalized linear models (GLMs) and some nonlinear models. We begin with an introductory example.

Consider a portion of the experiment on the growth of potato lesions described in Chapter 3. That experiment extended over several weeks; we consider just one week, the first, say. In that week there are 16 observations, consisting of the extent of lesions on each of four leaves, at four different temperatures. Let y_{ij} be the average log diameter of the lesions on leaf j at temperature i, for $j = 1, 2, \ldots, 4$ and $i = 1, 2, \ldots, 4$. Then for μ being an overall mean, and τ_i the effect of temperature on lesion number we could take, for each i

$$\mathrm{E}[y_{ij}] = \mu + \tau_i, \tag{4.1}$$

for $j = 1, 2, \ldots, 4$. For

$$\mathrm{E}[\mathbf{y}_i] = \mathrm{E}\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \begin{bmatrix} \mu + \tau_i \\ \mu + \tau_i \\ \mu + \tau_i \\ \mu + \tau_i \end{bmatrix} = (\mu + \tau_i)\mathbf{1}_4$$

we can define

$$\mathbf{y} = \left\{ {}_c\, \mathbf{y}_i \right\}_{i=1}^{4} \tag{4.2}$$

and then write

$$\mathrm{E}[\mathbf{y}] = \begin{bmatrix} \mathbf{1}_4 & \mathbf{1}_4 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_4 & \mathbf{0} & \mathbf{1}_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_4 & \mathbf{0} & \mathbf{0} & \mathbf{1}_4 & \mathbf{0} \\ \mathbf{1}_4 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_4 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} \tag{4.3}$$

$$= \left[\mathbf{1}_{16} \quad \left\{ {}_d\, \mathbf{1}_4 \right\}_{i=1}^{4} \right] \boldsymbol{\beta} \qquad \text{for} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} \tag{4.4}$$

$$= \mathbf{X}\boldsymbol{\beta} \qquad \text{for} \qquad \mathbf{X} = \left[\mathbf{1}_{16} \quad \left\{ {}_d\, \mathbf{1}_4 \right\}_{i=1}^{4} \right]. \tag{4.5}$$

Throughout all this the μ and τs are taken as fixed effects which we wish to estimate. That is what is now considered. But we go no further with this example, using (4.5) simply as a base from which to describe a general model.

4.1 A GENERAL MODEL

We think of dealing with N items of data, arrayed as a vector $\mathbf{y}$ of order $N \times 1$, and we take the basic (vector) equation of a model to be

$$\mathrm{E}[\mathbf{y}] = \boldsymbol{\mu} \tag{4.6}$$

where, for example, $\boldsymbol{\mu}$ may have the form $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ as in (4.5). We turn to this form in Section 4.2.

With $\mathbf{y}$ being data, we think of its elements as being realized values of some random variable which would traditionally be denoted as $\mathbf{Y}$.

For simplicity we abandon this notational distinction and use $\mathbf{y}$ both as a realization of a random vector and as the random vector itself. In this latter sense we attribute to $\mathbf{y}$ a variance-covariance matrix $\mathbf{V}$ and write

$$\text{var}(\mathbf{y}) = \mathbf{V}. \tag{4.7}$$

Then, to combine (4.6) and (4.7) we write

$$\mathbf{y} \sim (\boldsymbol{\mu}, \mathbf{V}), \tag{4.8}$$

meaning that $\mathbf{y}$ has mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{V}$.

Statement (4.8) is very general. It does no more than assign a symbol $\boldsymbol{\mu}$ to the mean $\mathbf{y}$ and another symbol $\mathbf{V}$ to the matrix of var($\mathbf{y}$). As they stand, $\boldsymbol{\mu}$ and $\mathbf{V}$ are nothing more than symbols. $\boldsymbol{\mu}$ has N elements and $\mathbf{V} = \mathbf{V}'$ has $N(N+1)/2$ unique elements. But there are only N data values; so without describing (modeling) $\boldsymbol{\mu}$ and $\mathbf{V}$ in terms of less than N parameters, $\boldsymbol{\mu}$ and $\mathbf{V}$ cannot be estimated. Thus we have to specify $\boldsymbol{\mu}$ and $\mathbf{V}$ in terms of underlying parameters appropriate to the nature of the data being studied. And this is just what we do for the different forms of models, LMs, LMMs, GLMs, GLMMs and some nonlinear models. We start with LMs.

4.2 A LINEAR MODEL FOR FIXED EFFECTS

Special forms of $\boldsymbol{\mu}$ and $\mathbf{V}$ for the traditional linear model for fixed effects are

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{4.9}$$

and

$$\mathbf{V} = \sigma^2\mathbf{I}_N. \tag{4.10}$$

$\boldsymbol{\beta}$ in (4.9) is a $p \times 1$ vector of unknown fixed effects and $\mathbf{X}$ is a known matrix, of order $N \times p$. Thus (4.8) and (4.9) give

$$\text{E}[\mathbf{y}] = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{4.11}$$

a vector of linear combinations of fixed effects. And from (4.7) and (4.10)

$$\text{var}(\mathbf{y}) = \mathbf{V} = \sigma^2\mathbf{I}, \tag{4.12}$$

which means that the variance of every element of $\mathbf{y}$ is taken as being the same, namely σ^2, and the covariance between every pair of elements is taken as being zero. In summary, we therefore have for LMs

$$\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}). \tag{4.13}$$

The equation $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ of (4.11) is called the *model equation* and $\mathbf{X}$ is the *model matrix*. In many situations $\mathbf{X}$ will have elements which are all 0 or 1, in which case $\mathbf{X}$ is known as an *incidence matrix*. But it is perfectly permissible for $\mathbf{X}$ to also have columns of observed or measured variables, such as predictor variables in regression (e.g., Chapter 3) or concomitant variables in analysis of covariance.

Notice that although (4.11) and (4.12) specify the mean and variance-covariance structure for $\mathbf{y}$ of (4.13), the actual form of the distribution is not specified in (4.13).

4.3 MAXIMUM LIKELIHOOD UNDER NORMALITY

Although estimating the $\boldsymbol{\beta}$ of (4.13) is often done by ordinary least squares (OLSE) or generalized least squares (GLSE), neither of which demand having an underlying distribution for $\mathbf{y}$, we follow the general approach of this book and use maximum likelihood (ML). This has the particular merit of simultaneously providing an estimator not only for $\boldsymbol{\beta}$ but also one for σ^2 of (4.13).

Starting with the assumption that $\mathbf{y}$ follows a multivariate normal distribution,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

the likelihood function is

$$L = L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = \frac{\exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I}/\sigma^2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]}{(2\pi\sigma^2)^{\frac{1}{2}N}} \tag{4.14}$$

and so the log likelihood is

$$l = \log L = -\tfrac{1}{2}N\log(2\pi) - \tfrac{1}{2}N\log\sigma^2 - \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma^2. \tag{4.15}$$

Denoting $\partial l/\partial\boldsymbol{\beta}$ by $l_{\boldsymbol{\beta}}$ and $\partial l/\partial\sigma^2$ by l_{σ^2}, it is easily found that

$$l_{\boldsymbol{\beta}} = \frac{\mathbf{X}'(\mathbf{y} - \mathrm{E}[\mathbf{y}])}{\sigma^2} = \frac{\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2} \tag{4.16}$$

and

$$l_{\sigma^2} = \frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\sigma^2} - \frac{N}{2\sigma^2}. \tag{4.17}$$

Equating l_β to zero, and in doing so denoting β by $\hat{\beta}$ gives $\mathbf{X}'\hat{\mathrm{E}}[\mathbf{y}] = \mathbf{X}'\mathbf{y}$ or

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}. \tag{4.18}$$

Equations (4.18) are known as *normal equations*, from which

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \text{if} \quad (\mathbf{X}'\mathbf{X})^{-1} \text{ exists.} \tag{4.19}$$

And from equating l_{σ^2} to zero, with $\hat{\sigma}^2$ in place of σ^2 we get

$$\hat{\sigma}^2 = (\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta})/N. \tag{4.20}$$

These are the ML estimators; they are not just solutions, because $\hat{\beta}$ lies in the same range as β, namely $-\infty < \beta < \infty$, and $\hat{\sigma}^2$ is non-negative, as is σ^2.

The first thing to notice about $\hat{\beta}$ of (4.19) is that it exists only if $(\mathbf{X}'\mathbf{X})^{-1}$ exists: and this requires $\mathbf{X}_{N\times p}$ to have full column rank p. This is very restrictive, because in many situations $\mathbf{X}$ does not have rank p.

4.4 SUFFICIENT STATISTICS

Working from (4.14) we can rewrite the density of $\mathbf{y}$ as

$$f_{\mathbf{y}}(\mathbf{y}) = (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\right\} \tag{4.21}$$

$$= (2\pi\sigma^2)^{-N/2}\exp\left\{\frac{1}{\sigma^2}\beta'\mathbf{X}'\mathbf{y} - \frac{1}{2\sigma^2}\mathbf{y}'\mathbf{y} - \frac{1}{2\sigma^2}\beta'\mathbf{X}'\mathbf{X}\beta\right\}.$$

To identify the sufficient statistics we define $\theta' = (\beta', \sigma^2)'$ as the parameter vector and define the following functions to match with the definition in Section S.3 of Appendix S:

$$d(\theta) = (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\beta'\mathbf{X}'\mathbf{X}\beta\right),$$

$$h(\mathbf{y}) = 1,$$

$$
\begin{aligned}
(T_1(\mathbf{y}), T_2(\mathbf{y}), \ldots, T_p(\mathbf{y}))' &= \mathbf{X}'\mathbf{y}, \\
T_{p+1}(\mathbf{y}) &= \mathbf{y}'\mathbf{y}, \\
(\nu_1(\boldsymbol{\theta}), \ldots, \nu_p(\boldsymbol{\theta}))' &= \boldsymbol{\beta}'/\sigma^2, \text{ and} \\
\nu_{p+1}(\boldsymbol{\theta}) &= -\frac{1}{2\sigma^2}.
\end{aligned}
$$

This results in the distribution of $\mathbf{y}$ being in the exponential family and hence the sufficient statistic is $(\mathbf{y}'\mathbf{X} \ \mathbf{y}'\mathbf{y})'$. As expected, the maximum likelihood estimators are functions of the sufficient statistics.

4.5 MANY APPARENT ESTIMATORS

a. General result

When $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist there is no longer just one solution $\hat{\boldsymbol{\beta}}$ given by (4.19). Instead there is an infinite number of solutions to (4.18) of the form

$$\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} \tag{4.22}$$

for $(\mathbf{X}'\mathbf{X})^-$ being any matrix satisfying

$$\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}. \tag{4.23}$$

For notational convenience we use $\mathbf{G}$ for $(\mathbf{X}'\mathbf{X})^-$:

$$\mathbf{G} \equiv (\mathbf{X}'\mathbf{X})^-. \tag{4.24}$$

By virtue of the nature of (4.23), we call $(\mathbf{X}'\mathbf{X})^-$ a *generalized inverse* of $\mathbf{X}'\mathbf{X}$. For $\mathbf{X}$ of full column rank, $(\mathbf{X}'\mathbf{X})^-$ of (4.24) is $(\mathbf{X}'\mathbf{X})^{-1}$, which exists, and we use $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ as the estimator of $\boldsymbol{\beta}$. But when $\mathbf{X}$ has less than full column rank, $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist and there are many matrices $(\mathbf{X}'\mathbf{X})^-$ satisfying (4.24). Thus there are many solutions $\boldsymbol{\beta}^0$ available from (4.22). That is why the symbol $\boldsymbol{\beta}^0$ is used in (4.22), as emphasis for distinguishing the many solutions $\boldsymbol{\beta}^0$ when $\mathbf{X}$ is less than full column rank from the solitary $\hat{\boldsymbol{\beta}}$ when $\mathbf{X}$ is of full column rank.

Note: From this point onward we assume $\mathbf{X}$ has less than full column rank, unless otherwise stated.

b. Mean and variance

Not only are there numerous values of $\boldsymbol{\beta}^0$ for any given $\mathbf{X}$, but none of them is unbiased for $\boldsymbol{\beta}$. With $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ of (4.11) it is easily seen that the expected value of $\boldsymbol{\beta}^0$ is $\mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$:

$$\mathrm{E}[\boldsymbol{\beta}^0] = \mathrm{E}[\mathbf{G}\mathbf{X}'\mathbf{y}] = \mathbf{G}\mathbf{X}'\mathrm{E}[\mathbf{y}] = \mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \tag{4.25}$$

In general this is not $\boldsymbol{\beta}$; it is $\boldsymbol{\beta}$ if $\mathbf{G}\mathbf{X}'\mathbf{X}$ equals $\mathbf{I}$, but this occurs only when $\mathbf{X}'\mathbf{X}$ is non-singular, in which case $\mathbf{G}$ is $(\mathbf{X}'\mathbf{X})^{-1}$.

The variance of $\boldsymbol{\beta}^0$ is

$$\mathrm{var}(\boldsymbol{\beta}^0) = \mathrm{var}(\mathbf{G}\mathbf{X}'\mathbf{y}) = \mathbf{G}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\mathbf{G}' = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\sigma^2. \tag{4.26}$$

This does not simplify to $\mathbf{G}\sigma^2$, which one might expect on the basis of it being $(\mathbf{X}'\mathbf{X})^{-1}$ when that exists.

In view of there being numerous solutions $\boldsymbol{\beta}^0$, with none of them unbiased for $\boldsymbol{\beta}$, one well might wonder what use there is for any $\boldsymbol{\beta}^0$. Fortunately, there is an important invariance property pertaining to $\mathbf{X}\boldsymbol{\beta}^0$ which provides widespread applicability and utility.

c. Invariance properties

The infinity of values $\boldsymbol{\beta}^0$, together with the dependence on $\mathbf{G}$ of $\mathrm{E}[\boldsymbol{\beta}^0]$ and $\mathrm{var}(\boldsymbol{\beta}^0)$ in (4.25) and (4.26), clearly negates using any $\boldsymbol{\beta}^0$ as an estimator of $\boldsymbol{\beta}$. But three standard properties of $\mathbf{G}$ (Section M.4 of Appendix M) do provide useful results that are invariant to $\mathbf{G}$ (for given $\mathbf{X}$, of course). These results are

$$\mathbf{G}' \text{ is a generalized inverse of } \mathbf{X}'\mathbf{X}; \tag{4.27}$$

$$\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}\mathbf{G}'\mathbf{X}' \text{ is invariant to } \mathbf{G}; \tag{4.28}$$

and

$$\mathbf{X} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}. \tag{4.29}$$

These results lead to the following useful properties involving $\mathbf{X}\boldsymbol{\beta}^0$. First, the predicted mean of $\mathbf{y}$ is

$$\widehat{\mathrm{E}[\mathbf{y}]} = \hat{\boldsymbol{\mu}} = \mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} \tag{4.30}$$

and is invariant to $\mathbf{G}$. Thus for every $\boldsymbol{\beta}^0 = \mathbf{G}\mathbf{X}'\mathbf{y}$, no matter what $\mathbf{G}$ is used, the value of $\mathbf{X}\boldsymbol{\beta}^0$ does not depend on $\mathbf{G}$. Second, in place of (4.25),

$$\mathrm{E}[\hat{\boldsymbol{\mu}}] = \mathrm{E}[\mathbf{X}\boldsymbol{\beta}^0] = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} \tag{4.31}$$

from (4.29); and third,

$$\mathrm{var}(\hat{\boldsymbol{\mu}}) = \mathrm{var}(\mathbf{X}\boldsymbol{\beta}^0) = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{I}\sigma^2\mathbf{X}\mathbf{G}'\mathbf{X}' = \mathbf{X}\mathbf{G}\mathbf{X}'\sigma^2 \tag{4.32}$$

is also invariant to $\mathbf{G}$.

d. Distributions

For

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\ \sigma^2\mathbf{I})$$

(4.25) and (4.26) give

$$\boldsymbol{\beta}^0 \sim \mathcal{N}(\mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta},\ \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\sigma^2), \tag{4.33}$$

whilst from (4.31) and (4.32)

$$\mathbf{X}\boldsymbol{\beta}^0 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\ \mathbf{X}\mathbf{G}\mathbf{X}'\sigma^2). \tag{4.34}$$

4.6 ESTIMABLE FUNCTIONS

a. Introduction

When we are interested in estimating functions of $\boldsymbol{\beta}$ we must distinguish between those which are functions just of $\mathbf{X}\boldsymbol{\beta}$ and those which are not. Because the parameter $\boldsymbol{\beta}$ affects the distribution of $\mathbf{y}$ only through $\mathbf{X}\boldsymbol{\beta}$, it is only functions of $\mathbf{X}\boldsymbol{\beta}$ which can be estimated satisfactorily. In that context consider a linear combination of elements of $\mathbf{X}\boldsymbol{\beta}$, say $\mathbf{t}'\mathbf{X}\boldsymbol{\beta}$. An unbiased, though perhaps not efficient, estimator of $\mathbf{t}'\mathbf{X}\boldsymbol{\beta}$ is $\mathbf{t}'\mathbf{y}$. Therefore we can estimate $\mathbf{q}'\boldsymbol{\beta}$ whenever $\mathbf{q}'$ is of the form $\mathbf{t}'\mathbf{X}$. But when $\mathbf{q}'$ cannot be written in the form $\mathbf{t}'\mathbf{X}$ it is not possible to estimate $\mathbf{q}'\boldsymbol{\beta}$ unbiasedly. Thus for estimating linear functions of $\boldsymbol{\beta}$ the only ones we can consider are those of the form $\mathbf{q}'\boldsymbol{\beta} = \mathbf{t}'\mathbf{X}\boldsymbol{\beta}$. Such functions are said to be *estimable functions*—functions of $\boldsymbol{\beta}$. Their characteristics are now described.

b. Definition

A linear combination of elements of β is $\mathbf{q}'\beta$ for some row vector $\mathbf{q}'$. It is called an *estimable function*, and is said to be *estimable*, under the following circumstances:

$$\mathbf{q}'\beta \text{ is estimable} \quad \text{iff} \quad \mathbf{q}'\beta = \mathbf{t}'\mathbf{X}\beta \quad \forall \quad \beta; \tag{4.35}$$

i.e.,

$$\mathbf{q}'\beta \text{ is estimable} \quad \text{iff} \quad \mathbf{q}' = \mathbf{t}'\mathbf{X} \text{ for some } \mathbf{t}'. \tag{4.36}$$

c. Properties

Three important properties of estimable functions are as follows:

(1) $\mathrm{E}[y_k]$ is estimable for any element y_k of $\mathbf{y}$. This is so because for

$$\mathbf{t}' = (\text{row of zeros except } k\text{th element being 1})$$

we have

$$\mathrm{E}[y_k] = \mu_k = (k\text{th row of } \mathbf{X})\beta = (\mathbf{t}'\mathbf{X})\beta,$$

which satisfies (4.35). Thus the expected value of each observation is estimable.

(2) Linear combinations of estimable functions are estimable. Suppose $\mathbf{q}_1'\beta = \mathbf{t}_1'\mathbf{X}\beta$ and $\mathbf{q}_2'\beta = \mathbf{t}_2'\mathbf{X}\beta$ are estimable functions. Combining them using scalars c_1 and c_2 gives

$$c_1\mathbf{q}_1\beta + c_2\mathbf{q}_2\beta = c_1\mathbf{t}_1'\mathbf{X}\beta + c_2\mathbf{t}_2'\mathbf{X}\beta = (c_1\mathbf{t}_1' + c_2\mathbf{t}_2')\mathbf{X}\beta = \mathbf{t}_*'\mathbf{X}\beta$$

for $\mathbf{t}_*' = c_1\mathbf{t}_1' + c_2\mathbf{t}_2'$, thus demonstrating estimability.

(3) Putting properties (1) and (2) together enables the establishment of functions (linear combinations of elements of β) that are estimable without having to ascertain the corresponding vectors $\mathbf{t}'$. This is illustrated in the example of Section 4.7.

d. Estimation

With $\mathbf{q}' = \mathbf{t}'\mathbf{X}$ of (4.36) and $\mathbf{X}\beta^0 = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}$ of (4.30) being invariant to $\mathbf{G}$, we have the ML estimator of estimable $\mathbf{q}'\beta$ as

$$\widehat{\mathbf{q}'\beta} = \mathbf{t}'\mathbf{X}\beta^0 = \mathbf{q}'\beta^0 = \mathbf{t}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y}, \tag{4.37}$$

invariant to $\mathbf{G}$. A notable feature of this is the second equality, that the estimator of estimable $\mathbf{q}'\beta$ is $\mathbf{q}'\beta^0$, the same linear combination of elements of β^0 as is the estimable function $\mathbf{q}'\beta$ of β. Furthermore, under normality of $\mathbf{y}$, using (4.37), we have

$$\widehat{\mathbf{q}'\beta} = \mathbf{q}'\beta^0 \sim \mathcal{N}(\mathbf{q}'\beta,\ \mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2). \tag{4.38}$$

Since the ML estimator of $\mathbf{q}'\beta$ is unbiased and based on the sufficient statistic (Section 4.4) it is a uniform minimum variance unbiased (UMVU) estimator.

The invariance of $\mathbf{q}'\beta^0$ and of its variance to different $\mathbf{G}$ when $\mathbf{q}'\beta$ is an estimable function are two eminently practical features of an estimable function. They totally avoid the impracticality of β^0 as an estimator of β through it and its variance being functions of $\mathbf{G}$ for which there is an infinite number of values.

4.7 A NUMERICAL EXAMPLE

For the sole purpose of numerically illustrating some of the preceding results, suppose for the 1-way classification of Chapter 2 we have the following data, for three classes with 3, 2 and 1 observations.

$i=1$	$i=2$	$i=3$
72	48	36
36	12	
12		
120	60	36

Then for

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{bmatrix} = \begin{bmatrix} 72 \\ 36 \\ 12 \\ 48 \\ 12 \\ 36 \end{bmatrix},$$

$$\mathbf{X}\beta = \begin{bmatrix} \mu+\alpha_1 & & \\ \mu+\alpha_1 & & \\ \mu+\alpha_1 & & \\ \mu & +\alpha_2 & \\ \mu & +\alpha_2 & \\ \mu & & +\alpha_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot \\ 1 & \cdot & 1 & \cdot \\ 1 & \cdot & 1 & \cdot \\ 1 & \cdot & \cdot & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}. \quad (4.39)$$

The resulting normal equations, $\mathbf{X}'\mathbf{X}\beta^0 = \mathbf{X}'\mathbf{y}$ of (4.18), are therefore

$$\begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & \cdot & \cdot \\ 2 & \cdot & 2 & \cdot \\ 1 & \cdot & \cdot & 1 \end{bmatrix} \begin{bmatrix} \mu^0 \\ \alpha_1^0 \\ \alpha_2^0 \\ \alpha_3^0 \end{bmatrix} = \begin{bmatrix} 216 \\ 120 \\ 60 \\ 36 \end{bmatrix}. \quad (4.40)$$

Four different matrices $\mathbf{G} = (\mathbf{X}'\mathbf{X})^-$ are

$$\mathbf{G}_1 = \begin{bmatrix} 0 & \cdot & \cdot & \cdot \\ \cdot & \frac{1}{3} & \cdot & \cdot \\ \cdot & \cdot & \frac{1}{2} & \cdot \\ \cdot & \cdot & \cdot & \frac{1}{1} \end{bmatrix}, \quad \mathbf{G}_2 = \begin{bmatrix} 1 & -1 & -1 & \cdot \\ -1 & 1\frac{1}{3} & 1 & \cdot \\ -1 & 1 & 1\frac{1}{2} & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix},$$

$$\mathbf{G}_3 = \frac{1}{18}\begin{bmatrix} \cdot & 2 & 3 & 6 \\ \cdot & 4 & -3 & -6 \\ \cdot & -2 & 6 & -6 \\ \cdot & -2 & -3 & 12 \end{bmatrix} \text{ and } \mathbf{G}_4 = \frac{1}{54}\begin{bmatrix} 17 & -11 & -8 & 1 \\ -11 & 23 & 2 & -7 \\ -8 & 2 & 26 & -10 \\ 1 & -7 & -10 & 35 \end{bmatrix}.$$

Post-multiplying these by $\mathbf{X}'\mathbf{y} = [216 \quad 120 \quad 60 \quad 36]'$ gives solutions $\beta^0 = [\mu^0 \quad \alpha_1^0 \quad \alpha_2^0 \quad \alpha_3^0]'$ as

$$\beta_1^0 = \begin{bmatrix} 0 \\ 40 \\ 30 \\ 36 \end{bmatrix}, \quad \beta_2^0 = \begin{bmatrix} 36 \\ 4 \\ -6 \\ 0 \end{bmatrix} \quad \text{and} \quad \beta_3^0 = \beta_4^0 = \begin{bmatrix} 35\frac{1}{3} \\ 4\frac{2}{3} \\ -5\frac{1}{3} \\ \frac{2}{3} \end{bmatrix}. \quad (4.41)$$

Note that β_2^0 has $\alpha_3^0 = 0$ (i.e., the last effect has solution zero), a characteristic seen in SAS GLM and Proc MIXED outputs. And $\beta_3^0 = \beta_4^0$ has $\alpha_1^0 + \alpha_2^0 + \alpha_3^0 = 0$, a feature of some other computing software.

It is also interesting to note that two different G-matrices can give the same $\boldsymbol{\beta}^0 = \mathbf{G}\mathbf{X}'\mathbf{y}$.

Demonstrating (4.28) we find that $\mathbf{X}\mathbf{G}\mathbf{X}'$ for each of the four $\mathbf{G}$s is

$$\mathbf{X}\mathbf{G}\mathbf{X}' = \begin{bmatrix} \bar{\mathbf{J}}_3 & \cdot & \cdot \\ \cdot & \bar{\mathbf{J}}_2 & \cdot \\ \cdot & \cdot & \bar{\mathbf{J}}_1 \end{bmatrix}$$

where $\bar{\mathbf{J}}_n$ is $n \times n$ with every element $1/n$. It is then easily verified that $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}$ of (4.29). Next, from (4.30) it is easily seen that $(\mathbf{X}\boldsymbol{\beta}^0)' = [40 \quad 40 \quad 40 \quad 30 \quad 30 \quad 36]$ for each $\boldsymbol{\beta}^0$.

By property (1) of Section 4.6c, having $\mathrm{E}[y_{ij}] = \mu + \alpha_i$ means that $\mu + \alpha_i$ is estimable; and from each $\boldsymbol{\beta}^0$ we find that the MLE of $\mu + \alpha_1$ is 40. For example, $\mu^0 + \alpha_1^0$ from $\boldsymbol{\beta}_1^0$ is $0 + 40 = 40$, from $\boldsymbol{\beta}_2^0$ it is $36 + 4 = 40$ and from $\boldsymbol{\beta}_3^0$ it is $35\frac{1}{3} + 4\frac{2}{3} = 40$. Also by property (2) $\alpha_1 - \alpha_2$ is estimable because $\alpha_1 - \alpha_2 = \mu + \alpha_1 - (\mu + \alpha_2)$; and each $\boldsymbol{\beta}^0$ gives $\alpha_1^0 - \alpha_2^0 = 10$. These calculations demonstrate for estimable $\mathbf{q}'\boldsymbol{\beta}^0$ the invariance of $\mathbf{q}'\boldsymbol{\beta}^0$ to $\boldsymbol{\beta}^0$.

Writing estimable $\mu + \alpha_1$ as $\mathbf{q}'\boldsymbol{\beta}$ for $\mathbf{q}' = [1 \quad 1 \quad 0 \quad 0]$ it will then be found, using (4.38), that

$$\mathrm{var}(\widehat{\mathbf{q}'\boldsymbol{\beta}}) = \mathrm{var}(\mathbf{q}'\boldsymbol{\beta}^0) = \mathbf{q}'\mathbf{G}\mathbf{q}\sigma^2 = \tfrac{1}{3}\sigma^2$$

no matter which $\mathbf{G}$ is used.

Remark: Derivation of the four G-matrices is as follows: $\mathbf{G}_1$ and $\mathbf{G}_2$ are based on the regular inverse of the lower right and upper left (respectively) 3×3 submatrices. $\mathbf{G}_3$ uses the formula for $\mathbf{G}_r$ in Searle (1987, p. 307) and $\mathbf{G}_4$ uses $(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}$ discussed in Searle (1971, p. 23) and more thoroughly in Searle (1999).

4.8 ESTIMATING RESIDUAL VARIANCE

a. Estimation

Equation (4.20) shows the ML estimator for σ^2 of $\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ when $\mathbf{X}'\mathbf{X}$ is nonsingular. In that equation replacing $\hat{\boldsymbol{\beta}}$ by $\boldsymbol{\beta}^0$ gives the ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$ when $\mathbf{X}'\mathbf{X}$ is singular as follows,

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)}{N} = \frac{\mathrm{SSE}}{N}. \tag{4.42}$$

The numerator of $\hat{\sigma}^2_{\mathrm{ML}}$ is denoted by SSE, the residual (or error) sum of squares in the context of analysis of variance. For convenience we retain that label but without reference to analysis of variance.

It is of interest to see if $\hat{\sigma}^2_{\mathrm{ML}}$ is unbiased for σ^2. To do this we use the result in Section S.1b of Appendix S for the expected value of a quadratic form. Then after simplifying SSE to be

$$\mathrm{SSE} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0) = \mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y} \tag{4.43}$$

we find, for

$$r_{\mathbf{X}} = \text{rank of } \mathbf{X} \tag{4.44}$$

that

$$\mathrm{E}[\mathrm{SSE}] = \mathrm{E}[\mathbf{y}'(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y}] = (N - r_{\mathbf{X}})\sigma^2. \tag{4.45}$$

Note that this result does not rely on any distributional form for $\mathbf{y}$, only on $\mathbf{y}$ having mean $\mathbf{X}\boldsymbol{\beta}$ and variance $\sigma^2\mathbf{I}$. Then (4.45) gives

$$\mathrm{E}[\hat{\sigma}^2_{\mathrm{ML}}] = \frac{\mathrm{E}[\mathrm{SSE}]}{N} = \left(1 - \frac{r_{\mathbf{X}}}{N}\right)\sigma^2.$$

Thus the ML estimator of σ^2 is biased downward.

On the other hand, dividing SSE by $(N - r_{\mathbf{X}})$ gives an unbiased estimator

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{N - r_{\mathbf{X}}} \qquad \text{with} \qquad \mathrm{E}[\hat{\sigma}^2] = \sigma^2. \tag{4.46}$$

This is, of course, the usual estimator used in analysis of variance, and being based on the sufficient statistic is a UMVU estimator.

b. Distribution of estimators

SSE is a quadratic form; and in Section S.2c of Appendix S is the following important theorem concerning the distribution of quadratic forms.

> **Theorem.** When $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ with $\mathbf{V}$ nonsingular then $\mathbf{y}'\mathbf{A}\mathbf{y}$ is distributed as a non-central χ^2 with degrees of freedom $\nu = \mathrm{rank}(\mathbf{A}\mathbf{V})$ and non-centrality parameter $\frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$, if and only if $\mathbf{A}\mathbf{V}$ is idempotent.

In applying this theorem there is also the extension that whenever $\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = 0$, the distribution becomes a usual (central) χ^2 distribution on ν degrees of freedom, which is denoted by χ^2_ν.

In applying the preceding theorem to SSE of (4.43) it will be found (E 4.12) that

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{N-r_{\mathbf{X}}}. \tag{4.47}$$

Therefore

$$\hat{\sigma}^2 = \frac{\text{SSE}}{N - r_{\mathbf{X}}} \sim \left(\frac{\sigma^2}{N - r_{\mathbf{X}}}\right) \chi^2_{N-r_{\mathbf{X}}}, \tag{4.48}$$

meaning by this that the distribution of $\hat{\sigma}^2$ is a scalar multiple of $\chi^2_{N-r_{\mathbf{X}}}$, that scalar being $\sigma^2/(N - r_{\mathbf{X}})$. And from (4.48)

$$\text{var}(\hat{\sigma}^2) = \frac{2\sigma^4}{N - r_{\mathbf{X}}}, \tag{4.49}$$

and from (4.42)

$$\hat{\sigma}^2_{\text{ML}} \sim \left(\frac{\sigma^2}{N}\right) \chi^2_{N-r_{\mathbf{X}}} \tag{4.50}$$

and so

$$\text{var}\left(\hat{\sigma}^2_{\text{ML}}\right) = \frac{2\sigma^4(N - r_{\mathbf{X}})}{N^2} = \frac{2\sigma^4}{N}\left(1 - \frac{r_{\mathbf{X}}}{N}\right). \tag{4.51}$$

These results, (4.48) and (4.50), rely on the normality of $\mathbf{y}$. An alternative expression comes from the general result (applicable to all ML estimators, regardless of the assumed distribution) that ML estimators have asymptotic normal distributions with variance structure given by the inverse of the information matrix. In the case of $\hat{\sigma}^2_{\text{ML}}$ this yields an asymptotic variance of $\hat{\sigma}^2_{\text{ML}}$ of $2\sigma^4/N$. And this is, of course, close in value to (4.51) when $r_{\mathbf{X}}/N$ is small.

4.9 COMMENTS ON 1- AND 2-WAY CLASSIFICATIONS

Section 4.7 numerically illustrates some of the basic properties surrounding the estimation of β from $\text{E}[\mathbf{y}] = \mathbf{X}\beta$. Here, for the 1-way classification, we describe some of its general results. For the 2-way classification we merely give a hint as to possible complications. Both of the 1-way and 2-way classifications are dealt with in great detail in a variety of books (e.g., Searle, 1987, 1997).

a. The 1-way classification

The example of Section 4.7 is that of a 1-way classification with unbalanced data, i.e., not all the same number of observations in the

subclasses. Equations (4.40) are an example of the normal equations which for the model equation

$$\mathrm{E}[y_{ij}] = \mu + \alpha_i \qquad \text{for} \qquad i = 1, 2, \ldots, a$$

always take the form

$$\begin{bmatrix} N & \{_r\, n_i\} \\ \{_c\, n_i\} & \{_d\, n_i\} \end{bmatrix} \begin{bmatrix} \mu \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} y_{\cdot\cdot} \\ \{_c\, y_{i\cdot}\} \end{bmatrix} \quad \text{for} \quad i = 1, 2, \ldots, a.$$

And the easiest solution, exemplified by $\boldsymbol{\beta}_1^0$ of (4.41), using

$$\mathbf{G} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \{_d\, 1/n_i\} \end{bmatrix}$$

gives

$$\mu^0 = 0 \qquad \text{and} \qquad \alpha_i^0 = \bar{y}_{i\cdot}\,.$$

Thus with $\mathrm{E}[y_{ij}] = \mu + \alpha_i$ being estimable its ML estimator is

$$\widehat{\mu + \alpha}_i = \mu^0 + \alpha_i^0 = 0 + \bar{y}_{i\cdot} = \bar{y}_{i\cdot}\,. \tag{4.52}$$

Also, with $\alpha_i - \alpha_k = \mu + \alpha_i - (\mu + \alpha_k)$ being estimable, its estimator is

$$\widehat{\alpha_i - \alpha_k} = \alpha_1^0 - \alpha_k^0 = \bar{y}_{i\cdot} - \bar{y}_{k\cdot};$$

and the sampling variance of $\alpha_i^0 - \alpha_k^0$ from (4.38) is $(1/n_i + 1/n_k)\sigma^2$.

An alternative model, simpler than $\mathrm{E}[y_{ij}] = \mu + \alpha_i$, is $\mathrm{E}[y_{ij}] = \mu_i$, often called the *cell means* model. In using it, $\mathbf{X}$ of (4.39) would be changed to exclude its first column, and $\mathbf{X}'\mathbf{X}$ and $\mathbf{G}$ would have their first row and column excluded. This change in the model is effectively equating $\mu + \alpha_i$ and μ_i. Hence $\hat{\mu}_i = \widehat{\mu + \alpha}_i = \bar{y}_{i\cdot}$ from (4.52).

b. The 2-way classification

The example in Section 1.3b, where y_{ijk} represented the rating of the ith cartoon type by the kth person in the jth group, suggests using

$$\mathrm{E}[y_{ijk}] = \mu + \alpha_i + \beta_j. \tag{4.53}$$

Estimation details for that model are available in many places (e.g., Searle 1971, 1987): Particularly for unbalanced data, those details are

extensive and need not occupy us here. We confine attention to estimability. Suppose we are interested in estimating the mean rating for the ith cartoon type. A reasonable estimator for this might be $\bar{y}_{i\cdot\cdot}$, similar to (4.52). If we define $\mu_i = \mu + \alpha_i$ so that

$$\mathrm{E}[y_{ijk}] = \mu_i + \beta_j, \tag{4.54}$$

then

$$\mathrm{E}[\bar{y}_{i\cdot\cdot}] = \mathrm{E}[\sum_j \sum_k (\mu_i + \beta_j)/n_{i\cdot}] = \mu_i + \sum_j n_{ij}\beta_j/n_{i\cdot},$$

where n_{ij} is the number of observations at the intersection of row i and row j. Thus $\bar{y}_{i\cdot\cdot}$ is not an unbiased estimator of μ_i as might have been expected. This illustrates how careful one must be in drawing what seems like an "obvious" conclusion about what it is that some estimators are estimating. In many cases they are not estimating what one might think is "obvious". However, if one follows the $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}'\mathbf{y}$ estimation procedure one finds that $\alpha_i - \alpha_k$ and $\beta_j - \beta_l$ are estimable. On the other hand, when an interaction effect is added to (4.53) so that

$$\mathrm{E}[y_{ijk}] = \mu_{ij} \equiv \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{4.55}$$

then $\hat{\mu}_{ij} = \bar{y}_{ij\cdot}$ is an estimator of $\mu + \alpha_i + \beta_j + \gamma_{ij}$ (providing $n_{ij} \neq 0$). But then it is impossible to estimate $\alpha_i - \alpha_k$ because the interaction terms can never be gotten rid of.

4.10 TESTING LINEAR HYPOTHESES

The general formulation of a linear hypothesis concerning $\boldsymbol{\beta}$ is

$$H: \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}. \tag{4.56}$$

$\mathbf{K}'$ must satisfy three conditions:

1. $\mathbf{K}' = \mathbf{T}'\mathbf{X}$ for some $\mathbf{T}'$, so that $\mathbf{K}'\boldsymbol{\beta}$ is estimable, with unbiased estimator $\mathbf{K}'\boldsymbol{\beta}^0$ invariant to $\boldsymbol{\beta}^0$.
2. $\mathbf{K}'$ must have full row rank—so that $\mathbf{K}'\boldsymbol{\beta}$ contains no redundant elements.
3. $\mathbf{K}'$ must have no more than $r_{\mathbf{X}}$ rows: i.e., the row rank of $\mathbf{K}'$ cannot exceed the rank of $\mathbf{X}$.

We show two ways of deriving a test.

a. Using the likelihood ratio

When $\boldsymbol{\theta}$ represents the vector of parameters in a model, we use $L(\boldsymbol{\theta})$ as the likelihood. Then for $\hat{\boldsymbol{\theta}}$ being the ML estimator of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}_0$ the ML estimator under the hypothesis, the likelihood ratio is $L(\hat{\boldsymbol{\theta}}_0)/L(\hat{\boldsymbol{\theta}})$, as discussed in Section 2.5b–iii. With this notation, and $\boldsymbol{\theta}' = [\boldsymbol{\beta}' \ \sigma^2]$, we have $L(\boldsymbol{\theta})$ given in (4.14). For $\hat{\boldsymbol{\theta}}$, (4.22) and (4.42) give the values of $\boldsymbol{\beta}$ and σ^2 that maximize $L(\boldsymbol{\beta}, \sigma^2)$ as $\boldsymbol{\beta}^0 = \mathbf{GX}'\mathbf{y}$ and $\hat{\sigma}^2_{\mathrm{ML}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)/N$. Using these in (4.14) in place of $\boldsymbol{\beta}$ and σ^2 gives

$$L(\hat{\boldsymbol{\theta}}) = L(\boldsymbol{\beta}^0, \hat{\sigma}^2_{\mathrm{ML}}) = \left[\frac{N/e}{2\pi(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)}\right]^{\frac{1}{2}N}. \tag{4.57}$$

Deriving $L(\hat{\boldsymbol{\theta}}_0)$ requires maximizing $L(\boldsymbol{\beta}, \sigma^2)$ subject to $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$. The results of doing this (E 4.13) are that for $\hat{\boldsymbol{\theta}}_0$

$$\boldsymbol{\beta}^0_0 = \boldsymbol{\beta}^0 - \mathbf{GK}(\mathbf{K}'\mathbf{GK})^{-1}(\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m}) \tag{4.58}$$

and

$$\hat{\sigma}^2_0 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0_0)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0_0)/N. \tag{4.59}$$

Substituting these values in place of $\boldsymbol{\beta}$ and σ^2 in (4.14) gives, after more algebra,

$$L(\hat{\boldsymbol{\theta}}_0) = L(\boldsymbol{\beta}^0_0, \hat{\sigma}^2_0) = \left[\frac{N/e}{2\pi(\mathrm{SSE} + Q)}\right]^{\frac{1}{2}N} \tag{4.60}$$

for

$$Q = (\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m})'(\mathbf{K}'\mathbf{GK})^{-1}(\mathbf{K}'\boldsymbol{\beta}^0 - \mathbf{m}). \tag{4.61}$$

Then the likelihood ratio reduces to

$$\Lambda = \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})} = \left[\frac{1}{1 + Q/\mathrm{SSE}}\right]^{\frac{1}{2}N}. \tag{4.62}$$

Clearly this ratio, (4.62), is a single-valued function of Q/SSE, decreasing monotonically when Q/SSE increases. Therefore Q/SSE can be used as a test statistic in place of (4.62). Moreover, by the same reasoning, instead of Q/SSE one can use (with degrees of freedom being abbreviated *df*)

$$\frac{Q}{df \text{ for } Q} \Big/ \frac{\mathrm{SSE}}{df \text{ for SSE}} \tag{4.63}$$

as the test statistic.

Then, using the theorem of Section 4.8b one can show (E 4.18) that Q has a non-central χ^2 distribution and SSE has a central χ^2 distribution. Furthermore, Q and SSE are independent, as may be established by using the following theorem.

> **Theorem.** For $\mathbf{y} \sim \mathcal{N}(\mu, \mathbf{V})$ with $\mathbf{V}$ nonsingular, the quadratic forms $\mathbf{y'Ay}$ and $\mathbf{y'By}$ are independent if and only if $\mathbf{AVB} = \mathbf{0}$.

These χ^2 and independence properties of Q and SSE result in our being able to use (4.63) as an F-statistic for testing $H: \mathbf{K'}\beta = \mathbf{m}$; its degrees of freedom are $r(\mathbf{K'})$ and $N - r_{\mathbf{X}}$.

4.11 t-TESTS AND CONFIDENCE INTERVALS

When $\mathbf{K'}$ of the hypothesis $H: \mathbf{K'}\beta = \mathbf{m}$ has just a single row $\mathbf{k'}$, then Q of (4.61) becomes

$$\frac{(\mathbf{k'}\beta^0 - \mathbf{m})^2}{\mathbf{k'Gk}[\text{SSE}/(N - r_{\mathbf{X}})]} = \frac{(\mathbf{k'}\beta^0 - \mathbf{m})^2}{\mathbf{k'Gk}\hat{\sigma}^2}$$

to be compared to the $\mathcal{F}$-distribution on 1 and $N-r$ degrees of freedom, where we use the notation $r = r_{\mathbf{X}} = \text{rank}(\mathbf{X})$ for this section. Now recall that when a variable is distributed as the t-distribution on n degrees of freedom its square is distributed as $\mathcal{F}_n^1$. Therefore

$$\frac{\mathbf{k'}\beta^0 - \mathbf{m}}{\hat{\sigma}\sqrt{\mathbf{k'Gk}}} \sim t_{N-r}$$

provides a test of $H: \mathbf{k'}\beta^0 = \mathbf{m}$. This t-test is also useful for one-sided alternatives, which is not so for the F-test.

Suppose $\mathbf{q'}\beta$ is estimable; then from (4.38) we have the $100(1-\alpha)\%$ confidence interval on $\mathbf{q'}\beta$ as

$$\mathbf{q'}\beta^0 \pm \hat{\sigma} t_{N-r,\alpha/2}\sqrt{\mathbf{q'Gq}}$$

where $t_{N-r,\alpha/2}$ is defined by the probability statement $\text{P}\{t \geq t_{N-r,\alpha/2}\} = \alpha/2$ for t having the t-distribution with $N-r$ degrees of freedom. When $N-r$ is large, ($N-r \geq 100$, say) $z_{\alpha/2}$ may be used in place of $t_{N-r,\alpha/2}$, where $z_{\alpha/2}$ is defined by

$$(2\pi)^{-\frac{1}{2}} \int_{z_{\alpha/2}}^{\infty} e^{-\frac{1}{2}x^2} dx = \alpha/2.$$

From $\text{SSE}/\sigma^2 \sim \chi^2_{N-r}$ of (4.51) a confidence interval on σ^2 is

$$\frac{\text{SSE}}{\chi^2_{N-r,U}} \leq \sigma^2 \leq \frac{\text{SSE}}{\chi^2_{N-r,L}}$$

where the denominators are defined by

$$\text{P}\{\chi^2_{k,L} \leq \chi^2_k \leq \chi^2_{k,U}\} = 1 - \alpha.$$

4.12 UNIQUE ESTIMATION USING RESTRICTIONS

In the case of models of the form such as $\text{E}[y_{ij}] = \mu + \alpha_i$ for $i = 1, 2, 3$, readers will undoubtedly have encountered constraints on the $\hat{\alpha}_i$s of the form

$$\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3 = 0 \qquad \text{or} \qquad \hat{\alpha}_3 = 0. \tag{4.64}$$

The second of these (and extensions thereof) is especially familiar to users of SAS GLM software where its use is standard practice.

Each equation in (4.64) is what we call a *linear constraint* on the solution. Careful use of such constraints can eliminate having many solutions $\boldsymbol{\beta}^0$ to the normal equations, and instead can yield just a single solution, one that satisfies the constraint(s) imposed. That being so, we denote such a solution as $\hat{\boldsymbol{\beta}}$. A brief outline for deriving $\hat{\boldsymbol{\beta}}$ follows. Some of the details are available in Searle (1971, Section 1.5b) and a complete description is given in Searle (1999).

In order to have a unique solution of the ML equations means that for $\mathbf{X}_{N\times p}$ of rank $r = p - m$ we need to have $\mathbf{H}$ in $\mathbf{H}\hat{\boldsymbol{\beta}} = \mathbf{c}$ being of order $m \times p$ of full row rank m. Then ML leads to minimizing $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + 2\boldsymbol{\theta}'(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{c})$ which yields equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{H}' \\ \mathbf{H} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{c} \end{bmatrix}, \tag{4.65}$$

where $2\boldsymbol{\theta}$ is a vector of Lagrange multipliers.

After considerable algebra (Searle, 1999) (4.65) yields

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}(\mathbf{X}'\mathbf{y} + \mathbf{H}'\mathbf{c}). \tag{4.66}$$

Because we have already established that ML generally yields a $\boldsymbol{\beta}$ as $\mathbf{G}\mathbf{X}'\mathbf{y}$ for some $\mathbf{G}$ [and $(\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}$ is a $\mathbf{G}$] we cannot call (4.66) an ML estimator because of the $\mathbf{H}'\mathbf{c}$ term therein. But (4.66) is an ML estimator under the constraints $\mathbf{H}\hat{\boldsymbol{\beta}} = \mathbf{c}$. And if $\mathbf{c} = \mathbf{0}$, which is often the case, then $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H})^{-1}\mathbf{X}'\mathbf{y}$ is an ML estimator.

Instead of having $\mathbf{H}\hat{\boldsymbol{\beta}} = \mathbf{c}$ as constraints on solutions, suppose that we have $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ as restrictions on parameters ("restrictions" rather than "constraints" to distinguish parameters from solutions). Then on partitioning $\mathbf{H}$ as $[\mathbf{H}_1 \quad \mathbf{H}_2]$ with $\mathbf{H}_1^{-1}$ existing (after perhaps permuting columns of $\mathbf{H}$ to permit this), we can rewrite $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ as

$$\boldsymbol{\beta}_1 = \mathbf{H}_1^{-1}\mathbf{c} - \mathbf{H}_1^{-1}\mathbf{H}_2\boldsymbol{\beta}_2. \tag{4.67}$$

This can then be substituted into $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$, from which the ML estimator of $\boldsymbol{\beta}_2$ can be obtained as

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'(\mathbf{y} - \mathbf{X}_1\mathbf{H}_1^{-1}\mathbf{c}) \tag{4.68}$$

where $\mathbf{S} = \mathbf{X}_2 - \mathbf{X}_1\mathbf{H}_1^{-1}\mathbf{H}_2$. Replacing $\boldsymbol{\beta}_2$ in (4.67) with $\hat{\boldsymbol{\beta}}_2$ of (4.68) then gives $\hat{\boldsymbol{\beta}}_1$, and the resulting $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$ is identical to (4.66) (Searle, 1999).

4.13 EXERCISES

E 4.1 Write the linear model equation $\mathrm{E}[y_{ij}] = \mu + \alpha_i + \beta_j$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ in matrix form by identifying $\mathbf{X}$ and $\boldsymbol{\beta}$. For $\mathbf{X}$ use Kronecker product notation (Appendix M).

E 4.2 Consider two different forms of simple linear regression, namely

$$\mathrm{E}[y_i] = \alpha + \gamma x_i \qquad \text{and} \qquad \mathrm{E}[y_i] = \mu + \gamma(x_i - \bar{x})$$

for $i = 1, 2, \ldots, n$:

(a) For each model write $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$, specifying $\mathbf{X}$ and $\boldsymbol{\beta}$.

(b) Obtain $\mathbf{X}'\mathbf{X}$, $(\mathbf{X}'\mathbf{X})^{-1}$ and $\hat{\boldsymbol{\beta}}$.

(c) Obtain $\mathrm{var}(\hat{\boldsymbol{\beta}})$. Do you see any advantage of one model equation over the other?

(d) Obtain the variance of the estimated value of the mean of y_i at a new value of x, denoted as x^*.

E 4.3 For $\mathbf{X}^-$ being a generalized inverse of $\mathbf{X}$, and for any $\mathbf{z}$ of appropriate order, show for $\boldsymbol{\beta}^0$ of (4.22) that

$$\boldsymbol{\beta}_{\mathbf{z}} = \boldsymbol{\beta}^0 + (\mathbf{I} - \mathbf{X}^-\mathbf{X})\mathbf{z}$$

is a solution of the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$.

E 4.4 From Appendix M, quote the two calculus results used in deriving (4.16), and explain how they are used.

E 4.5 For $\mathbf{X} = \left[\mathbf{1}_N \quad \left\{{}_d\, \mathbf{1}_{n_i}\right\}_{i=1}^{5}\right]$ show that $\mathbf{XGX}'$ has the same form as the example of Section 4.7.

E 4.6 For $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ with $\mathbf{X}$ of full rank, show that all linear combinations of $\boldsymbol{\beta}$ are estimable.

E 4.7 For the example of Section 4.7, and $\eta = \alpha_1 + 2.7\alpha_2 - 3.7\alpha_3$:

(a) Explain why η is estimable.

(b) What is the ML estimate of η?

E 4.8 For $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\ \sigma^2\mathbf{I})$ derive (4.38) for estimable $\mathbf{q}'\boldsymbol{\beta}$.

E 4.9 For the 1-way classification $\mathrm{E}[y_{ij}] = \mu + \alpha_i$:

(a) Show that $\sum_i \lambda_i \alpha_i$ is estimable if and only if $\sum_i \lambda_i = 0$.

(b) For $i = 1, 2, 3$ and $j = 1, 2$ derive the generalized inverses of $\mathbf{X}'\mathbf{X}$ which give

$$\boldsymbol{\beta}_1^0 = \begin{bmatrix} \bar{y}_{\cdot\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot} \\ \bar{y}_{3\cdot} - \bar{y}_{\cdot\cdot} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta}_2^0 = \begin{bmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{3\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{3\cdot} \\ 0 \end{bmatrix}.$$

(c) Verify that your answers in part (b) are indeed generalized inverses of $\mathbf{X}'\mathbf{X}$.

E 4.10 Show that

$$\tfrac{1}{6}\begin{bmatrix} 1 & 16 & 9 & -6 \\ -1 & -14 & -9 & 6 \\ -1 & -16 & -6 & 6 \\ -1 & -16 & -9 & 12 \end{bmatrix}$$

(a) is non-singular;

(b) is a generalized inverse of $\mathbf{X}'\mathbf{X}$ in (4.40);

(c) gives a solution of (4.40) very different from the solutions in (4.41), but yields the same estimators of $\mu + \alpha_i$.

E 4.11 Show that the mean squared error of $\hat{\sigma}^2$ is $\mathrm{var}(\hat{\sigma}^2)$ but that of $\hat{\sigma}^2_{\mathrm{ML}}$ is $\sigma^4[2N - 1 + (r_{\mathbf{X}} - 1)^2]/N$.

E 4.12 Derive (4.47).

E 4.13 Derive (4.58).

E 4.14 Derive (4.60).

E 4.15 Derive (4.62).

E 4.16 Simplify (4.63) for $r(\mathbf{K}') = 1$, and relate it to t of Section 4.11.

E 4.17 Maximize $L(\boldsymbol{\beta}, \sigma^2)$ subject to $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ to yield (4.58) and (4.59).

E 4.18 Show that Q and SSE have the distribution described below (4.63).

Chapter 5

GENERALIZED LINEAR MODELS (GLMs)

5.1 INTRODUCTION

Models for the analysis of non-normal data using nonlinear models have a long history. The use of probit regression for a binary response is a classic example. The word *probit* was traced by David (1995) as far back as Bliss (1934). Finney (1952) attributes the actual origin of probit regression to psychologists in the late 1800s.

In an early example of probit regression, Bliss (1934) describes an experiment in which nicotine is applied to aphids and the proportion killed is recorded (how is that for an early antismoking message?). As an appendix to a paper Bliss wrote a year later (Bliss, 1935), Fisher (1935) outlines the use of maximum likelihood to obtain estimates of the probit model.

However it was years before maximum likelihood estimation for probit models caught on. Finney (1952), in an appendix entitled "Mathematical basis of the probit method" gives some of the rational for maximum likelihood and motivates a computational method that he spends six pages describing in a different appendix.

More specifically, if we let p_i denote the probability of a success for the ith observation, the probit model is given by

$$\begin{aligned} y_i &\sim \text{indep. Bernoulli}(p_i) \\ p_i &= \Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right), \end{aligned} \tag{5.1}$$

where $\mathbf{x}_i'$ denotes the ith row of a matrix of predictors and $\Phi(\cdot)$ is the standard normal c.d.f. Considering the scalar functions applied elementwise to the vectors, we can rewrite (5.1) as

$$\begin{aligned} \mathbf{y} &\sim \text{indep. Bernoulli}(\mathbf{p}) \\ \mathbf{p} &= \Phi(\mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{5.2}$$

or equivalently

$$\Phi^{-1}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}$ is the model matrix. The use of the inverse standard normal c.d.f., known as the probit, to transform the mean of $\mathbf{y}$ to the linear predictor is attractive on two counts. First, it expands the range of $\mathbf{p}$ from [0,1] to the whole real line, making it more reasonable to assume a model of the form $\mathbf{X}\boldsymbol{\beta}$. Second, in many problems, the sigmoidal form of $\mathbf{p}$ as a function of the covariates is often observed in practice.

Finney (1952) suggested calculating an estimate of $\boldsymbol{\beta}$ via an iteratively weighted least squares algorithm. He recommended using *working probits* which he defined (ignoring the shift of five units historically used to keep all the calculations positive) as

$$t_i = \mathbf{x}_i'\boldsymbol{\beta} + \frac{y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta})}{\phi(\mathbf{x}_i'\boldsymbol{\beta})}, \tag{5.3}$$

where $\phi(\cdot)$ is the standard normal probability density function (p.d.f.). The working probits for a current value of $\boldsymbol{\beta}$ were regressed on the predictors using weights given by $\dfrac{[\phi(p_i)]^2}{\Phi(p_i)[1-\Phi(p_i)]}$ (see E 5.1) in order to get the new value of $\boldsymbol{\beta}$. This algorithm was iterated until convergence (or at least until the computer – a person! – got tired of performing the calculations).

Nelder and Wedderburn (1972) recognized that the working probits could be generalized in a straightforward way to unify an entire collection of maximum likelihood problems. This *generalized linear model* (GLM) could handle probit or logistic regression, Poisson regression, log-linear models for contingency tables, variance components estimation from ANOVA mean squares and many other problems in the same way.

They replaced $\Phi^{-1}(\cdot)$ with a general *link* function, $g(\cdot)$, which transforms (or links) the mean of y_i to the linear predictor. With $g_\mu(\mu)$ representing $\partial g(\mu)/\partial\mu$, they then defined a *working variate* via

$$\begin{aligned} t_i &\equiv g(\mu_i) + g_\mu(\mu_i)(y_i - \mu_i) \\ &= \mathbf{x}_i'\boldsymbol{\beta} + g_\mu(\mu_i)(y_i - \mu_i). \end{aligned} \tag{5.4}$$

Since the second term on the right-hand side of (5.4) has expectation zero it can be regarded as an *error term* so that t_i follows a linear model, albeit with unequal variances which depend on the unknown $\boldsymbol{\beta}$. This suggests using (5.4) just like (5.3): regress $\mathbf{t}$ on $\mathbf{X}$ using a weighted linear regression (more details are given in Section 5.4e) and iterate until the estimates of $\boldsymbol{\beta}$ stabilize.

More important, it made possible a style of thinking which freed the data analyst from necessarily looking for a transformation which simultaneously achieved linearity in the predictors and normality of the distribution (as in Box and Cox, 1962).

What advantages does this have? First, it unifies what appear to be very different methodologies, which helps us to understand, use and (for those of us in the business) teach the techniques. Second, since the right-hand side of the model equation is a linear model after applying the link, many of the standard ways of thinking about linear models carry over to GLMs.

5.2 STRUCTURE OF THE MODEL

Building a generalized linear model involves three decisions:

1. What is the distribution of the data (for fixed values of the predictors and possibly after a transformation)?

2. What function of the mean will be modeled as linear in the predictors?

3. What will the predictors be?

a. Distribution of y

Typically the vector $\mathbf{y}$ is assumed to consist of independent measurements from a distribution with density from the exponential family or

similar to the exponential family:

$$\begin{aligned} y_i &\sim \text{indep. } f_{Y_i}(y_i) \\ f_{Y_i}(y_i) &= \exp\{[y_i\gamma_i - b(\gamma_i)]/\tau^2 - c(y_i,\tau)\}, \end{aligned} \tag{5.5}$$

where, for convenience, we have written the distribution in what is called *canonical form.* For example, for the probit model, the data would be independent Bernoulli so that $f_{Y_i}(y_i)$ would be $p_i^{y_i}(1-p_i)^{1-y_i}$, where p_i is the probability of a success and $\gamma_i = \log[p_i/(1-p_i)]$. Most commonly-used distributions can be written in the form (5.5) (see E 5.2).

b. Link function

We typically want to relate the parameters of the distribution to various predictors. We do so by modeling a transformation of the mean, μ_i, which would be some function of γ_i, as a linear model in the predictors:

$$\begin{aligned} \mathrm{E}[y_i] &= \mu_i \\ g(\mu_i) &= \mathbf{x}_i'\boldsymbol{\beta}, \end{aligned} \tag{5.6}$$

where $g(\cdot)$ is a known function, called the *link function* (since it links together the mean of y_i and the linear form of predictors), $\mathbf{x}_i'$ is the ith row of the model matrix, and $\boldsymbol{\beta}$ is the parameter vector in the linear predictor. In the probit example $g(\mu) = \Phi^{-1}(\mu)$ and $\mu = 1/(1 + \exp[-\gamma])$.

c. Predictors

In practice, of course, one must make decisions as to which predictors to include on the right-hand side of (5.6) and in what form to include them. For example, in the classic paper of Bliss (1934) the suggested predictor of survival is log nicotine dose as opposed to nicotine itself.

A key point in using GLMs is that many of the considerations in modeling are the same as for LMMs since the right-hand sides of the model equations for the mean are the same. For example, issues of how to represent predictors and interactions, whether and how to model nonlinear relationships and (as we will see in Chapter 8) the incorporation of random factors.

d. Linear models

This generalized class of models subsumes the linear model of Chapter 4 as a special case. The normal distribution can be written in the form (5.5) by defining:

$$\begin{aligned} \gamma_i &= \mu_i \\ b(\gamma_i) &= \tfrac{1}{2}\mu_i^2 \\ \tau^2 &= \sigma^2 \\ c(y_i, \tau) &= \tfrac{1}{2}\log 2\pi\sigma^2 + \tfrac{1}{2}y_i^2/\sigma^2. \end{aligned} \tag{5.7}$$

With $g(\mu_i) = \mu_i$ and $\mu_i = \mathbf{x}_i'\beta$ we generate the linear model of Section 4.3.

5.3 TRANSFORMING VERSUS LINKING

In its earliest incarnations, probit analysis was little more than a transformation technique. It was realized that the frequent sigmoidal shape in plots of observed proportions of successes plotted against a predictor x could be made into a straight line by applying a transformation corresponding to the inverse of the normal c.d.f. However, one of the main ideas of GLMs is to get away from the idea of transforming the data. The strategy, then, is to apply a link function to the mean of the response and fit the resulting model by the method of maximum likelihood.

5.4 ESTIMATION BY MAXIMUM LIKELIHOOD

a. Likelihood

The log likelihood for (5.5) is given by

$$l = \sum_{i=1}^{n}[y_i\gamma_i - b(\gamma_i)]/\tau^2 - \sum_{i=1}^{n} c(y_i, \tau). \tag{5.8}$$

b. Some useful identities

Before we derive the maximum likelihood equations it is useful to establish some identities. These flow from the results

$$\mathrm{E}\left[\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\right] = 0, \tag{5.9}$$

and

$$\mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\right) = -\mathrm{E}\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \gamma_i^2}\right], \tag{5.10}$$

which require regularity conditions (Casella and Berger, 1990, p. 308). Using (5.5) in (5.9) gives

$$\mathrm{E}\left[\left\{y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\} \Big/ \tau^2\right] = 0 \tag{5.11}$$

or

$$\mathrm{E}[y_i] = \mu_i = \frac{\partial b(\gamma_i)}{\partial \gamma_i}. \tag{5.12}$$

And using (5.5) in (5.10) we obtain

$$\mathrm{var}\left(\left\{y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\} \Big/ \tau^2\right) = -\mathrm{E}\left[-\frac{1}{\tau^2}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right], \tag{5.13}$$

which, using (5.12) gives

$$\mathrm{var}\left(\frac{y_i - \mu_i}{\tau^2}\right) = \frac{1}{\tau^2}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}$$

or

$$\begin{aligned} \mathrm{var}(y_i) &= \tau^2 \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2} \\ &\equiv \tau^2 v(\mu_i), \end{aligned} \tag{5.14}$$

wherein we define $v(\mu_i)$ as $\partial^2 b(\gamma_i)/\partial \gamma_i^2$. Note that $v(\mu_i)$ is often called the *variance function*, since it indicates how the variance of y_i depends on the mean of y_i. Two other useful identities are

$$\frac{\partial \gamma_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \gamma_i}\right)^{-1} = \left(\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right)^{-1} = \frac{1}{v(\mu_i)} \tag{5.15}$$

and, using the chain rule and (5.6),

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial g(\mu_i)} \frac{\partial g(\mu_i)}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} \frac{\partial \mathbf{x}_i' \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}$$

$$= \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} \mathbf{x}_i' \ . \tag{5.16}$$

As an illustration of these results, consider the linear model in Section 5.1d. With subscripts denoting derivatives we have $b_\gamma(\gamma_i)$ equal to μ_i, the mean, and $b_{\gamma\gamma}(\gamma_i) = 1$ so that, from (5.14), $\text{var}(y_i) = \tau^2 b_{\gamma\gamma}(\gamma_i) = \sigma^2$, as expected. Also, $\partial\gamma_i/\partial\mu_i = \partial\mu_i/\partial\mu_i = 1 = v(\mu_i)^{-1}$, verifying (5.15) and, with $g_\mu(\mu_i) = 1$, $\partial\mu_i/\partial\boldsymbol{\beta} = \mathbf{x}_i'$ as in (5.16). Note that the normal distribution has an unusual feature among distributions given by (5.5): its variance is a constant and not a function of the mean.

c. Likelihood equations

We are now in a position to derive the maximum likelihood equations for $\boldsymbol{\beta}$. From (5.8) we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2} \sum \left[y_i \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} - \frac{\partial b(\gamma_i)}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} \right]$$

$$= \frac{1}{\tau^2} \sum (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} \qquad \text{using (5.12)}$$

$$= \frac{1}{\tau^2} \sum (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \qquad \text{using the chain rule}$$

$$= \frac{1}{\tau^2} \sum \frac{(y_i - \mu_i)}{v(\mu_i) g_\mu(\mu_i)} \mathbf{x}_i' \qquad \text{using (5.15) and (5.16)}$$

$$= \frac{1}{\tau^2} \sum (y_i - \mu_i) w_i g_\mu(\mu_i) \mathbf{x}_i', \tag{5.17}$$

upon defining $w_i = [v(\mu_i) g_\mu^2(\mu_i)]^{-1}$.

We can write this in matrix notation as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2} \mathbf{X}' \mathbf{W} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}), \tag{5.18}$$

with $\mathbf{W} = \left\{ {}_d w_i \right\}$ and $\mathbf{\Delta} = \left\{ {}_d g_\mu(\mu_i) \right\}$.

The ML equations are thus given by

$$\mathbf{X}'\mathbf{W}\mathbf{\Delta}\mathbf{y} = \mathbf{X}'\mathbf{W}\mathbf{\Delta}\boldsymbol{\mu}, \tag{5.19}$$

where $\mathbf{W}, \mathbf{\Delta}$ and $\boldsymbol{\mu}$ involve the unknown $\boldsymbol{\beta}$. Typically these are nonlinear functions of $\boldsymbol{\beta}$ and so (5.19) cannot be solved analytically.

For example, for the probit model of (5.2), the log likelihood and its derivative are

$$l = \sum \left(y_i \left\{ \log \Phi(\mathbf{x}_i'\boldsymbol{\beta}) - \log[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] \right\} + \log[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})] \right) \tag{5.20}$$

and

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum \left[y_i \left(\frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})} \mathbf{x}_i' + \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})}{1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})} \mathbf{x}_i' \right) - \frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})}{1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})} \mathbf{x}_i' \right] \\ &= \sum \frac{[y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]\phi(\mathbf{x}_i'\boldsymbol{\beta})}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]} \mathbf{x}_i' \\ &= \sum \frac{(y_i - \mu_i)\phi(\mathbf{x}_i'\boldsymbol{\beta})}{\mu_i(1 - \mu_i)} \mathbf{x}_i'. \end{aligned} \tag{5.21}$$

Identifying $b(\gamma_i)$ of (5.5) as $\log(1+e^{\gamma_i})$ so that $b_\gamma(\gamma_i) = (1+e^{-\gamma_i})^{-1} = \mu_i$ and $b_{\gamma\gamma}(\gamma_i) = \mu_i(1 - \mu_i)$, it is straightforward (see E 5.4) to show that (5.21) is of the form of (5.18).

For solving the ML equations or for deriving the large-sample variance of $\hat{\boldsymbol{\beta}}$, it is useful to have the expected value of the second derivative of the log likelihood:

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} = -\frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{\Delta} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'} + \frac{1}{\tau^2} \mathbf{X}' \frac{\partial \mathbf{W}\mathbf{\Delta}}{\partial \boldsymbol{\beta}'} (\mathbf{y} - \boldsymbol{\mu}) \tag{5.22}$$

so that

$$\begin{aligned} -\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} \right] &= \frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{\Delta} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}'} + \mathbf{0} \\ &= \frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{\Delta}\mathbf{\Delta}^{-1}\mathbf{X} \quad \text{using (5.16)} \\ &= \frac{1}{\tau^2} \mathbf{X}'\mathbf{W}\mathbf{X}, \end{aligned} \tag{5.23}$$

where, again, $\mathbf{W} = \left\{ {}_d w_i \right\} = \left\{ {}_d [v(\mu_i) g_\mu^2(\mu_i)]^{-1} \right\}$.

d. Large-sample variances

To derive the large-sample variance of $\hat{\boldsymbol{\beta}}$ we first note that

$$\begin{aligned} -\mathrm{E}\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \tau^2}\right] &= -\mathrm{E}\left[\frac{\partial}{\partial \tau^2}\frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})\right] \\ &= \frac{1}{\tau^4}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}\mathrm{E}\left[\mathbf{y}-\boldsymbol{\mu}\right] \qquad (5.24) \\ &= \mathbf{0}, \end{aligned}$$

so that estimation of τ^2 does not affect the large-sample variance of $\hat{\boldsymbol{\beta}}$. The usual large-sample arguments (see Section S.4c of Appendix S), along with (5.23) and (5.24), show that (see E 5.6)

$$\mathrm{var}_\infty(\hat{\boldsymbol{\beta}}) = \tau^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}, \qquad (5.25)$$

where var_∞ indicates the limiting or asymptotic variance.

e. Solving the ML equations

Solution of the ML equations, (5.19), for $\boldsymbol{\beta}$ is usually performed by an iterative weighted least squares method. This can be derived as an example of the use of Fisher scoring (Searle et al., 1992, p. 295). Fisher scoring is an iterative method for maximizing a likelihood and it takes the form

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \mathbf{I}(\boldsymbol{\theta}^{(m)})^{-1}\left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}, \qquad (5.26)$$

where (m) indicates the mth iteration, $\mathbf{I}(\boldsymbol{\theta})$ is the information matrix and $\boldsymbol{\theta}$ is the entire parameter vector.

Using (5.24), (5.23), and (5.18), the portion of the equation for $\boldsymbol{\beta}$ (see E 5.7) is of the form

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}), \qquad (5.27)$$

where it is understood that $\mathbf{W}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\mu}$ are evaluated at $\boldsymbol{\beta}^{(m)}$.

How does this relate to the working variate of (5.4)? We have

$$\mathbf{t} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) \qquad (5.28)$$

so that, with the use of (5.14)

$$\mathrm{var}(\mathbf{t}) = \mathrm{var}[\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})] = \left\{{}_d\tau^2 v(\mu_i) g'^2_\mu(\mu_i)\right\} = \tau^2\mathbf{W}^{-1}, \qquad (5.29)$$

so a weighted regression of $\mathbf{t}$ on $\mathbf{X}$ using weights equal to the inverse of the variance of $\mathbf{t}$ gives

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}[\mathbf{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})] \\ &= \boldsymbol{\beta}^{(m)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}), \qquad (5.30)\end{aligned}$$

which is the same as (5.27).

f. Example: Potato flour dilutions

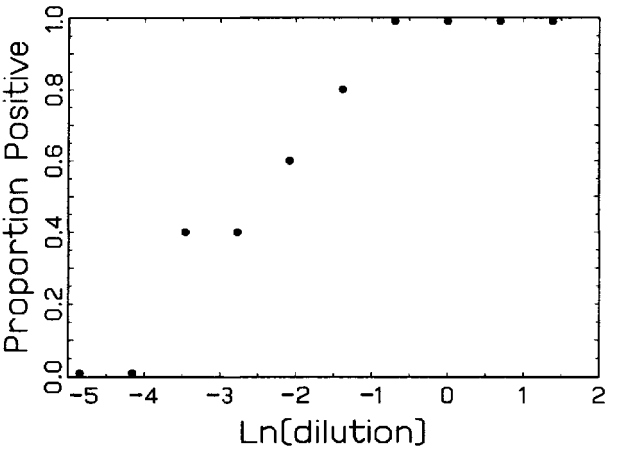
Finney (1971) gives an example of the growth of spores in a potato flour suspension. For each of 10 dilutions, five plates are tested for positive growth. The data are given in Table 5.1. As the flour suspensions get more concentrated, the probability of growth (i.e., proportion of positive plates) increases. Figure 5.1 shows that the probability of response, as a function of the natural logarithm of dilution, follows a roughly sigmoidal shape, so we might entertain a logistic regression model. Let y_i denote the number of plates out of five that show a positive response. A possible model is

Table 5.1: Potato Flour Data

Dilution	Spore Growth		Proportion
(g/100 ml	No. of Plates	No. Positive	of Residual Plates
1/128	5	0	0.0
1/64	5	0	0.0
1/32	5	2	0.4
1/16	5	2	0.4
1/8	5	3	0.6
1/4	5	4	0.8
1/2	5	5	1.0
1	5	5	1.0
2	5	5	1.0
4	5	5	1.0

$$\begin{aligned}\mathrm{E}[y_i] &= 5\pi(x_i) = 5\frac{1}{1+e^{-(\alpha+\beta x_i)}} \qquad (5.31)\\ y_i &\sim \text{indep. binomial } [5, \pi(x_i)].\end{aligned}$$

Figure 5.1: Proportion of positive spore growth plotted against log dilution for the potato flour data.



The log likelihood for this model is given by

$$l = \sum \left[\log \binom{5}{y_i} + y_i(\alpha + \beta x_i) - 5\log(1 + e^{\alpha+\beta x_i}) \right]$$

$$= c + \alpha \sum y_i + \beta \sum y_i x_i - 5 \sum \log(1 + e^{\alpha+\beta x_i}), \qquad (5.32)$$

where $c = \sum \binom{5}{y_i}$ is a function of the y_i but not of α and β. The log likelihood is shown as a function of α and β in Figure 5.2. The ML equations are thus given by

$$\sum y_i = \sum \frac{5}{1 + e^{-(\hat{\alpha}+\hat{\beta}x_i)}}$$

$$\sum y_i x_i = \sum \frac{5x_i}{1 + e^{-(\hat{\alpha}+\hat{\beta}x_i)}}. \qquad (5.33)$$

With $\sum y_i = 31$ and $\sum y_i x_i \doteq -17.329$ it is merely tedious arithmetic to verify that $\hat{\alpha} = 4.17$ and $\hat{\beta} = 1.62$ solve these equations to within rounding error. Figure 5.3 plots the data and fitted values.

To illustrate the large-sample variance calculation note that

$$\tau^2 = 1$$

Figure 5.2: Log likelihood plotted against parameters for the potato flour data.

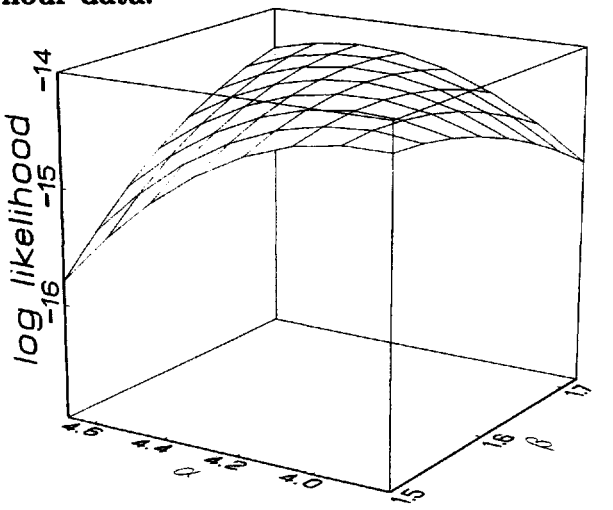
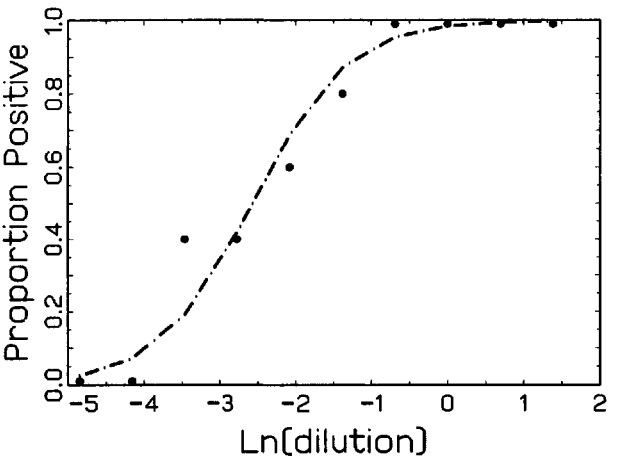


Figure 5.3: Proportion positive versus log dilution for the potato flour data.



$$v(\mu_i) = \mu_i(1-\mu_i)$$

$$g_\mu(\mu_i) = 1/v(\mu_i)$$

so that $\mathbf{W} = \left\{{}_d\mu_i(1-\mu_i)\right\}$. We thus have

$$\mathbf{X'WX} = \begin{bmatrix} \sum \mu_i(1-\mu_i) & \sum x_i\mu_i(1-\mu_i) \\ \sum x_i\mu_i(1-\mu_i) & \sum x_i^2\mu_i(1-\mu_i) \end{bmatrix}$$

$$= \begin{bmatrix} 4.38306 & -11.09943 \\ -11.09943 & 32.89365 \end{bmatrix}$$

with inverse

$$(\mathbf{X'WX})^{-1} = \begin{bmatrix} 1.56805 & 0.52911 \\ 0.52911 & 0.20894 \end{bmatrix}.$$

This gives

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{AN}\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1.56805 & 0.52911 \\ 0.52911 & 0.20894 \end{pmatrix} \right].$$

5.5 TESTS OF HYPOTHESES

a. Likelihood ratio tests

Likelihood ratio tests follow the usual prescription of comparing the maximized values of the log likelihood both under H_0 and not restricted to H_0. If the difference is large (i.e., the unrestricted model fit is much better) then H_0 is rejected.

When there are multiple parameters we will often be interested in hypotheses concerning only a subset of the parameters. Accordingly, let the parameter vector $\boldsymbol{\theta}$ be partitioned into two components $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ and suppose interest focuses on $\boldsymbol{\theta}_1$ while $\boldsymbol{\theta}_2$ is left unspecified. $\boldsymbol{\theta}_2$ is often called a *nuisance parameter*. Either or both of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ could be vector-valued and, if the entire parameter vector is of interest, $\boldsymbol{\theta}_2$ could be null.

Suppose our hypothesis is of the form $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$, where $\boldsymbol{\theta}_{1,0}$ is a specified value of $\boldsymbol{\theta}_1$, and let $\hat{\boldsymbol{\theta}}_{2,0}$ be the MLE of $\boldsymbol{\theta}_2$ under the restriction that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$. The likelihood ratio test statistic is given by

$$-2\log\Lambda = -2\left[l(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0}) - l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right], \tag{5.34}$$

where $\hat{\boldsymbol{\theta}}' = (\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2')$ and the large-sample critical region of the test is to reject H_0 in favor of the alternative when

$$-2\log\Lambda > \chi^2_{\nu,1-\alpha}, \tag{5.35}$$

where ν is the dimension of $\boldsymbol{\theta}_1$.

b. Wald tests

An alternative method of testing is to use the large-sample normality of the ML estimator in order to form a test. From standard results (Appendix S)

$$\hat{\boldsymbol{\theta}} \sim \mathcal{AN}[\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})], \tag{5.36}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information for $\hat{\boldsymbol{\theta}}$. Again, if we write $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$, and write conformably

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix} \tag{5.37}$$

then standard matrix algebra for partitioned matrices (Searle, 1982, p. 354) and multivariate normal calculations show that the large-sample variance of $\hat{\boldsymbol{\theta}}_1$ is given by

$$\text{var}_\infty(\hat{\boldsymbol{\theta}}_1) = \left(\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}\right)^{-1}. \tag{5.38}$$

To test $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$ we form the Wald statistic

$$W = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0})'[\text{var}_\infty(\hat{\boldsymbol{\theta}}_1)]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}), \tag{5.39}$$

which, under H_0, has the same large-sample χ^2 distribution as the LRT with degrees of freedom equal to the dimension of $\boldsymbol{\theta}_1$. More explicitly we would reject the $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$ if

$$W > \chi^2_{\nu,1-\alpha}. \tag{5.40}$$

Both the LRT and the Wald tests are available to test the same hypotheses and have the same limiting distribution. What are the differences? For large samples, and if the deviation from the null hypothesis is not too extreme, the two test statistics will give similar, though not identical results (Bishop et al., 1975, Sec. 14.9). However, for small samples or for extreme deviations, they can differ. Generally, investigations have shown (Cox and Hinkley, 1974; McCullagh and Nelder,

1989) that use of the large sample-distribution for the LRT gives a more accurate approximation for small and moderate-sized samples than for the Wald test. The LRT is thus to be preferred. The Wald test does, however, have a computational advantage since it does not require calculation of $\hat{\boldsymbol{\theta}}_{2,0}$.

c. Illustration of tests

We use the potato flour data to illustrate these tests for the null hypothesis $H_0: \beta = 0$, i.e., no relationship between spore growth and log dilution. To perform the likelihood ratio test we must maximize the likelihood under the null hypothesis, that is, when the probability of growth is constant. Under $H_0, \hat{\alpha} \doteq 0.4896$ (see E 5.5). We thus have

$$l(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0}) = l(0.50, 0) = -33.20$$

while

$$l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = l(\hat{\alpha}, \hat{\beta}) = l(4.17, 1.62) = -14.214.$$

The LRT statistic is thus $-2 \log \Lambda = -2[-33.20 - (-14.21)] = 37.88$. The statistic has 1 degree of freedom, which is the dimension of β. So we easily reject H_0 at any usual level of significance and the p-value is $\mathrm{P}\{\chi_1^2 \geq 37.88\} \doteq 0$.

The Wald test statistic uses $\hat{\beta} = 1.62$ from below (5.33) and $\mathrm{var}(\hat{\beta})_\infty = 0.2089$ from the end of Section 5.4. Substituting in (5.39) we then have $W = (1.62)(0.2089)^{-1}(1.62) = 1.62^2/0.2089 = 12.6$. This has a p-value of $\mathrm{P}\{\chi_1^2 \geq 12.6\} \doteq 0.0004$, which again corresponds to rejection of the null hypothesis at the usual significance levels. This illustrates that the two test statistics need not be numerically similar for large deviations from the null hypothesis. Of course, in such situations the same qualitative conclusion would ordinarily be reached.

d. Confidence intervals

Either the LRT or Wald test can be used to construct large-sample confidence intervals for $\boldsymbol{\theta}_1$. For the LRT we include in the confidence set all values $\boldsymbol{\theta}_1$ such that

$$-2\left[l(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2,1}) - l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right] \leq \chi^2_{\nu,1-\alpha}. \tag{5.41}$$

In (5.41) $\hat{\boldsymbol{\theta}}_{2,1}$ represents the MLE of $\boldsymbol{\theta}_2$ for each value of $\boldsymbol{\theta}_1$ checked for inclusion in the set.

For the Wald test we include in the confidence set all values of $\boldsymbol{\theta}_1$ such that

$$(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)'[\text{var}_\infty(\hat{\boldsymbol{\theta}}_1)]^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \leq \chi^2_{\nu,1-\alpha}. \tag{5.42}$$

The computational burden of the likelihood-based confidence interval is thus larger than that for the Wald-based interval. However, the small and moderate-sized sample performance of the LRT-based confidence region has generally been found to be better.

e. Illustration of confidence intervals

The likelihood-based confidence interval solves for the values of β such that

$$-2[l(\beta, \hat{\alpha}_\beta) - l(\hat{\alpha}, \hat{\beta})] \leq 3.84,$$

where $\hat{\alpha}_\beta$ denotes the MLE of α when β is fixed at some value. Numerical calculations give the interval as (0.90, 2.76).

The Wald-based confidence interval for β is straightforward since it is based on

$$\hat{\beta} \sim \mathcal{AN}(\beta, 0.2089),$$

which gives a confidence interval of $1.62 \pm 1.96(0.2089)^{1/2} = (0.72, 2.52)$. The LR based interval is approximately the same length as the Wald interval but is not symmetrically placed about the MLE, an indication of the non-normality of the sampling distribution.

5.6 MAXIMUM QUASI–LIKELIHOOD

a. Introduction

In some statistical investigations, such as the potato flour example of Section 5.5, we know the distribution of the data (binomial with $n = 5$ in that instance). In others we are less certain. For example, in analyzing data on costs of hospitalization we know the data are positive (though it would be nice to be paid for some hospital ordeals!) and they are invariably skewed right. With a little more experience with such data we would know that the variance increases with the mean and we might have a rough idea as to how quickly it increases. However, we are unlikely to know *a priori* exactly what distributional form is correct or even likely to fit well. But not knowing the distribution makes it impossible to construct a likelihood and thus to use such techniques as maximum likelihood and likelihood ratio tests.

It would therefore be useful to have inferential methods which work as well or almost as well as ML but without having to make specific distributional assumptions. This is the basic idea behind *quasi-likelihood*: to derive a likelihood-like quantity whose construction requires few assumptions.

What are the important characteristics of likelihood which are required to generate workable estimators? It turns out to be easier to mimic the properties of the derivative of the log likelihood (also called the *score function*) rather than the likelihood itself.

b. Definition

We define an analog of likelihood using (5.9) and (5.10), except that we differentiate with respect to μ_i instead of γ_i. First, from (5.9) we want

$$\mathrm{E}\left[\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right] = 0. \tag{5.43}$$

Then we observe that by the chain rule, what we will denote as v^* is

$$\begin{aligned} v^* = \mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) &= \mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\frac{\partial \gamma_i}{\partial \mu_i}\right) \\ &= \left[\mathrm{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \gamma_i}\right)\right]\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2 \end{aligned} \tag{5.44}$$

and using (5.10)

$$= \left(-E\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \gamma_i^2}\right]\right)\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2.$$

Now, by the nature of $f_{Y_i}(y_i)$ in (5.5), with $b(\gamma_i)$ containing no data this is

$$v^* = \frac{1}{\tau^2}\left[\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right]\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2, \tag{5.45}$$

and from the definition of $v(\mu_i)$ below (5.14) this becomes

$$\begin{aligned} v^* &= \frac{v(\mu_i)}{\tau^2}\left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2 \\ &= \frac{v(\mu_i)}{\tau^2}\frac{1}{v(\mu_i)^2} \qquad \text{from (5.15).} \end{aligned} \tag{5.46}$$

Thus

$$\text{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) = \frac{1}{\tau^2 v(\mu_i)}, \tag{5.47}$$

or, by (5.10) and using $\partial\mu_i$ in place of $\partial\gamma_i$,

$$\text{var}\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) = -\text{E}\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \mu_i^2}\right] = \frac{1}{\tau^2 v(\mu_i)}. \tag{5.48}$$

Observe that (5.43) and (5.48) are the analogs of (5.9) and (5.10).

We thus seek a quantity in place of $\partial \log f_{Y_i}(y_i)/\partial\mu_i$ which has properties (5.43) and (5.48). It is straightforward to verify (see E 5.8) that

$$q_i = \frac{y_i - \mu_i}{\tau^2 v(\mu_i)} \tag{5.49}$$

satisfies these same conditions, where we assume that $\text{var}(y_i) \propto v(\mu_i)$. The τ occurring in (5.49) is merely the (unspecified) constant of proportionality relating $\text{var}(y_i)$ to $v(\mu_i)$, which is not exactly the same as the τ that appears in the density (5.5). However, we will use the same notation since, as we see below, they play the same role.

Since the contribution to the log likelihood from y_i is the integral with respect to μ_i of $\partial \log f_{Y_i}(y_i)/\partial\mu_i$, we define the log quasi–likelihood via the contribution y_i makes to it:

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\tau^2 v(t)} dt, \tag{5.50}$$

which, by definition, has derivative with respect to μ_i equal to q_i. Finally, to find the *maximum quasi–likelihood* (MQL) estimator of $\boldsymbol{\beta}$ we solve the *maximum quasi–likelihood equations*

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum Q_i = \mathbf{0}. \tag{5.51}$$

Evaluating the derivative in (5.51) gives

$$\sum \frac{y_i - \mu_i}{\tau^2 v(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

which, using (5.16), is the same as

$$\sum \frac{y_i - \mu_i}{\tau^2 v(\mu_i) g_\mu(\mu_i)} \mathbf{x}_i' = \mathbf{0}, \tag{5.52}$$

or, in matrix notation,

$$\frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})=\mathbf{0}, \tag{5.53}$$

the same as (5.18). Note that by defining maximum quasi–likelihood estimators as solutions to the maximum quasi–likelihood equations, (5.51), we avoid a true maximization problem or even the definition of a quasi–likelihood or log quasi–likelihood itself.

In some ways this is a remarkable result. Q_i is constructed using only information about how the variance changes with the mean and nothing more. And, it is often the case that if we specify a mean-to-variance relationship, we obtain maximum quasi–likelihood equations which are exactly the same as those corresponding to a legitimate likelihood.

For example, suppose we are willing to assume the mean and variance are equal, so that what we build into quasi–likelihood is the fact that $v(\mu_i)=\mu_i$. Note that this allows the variance to be merely proportional to the mean rather than exactly equal to it, so that

$$\begin{aligned} Q_i &= \int_{y_i}^{\mu_i}\frac{y_i-t}{\tau^2 t}dt, \\ &= \frac{1}{\tau^2}(y_i\log t-t)\Big|_{y_i}^{\mu_i} \\ &= \frac{1}{\tau^2}(y_i\log\mu_i-\mu_i-y_i\log y_i+y_i) \end{aligned} \tag{5.54}$$

and the MQL equations for $\boldsymbol{\beta}$ are

$$\frac{\partial}{\partial\boldsymbol{\beta}}\sum(y_i\log\mu_i-\mu_i)=\mathbf{0} \tag{5.55}$$

(the other terms dropping out).

Instead of merely assuming that $v(\mu_i)=\mu_i$ suppose we make the assumption that $y_i\sim\text{Poisson}(\mu_i)$, which would actually force $\text{var}(y_i)=\mu_i$ as well. Then $\log f_{Y_i}(y_i)=y_i\log\mu_i-\mu_i-\log(y_i!)$ and the ML equations would be

$$\frac{\partial}{\partial\boldsymbol{\beta}}\sum(y_i\log\mu_i-\mu_i)=\mathbf{0}, \tag{5.56}$$

which are the same as the MQL equations, (5.55)! In this case MQL and ML would give exactly the same estimates and hence MQL would

be fully efficient. In other cases (see E 5.3) ML does not give equations of the form (5.19) and, in those cases, MQL may not be fully efficient. See exercise E 5.10 for some simple calculations and Firth (1987) for more detail.

MQL has important advantages over ML. To explain, consider again the specific situation of regression with a Poisson-distributed response. ML would assume $\text{var}(y_i) = v(\mu_i)$. However, in practice it is often true that data appear selected from a distribution in which the variance is larger than the mean. If the variance is proportional to the mean, the specification of the model under quasi–likelihood is still correct because the assumption is only that $\text{var}(y_i) = \tau^2 v(\mu_i)$; that is, $\text{var}(y_i)$ is proportional to $v(\mu_i)$, not necessarily equal.

Thus MQL affords us two degrees of robustness. First, we need not make a distributional assumption and second, we have only to specify the mean-to-variance relationship up to a proportionality constant which can be estimated from the data (see below).

Inference using MQL proceeds much as ML for β. Under mild conditions (McCullagh, 1983) it can be shown that

$$\tilde{\beta} \sim \mathcal{AN}\left[\beta, \tau^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right], \tag{5.57}$$

with $\tilde{\beta}$ being the MQL estimator of β and, as we defined before, $\mathbf{W} = \left\{{}_d\,[v(\mu_i)g_\mu^2(\mu_i)]^{-1}\right\}$.

However, τ is usually handled differently and estimated via a moment estimator (McCullagh and Nelder, 1989, p. 328):

$$\hat{\tau}^2 = \frac{1}{n-p}\sum \frac{(y_i - \hat{\mu}_i)^2}{v(\mu_i)}, \tag{5.58}$$

where n is the number of observations and p is the dimension of β.

5.7 EXERCISES

E 5.1 Show that $\dfrac{\phi(p_i)^2}{\Phi(p_i)[1-\Phi(p_i)]}$ is the inverse of an estimate of $\text{var}(t_i)$, where t_i is defined in (5.3).

E 5.2 Show that the binomial, Poisson and gamma distributions can be written in the form (5.5). *Hint for the gamma distribution*: Write the density in terms of the mean and coefficient of variation.

E 5.3 Suppose $y \sim \mathcal{N}(e^\theta, e^\theta)$, i.e., y is normal with equal mean and variance. Show that the distribution of y is *not* of the form (5.5).

E 5.4 Show that (5.21) can be written in the form (5.18).

E 5.5 Suppose $y_i \sim$ indep. Binomial(n,p) for $i = 1,2,\ldots,m$, where $p = 1/(1+e^{-\alpha})$. Show that the MLE of α is $\log[\sum y_i/(mn-\sum y_i)]$.

E 5.6 Using (5.24) verify that the large-sample variance of $\hat{\beta}$ is given by (5.25).

E 5.7 Derive (5.27) from (5.26).

E 5.8 Show that q_i of (5.49) satisfies (5.51), (5.52), and (5.53).

E 5.9 For binary (Bernoulli) and Poisson distributed data, in (5.19) show that $\mathbf{W\Delta} = \mathbf{I}$ and hence it simplifies to

$$\mathbf{X'y} = \mathbf{X'}\boldsymbol{\mu}.$$

E 5.10 *Efficiency of MQL*: Suppose that $y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1,2,\ldots n$, where $\log \mu_i = x_i\beta$ and $v(\mu_i) = \mu_i$. Calculate the ratio of the large-sample variances of $\tilde{\beta}$, the MQL estimator of β and $\hat{\beta}$, the MLE of β. For concreteness, assume that $n/2$ of the observations have $x_i = 5$ and $n/2$ are 10. Do the calculations for β equal to 0.1, 1, and 10.

Chapter 6

LINEAR MIXED MODELS (LMMs)

6.1 A GENERAL MODEL

a. Introduction

Chapter 4 deals with linear models (LMs), $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$, where elements of $\boldsymbol{\beta}$ are fixed effects, i.e., unknown constants. An example is $\mathrm{E}[y_{ij}] = \mu + \alpha_i$ where μ is a general mean and (in Section 1.3a) α_1 and α_2 represent effects on the response variable of a patient receiving the placebo or the drug progabide, respectively. Each of μ, α_1 and α_2 is a fixed effect, and in $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ the $\boldsymbol{\beta}$ is $[\mu \quad \alpha_1 \quad \alpha_2]'$.

In contrast, in Section 1.5a we discuss the model

$$\mathrm{E}[y_{ij}] = \mu + a_i + \beta_j + c_{ij} \tag{6.1}$$

where a_i is a random effect representing clinic i, β_j is a fixed effect for dose j of a drug, and c_{ij} is a random effect for interaction. This, with its mixture of fixed and random effects, is a *linear mixed model* (LMM). A special case of an LMM is when there are no fixed effects (except μ), whereupon it is called a *random model.*

In linear models, fixed effects are used for modeling the mean of $\mathbf{y}$ while random effects govern the variance-covariance structure of $\mathbf{y}$. In fact, a prime reason for having random effects is to simplify the otherwise difficult task of specifying the $N(N+1)/2$ distinct elements of $\mathrm{var}(\mathbf{y}_{N\times 1})$. Without using random effects we would have to deal with elements of $\mathrm{var}(\mathbf{y})$ being a variety of forms; but with random factors we can conveniently deal with variances and covariances attributable

to factors acknowledged to be affecting the data. Since the two kinds of effect (fixed and random) are different and so get treated differently when analyzing data, we need to know, for our data, how to decide for each factor whether it is to be deemed to be a fixed effects factor or a random effects factor. The making of this decision is discussed in Section 1.6. Having so decided, the procedures for an LMM are as follows.

b. Basic properties

The starting point for an LM is $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ being fixed effects; for an LMM we still use $\mathbf{X}\boldsymbol{\beta}$ for fixed effects but add to it $\mathbf{Zu}$ where $\mathbf{Z}$, like $\mathbf{X}$, is a known (model) matrix and $\mathbf{u}$ is the vector of random effects that occur in the data vector $\mathbf{y}$. Although the elements of $\mathbf{u}$ are random variables it is convenient to specify the model conditional on their unobservable but realized values. Thus we write not $\mathrm{E}[\mathbf{y}]$ as $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ but

$$\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \tag{6.2}$$

meaning that for the realized $\mathbf{u}$, (6.2) is the conditional mean. Were we to use $\mathbf{U}$ for random variables and $\mathbf{u}$ for their realized values, we would in place of $\mathrm{E}[\mathbf{y}|\mathbf{u}]$ write $\mathrm{E}[\mathbf{y}|\mathbf{U} = \mathbf{u}]$—but the clumsiness of this is distracting, so we stay with $\mathrm{E}[\mathbf{y}|\mathbf{u}]$.

In order to handle first and second moments of $\mathbf{y}$, those of $\mathbf{u}$ are needed. They get specified by

$$\mathbf{u} \sim (\mathbf{0}, \mathbf{D}), \text{ meaning that } \mathrm{E}[\mathbf{u}] = \mathbf{0} \text{ and } \mathrm{var}(\mathbf{u}) = \mathbf{D}. \tag{6.3}$$

There is no loss of generality in taking $\mathrm{E}[\mathbf{u}]$ to be $\mathbf{0}$, because if it was otherwise, $\mathrm{E}[\mathbf{u}] = \boldsymbol{\tau}$, say, then $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ could be rewritten as $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\tau} + \mathbf{Z}(\mathbf{u} - \boldsymbol{\tau})$. Defining $\mathbf{X}^* = [\mathbf{X}\ \ \mathbf{Z}]$ and $\boldsymbol{\beta}^* = [\boldsymbol{\beta}'\ \ \boldsymbol{\tau}']'$ gives $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\tau}$ as $\mathbf{X}^*\boldsymbol{\beta}^*$; and the further defining of $\mathbf{u} - \boldsymbol{\tau}$ as $\mathbf{u}^*$ gives $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{Zu}^*$ which, with $\mathrm{E}[\mathbf{u}^*] = \mathbf{0}$, has exactly the same form as $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$.

For specifying $\mathrm{var}(\mathbf{y})$ we have $\mathrm{var}(\mathbf{u}) = \mathbf{D}$ from (6.3) and now define

$$\mathrm{var}(\mathbf{y}|\mathbf{u}) = \mathbf{R}. \tag{6.4}$$

With $\mathrm{E}[\mathbf{u}] = \mathbf{0}$ applied to (6.2), this gives (see E 6.1)

$$\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta},\ \mathbf{ZDZ}' + \mathbf{R}), \tag{6.5}$$

showing that the fixed effects enter only the mean whereas the random effects model matrix and variance enter only the variance of $\mathbf{y}$.

6.2 ATTRIBUTING STRUCTURE TO VAR(y)

The expression for var(**y**) in (6.5) is

$$\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{ZDZ}' + \mathbf{R}. \tag{6.6}$$

Some simplifications are now described, using the following example for illustration.

a. Example

Suppose data are scores on a mathematics exam given to four ninth-grade classes from each of fifteen high schools in New York City. Aside from differences between boys and girls (which would be modelled by fixed effects) there will undoubtedly be three sources of variability: (i) among schools, (ii) among classes within each school, and (iii) among pupils within each class. Let the exam score of pupil k (of gender t) in class j of school i be y_{tijk}. Then a model equation could be

$$\text{E}[y_{tijk}|s_i, c_{ij}] = \beta_t + s_i + c_{ij}. \tag{6.7}$$

β_t is the fixed effect for gender t. The school effects, s_i for $i = 1, 2, \ldots, 15$, and the class effects c_{ij} for $j = 1, \ldots, 4$ for each school i, would be treated as random effects. So would p_{tijk}, representing everything not accounted for by β_i, s_i and c_{ij} for the individual pupil. Thus $\boldsymbol{\beta}$ of $\mathbf{X}\boldsymbol{\beta}$ in (6.2) will have two elements, β_m and β_f for male and female, respectively; and **u** of **Zu** will have the 15 s_i-effects and the 60 ($= 4 \times 15$) c_{ij}-effects.

b. Taking covariances between factors as zero

For the example it is convenient to partition **u** into two sub-vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, with $\mathbf{u}_1$ having all 15 s_i-effects as elements, and $\mathbf{u}_2$ having all 60 c_{ij}-effects. Thus

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \begin{bmatrix} \left\{ {}_c\, s_i \right\}_{i=1}^{15} \\ \left\{ {}_c \left\{ {}_c\, c_{ij} \right\}_{j=1}^{4} \right\}_{i=1}^{15} \end{bmatrix}.$$

Partition **Z** correspondingly as

$$\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2] \qquad \text{so that} \qquad \mathbf{Zu} = \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2.$$

Also partition $\mathbf{D}$ as

$$\mathbf{D} = \text{var}(\mathbf{u}) = \begin{bmatrix} \text{var}(\mathbf{u}_1) & \text{cov}(\mathbf{u}_1, \mathbf{u}_2') \\ \text{cov}(\mathbf{u}_2, \mathbf{u}_1') & \text{var}(\mathbf{u}_2) \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_2 \end{bmatrix},$$

with $\mathbf{D}_{21} = \mathbf{D}_{12}'$. Then (6.2) becomes

$$\text{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 \tag{6.8}$$

and $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$ gets to be

$$\mathbf{V} = \mathbf{Z}_1\mathbf{D}_1\mathbf{Z}_1' + \mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2' + \mathbf{Z}_1\mathbf{D}_{12}\mathbf{Z}_2' + \mathbf{Z}_2\mathbf{D}_{21}\mathbf{Z}_1' + \mathbf{R}. \tag{6.9}$$

$\mathbf{R}$ is defined in (6.4) as $\text{var}(\mathbf{y}|\mathbf{u})$. For the example this is the variance-covariance matrix of the p_{tijk}-terms described below (6.7). These represent not only the variability among pupils but also any variability not attributable to s_i and c_{ij}.

The preceding notations extend very directly from the two random factors of the example to having r random factors, so that with

$$\mathbf{u} = \left\{ {}_c\,\mathbf{u}_i \right\}_{i=1}^{r} \quad \text{and} \quad \mathbf{D} = \text{var}(\mathbf{u}) = \left\{ {}_m\,\mathbf{D}_{ii'} \right\}_{i,i'=1}^{r}$$

and

$$\mathbf{Z} = [\mathbf{Z}_1 \;\; \mathbf{Z}_2 \;\; \cdots \;\; \mathbf{Z}_r] = \left\{ {}_r\,\mathbf{Z}_i \right\}_{i=1}^{r},$$

(6.8) and (6.9) become

$$\text{E}[\mathbf{y}|\mathbf{u}] \;=\; \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{u}_i$$

and

$$\mathbf{V} \;=\; \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{D}_{ii}\mathbf{Z}_i' + \sum_{\substack{i=1 \\ i\neq i'}}^{r}\sum_{i'=1}^{r} \mathbf{Z}_i\mathbf{D}_{ii'}\mathbf{Z}_{i'}' + \mathbf{R} \tag{6.10}$$

where

$$\mathbf{D}_{ii} \;=\; \text{var}(\mathbf{u}_i) \qquad \text{and} \qquad \mathbf{D}_{ii'} = \text{cov}(\mathbf{u}_i, \mathbf{u}_{i'}').$$

Thus, in the example, $\mathbf{D}_{11}$ is the variance-covariance matrix of schools and $\mathbf{D}_{12}$ is the matrix of covariances between schools and classes. In point of fact, it is reasonable to take those covariances as zero. Some of

them are covariances between a school effect s_i and the class effects $c_{i'j}$ of classes in a different school; and there would seem to be no reason for thinking those covariances (correlations) are anything but zero. And other covariances in $\mathbf{D}_{12}$ are covariances between a school effect s_i and the class effects c_{ij} of classes within that school; and since we use random effects with the thought that they capture all the variability in the data, we assume those covariances are zero too; i.e., $\mathbf{D}_{12} = \mathbf{0}$. This extends very directly to the general case of (6.10), so that for $i \neq i'$ we take $\mathbf{D}_{ii'} = \mathbf{0}$. Hence, on writing $\mathbf{D}_i$ for $\mathbf{D}_{ii}$,

$$\mathbf{D} = \left\{ {}_d\, \mathbf{D}_i \right\}_{i=1}^{r} \qquad \text{and} \qquad \mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{D}_i \mathbf{Z}_i' + \mathbf{R}.$$

c. The traditional variance components model

– i. *Customary notation*

If in the example we assume that there is no covariance between schools and that schools exhibit homogeneity of variance, then for the 15 schools (6.9) has

$$\mathbf{D}_1 = \sigma_s^2 \mathbf{I}_{15}. \tag{6.11}$$

Similar assumptions for the four classes within school i would give

$$\text{var}[c_{i1} \quad c_{i2} \quad c_{i3} \quad c_{i4}]' = \sigma_i^2 \mathbf{I}_4.$$

And making the very reasonable assumption that the classes in one school are independent of those in every other school gives

$$\mathbf{D}_2 = \left\{ {}_d\, \text{var}(\mathbf{u}_i) \right\}_{i=1}^{15} = \left\{ {}_d\, \sigma_i^2 \mathbf{I}_4 \right\}_{i=1}^{15}.$$

An even simpler assumption is that the four classes within a school have the same variance for all 15 schools, i.e., that $\sigma_i^2 = \sigma_c^2 \ \ \forall\ i$, so giving

$$\mathbf{D}_2 = \sigma_c^2 \mathbf{I}_{60} \tag{6.12}$$

a form that is similar to (6.11). Thus (6.11) and (6.12), and a similar form for $\mathbf{R}$, namely

$$\mathbf{R} = \sigma^2 \mathbf{I}_{1200},$$

makes up the standard structure for the traditional variance components model.

The general case of r random effects factors then has

$$\mathbf{D} = \left\{ {}_d\, \sigma_i^2 \mathbf{I}_{q_i} \right\}_{i=1}^{r}$$

for random factor i having q_i effects in the data (i.e., $\mathbf{u}_i$ of order $q_i \times 1$) and

$$\mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2 + \sigma^2 \mathbf{I}_N \tag{6.13}$$

for $\mathbf{y}$ of order $N \times 1$.

– ii. *Amended notation*

An amendment to the preceding notation suggested by Hartley and Rao (1967) amounts to redefining $\mathbf{D}$ so as to include $\sigma^2 \mathbf{I}_N$. This is achieved by defining

$$\sigma_0^2 \equiv \sigma^2, \quad \mathbf{Z}_0 \equiv \mathbf{I}_N \quad \text{and} \quad q_0 \equiv N.$$

Then we define

$$\mathbf{D}_* = \begin{bmatrix} \sigma_0^2 \mathbf{I}_{q_0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} = \left\{ {}_d\, \sigma_i^2 \mathbf{I}_{q_i} \right\}_{i=0}^{r} \tag{6.14}$$

and from (6.13)

$$\mathbf{V} = \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2, \tag{6.15}$$

which can be written as

$$\mathbf{V} = \mathbf{Z}_* \mathbf{D}_* \mathbf{Z}_*' \quad \text{for} \quad \mathbf{Z}_* = [\mathbf{Z}_0 \ \ \mathbf{Z}_1 \ \ \mathbf{Z}_2 \cdots \mathbf{Z}_r] = [\mathbf{Z}_0 \ \ \mathbf{Z}]. \tag{6.16}$$

Corresponding to $\mathbf{Z}_0$ will be $\mathbf{u}_0$ of order $N \times 1$, familiarly thought of in the context of analysis of variance models as the residual error term.

The variances σ_i^2 for $i = 0, 1, \ldots, r$ are called variance components because they are the components of the variance of an individual observation; i.e., for the example

$$\text{var}(y_{tijk}) = \sigma_s^2 + \sigma_c^2 + \sigma^2.$$

d. An LMM for longitudinal data

Longitudinal data are successive observations on each of a collection of observational units (often people). An example is blood pressure measurements taken weekly on a group of patients. If $\mathbf{y}_i$ is the vector of n_i measurements on patient i, a model equation suggested by Laird and Ware (1982) is

$$\mathrm{E}[\mathbf{y}_i|\mathbf{u}_i] = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i$$

with vectors $\boldsymbol{\beta}$ and $\mathbf{u}_i$ consisting of fixed and random effects, respectively. $\boldsymbol{\beta}$ is the same for all patients and $\mathbf{u}_i$ is specific to patient i.

Suppose that there are m such patients. Then for

$$\mathbf{y} = \left\{{}_c\, \mathbf{y}_i\right\}, \quad \mathbf{X} = \left\{{}_c\, \mathbf{X}_i\right\}, \quad \mathbf{Z} = \left\{{}_d\, \mathbf{Z}_i\right\} \quad \text{and} \quad \mathbf{u} = \left\{{}_c\, \mathbf{u}_i\right\}.$$

we have

$$\left\{{}_c\, \mathrm{E}[\mathbf{y}_i|\mathbf{u}_i]\right\}_{i=1}^{m} = \mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}.$$

And the variance structure suggested by Laird and Ware (1982) is $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$ with

$$\mathbf{D}_{ii} = \mathbf{D}\ \forall\, i, \qquad \mathbf{D}_{ii'} = \mathbf{0}\ \forall\, i \neq i' \qquad \text{and} \qquad \mathbf{R} = \left\{{}_d\, \mathbf{R}_i\right\}$$

so that

$$\mathbf{V} = \mathrm{var}(\mathbf{y}) = \left\{{}_d\, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i\right\}.$$

More details for this model are described in Chapter 7.

6.3 ESTIMATING FIXED EFFECTS FOR V KNOWN

We take $\mathbf{y}$ to be normally distributed,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),$$

so that the log likelihood is

$$l = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{2}\log|\mathbf{V}| - \frac{N}{2}\log 2\pi. \tag{6.17}$$

Then from the chapter appendix (Section 6.12a–ii) we use

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \tag{6.18}$$

with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$. Making those substitutions in (6.18) and equating the result to $\mathbf{0}$ with $\boldsymbol{\beta}$ written as $\boldsymbol{\beta}^0$ gives

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad \text{so that} \qquad \boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{6.19}$$

Because $\boldsymbol{\beta}^0$ varies with the choice of $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$, we confine attention to $\mathbf{X}\boldsymbol{\beta}^0$ which is invariant because $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}$ (see Section M.4c of Appendix M) is. Thus

$$\mathrm{ML}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{6.20}$$

is the ML estimator of $\mathbf{X}\boldsymbol{\beta}$; and so $\boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta}^0$ is the ML estimator of $\boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta}$ for any $\boldsymbol{\lambda}$.

With $\mathrm{var}(\mathbf{y}) = \mathbf{V}$ it is easily seen that

$$\mathrm{var}(\mathbf{X}\boldsymbol{\beta}^0) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-\prime}\mathbf{X}'.$$

Then, because $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-\prime}$ is a generalized inverse of $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$, and also because of the invariance property referred to prior to (6.20), $\mathrm{var}(\mathbf{X}\boldsymbol{\beta}^0)$ reduces to

$$\mathrm{var}(\mathbf{X}\boldsymbol{\beta}^0) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'. \tag{6.21}$$

To test the null hypothesis $H_0: \mathbf{S}'\mathbf{X}\boldsymbol{\beta} = \mathbf{m}$, where $\mathbf{S}'$ is of full row rank ($r_{\mathbf{S}} \leq r_{\mathbf{X}}$), we can derive a chi-square statistic using

$$X^2 = (\mathbf{S}'\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{m})'[\mathbf{S}'\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{S}]^{-1}(\mathbf{S}'\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{m}). \tag{6.22}$$

Under H_0, X^2 has a central χ^2 distribution with $r_{\mathbf{S}} = \mathrm{rank}(\mathbf{S})$ degrees of freedom.

More typically $\mathbf{V}$ is known only up to a scalar multiple. To emphasize the connections with Chapter 5 and for simplicity of notation we therefore write $\mathbf{V}$ in terms of a weight matrix $\mathbf{W}$, which is the inverse of $\mathbf{V}$ up to a scalar multiple, i.e., $\mathbf{V} = \sigma^2\mathbf{W}^{-1}$, where $\mathbf{W}$ is assumed known. In such a case the following statistic can be derived as the likelihood ratio test and is also the uniformly most powerful invariant test (Lehmann, 1986):

$$F = \frac{(\mathbf{S}'\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{m})'[\mathbf{S}'\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^-\mathbf{X}'\mathbf{S}]^{-1}(\mathbf{S}'\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{m})}{r_{\mathbf{S}}\hat{\sigma}^2}, \tag{6.23}$$

where

$$\hat{\sigma}^2 = \frac{\mathbf{y}'[\mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^-\mathbf{X}'\mathbf{W}]\mathbf{y}}{N - r_{\mathbf{X}}}.$$

Under the null hypothesis, F has an $\mathcal{F}$-distribution on $r_{\mathbf{S}}$ and $N - r_{\mathbf{X}}$ degrees of freedom. The null hypothesis is rejected at significance level α when F exceeds $\mathcal{F}^{r_{\mathbf{S}}}_{N-r_{\mathbf{X}},1-\alpha}$.

6.4 ESTIMATING FIXED EFFECTS FOR V UNKNOWN

a. Estimation

With $\mathbf{V}$ unknown but not being a function of $\boldsymbol{\beta}$, the log likelihood function l of (6.17) has to be maximized with respect to elements of both $\boldsymbol{\mu}$ and $\mathbf{V}$. For $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\theta} = \boldsymbol{\beta}$, setting $\partial l/\partial\boldsymbol{\theta}$ to $\mathbf{0}$ will lead to the same result for $\boldsymbol{\beta}^0$ as in (6.19), only with $\mathbf{V}$ therein being replaced by the solution $\hat{\mathbf{V}}$ coming from maximizing l with respect to the parameters in $\mathbf{V}$. No matter what $\hat{\mathbf{V}}$ is, the ML estimator of $\mathbf{X}\boldsymbol{\beta}$ will be

$$\mathrm{ML}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \tag{6.24}$$

where we here introduce the symbol $\hat{\boldsymbol{\beta}}$ to represent $\boldsymbol{\beta}^0$ of (6.19) but with $\mathbf{V}$ replaced by $\hat{\mathbf{V}}$, which is $\mathbf{V}$ with its parameters replaced by their ML estimators. The ML equations for $\mathbf{V}$ are obtained from equating to $\mathbf{0}$ the expression

$$\frac{\partial l}{\partial \varphi_k} = -\tfrac{1}{2}\left[\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \varphi_k}\right) - (\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \varphi_k}\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] \tag{6.25}$$

of (6.70) using $\boldsymbol{\varphi}$ for each parameter in $\mathbf{V}$; and in doing this $\boldsymbol{\mu}$ is replaced by $\mathbf{X}\hat{\boldsymbol{\beta}}$ of (6.24). On writing

$$\left.\frac{\partial \mathbf{V}}{\partial \varphi_k}\right|_{\mathbf{V}=\hat{\mathbf{V}}} \quad \text{as} \quad \hat{\mathbf{V}}_{\varphi_k}$$

this gives

$$\mathrm{tr}(\hat{\mathbf{V}}^{-1}\hat{\mathbf{V}}_{\varphi_k}) = \mathbf{y}'\hat{\mathbf{P}}\hat{\mathbf{V}}_{\varphi_k}\hat{\mathbf{P}}\mathbf{y} \tag{6.26}$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \tag{6.27}$$

and $\hat{\mathbf{P}}$ is $\mathbf{P}$ with $\mathbf{V}$ replaced by $\hat{\mathbf{V}}$. For the case of the parameters in $\mathbf{V}$ being variance components, as in (6.15), we describe ML estimation in Section 6.8.

b. Sampling variance

Instead of dealing with the variance (matrix) of a vector $\mathbf{X}\hat{\boldsymbol{\beta}}$ we consider the simpler case of the scalar $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}$ for estimable $\boldsymbol{\ell}'\boldsymbol{\beta}$ (i.e., $\boldsymbol{\ell}' = \mathbf{t}'\mathbf{X}$ for some $\mathbf{t}'$).

For known $\mathbf{V}$ we have from (6.21) that $\mathrm{var}(\boldsymbol{\ell}'\boldsymbol{\beta}^0) = \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}$. A replacement for this when $\mathbf{V}$ is not known is to use $\boldsymbol{\ell}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}$,

which is an estimate of $\text{var}(\boldsymbol{\ell}'\boldsymbol{\beta}^0) = \text{var}[\boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}]$. But this is *not* an estimate of $\text{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}) = \text{var}[\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^-\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}]$. The latter requires taking account of the variability in $\hat{\mathbf{V}}$ as well as that in $\mathbf{y}$. To deal with this, Kackar and Harville (1984, p. 854) observe that (in our notation) $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}} - \boldsymbol{\ell}'\boldsymbol{\beta}$ can be expressed as the sum of two independent parts, $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}} - \boldsymbol{\ell}'\boldsymbol{\beta}^0$ and $\boldsymbol{\ell}'\boldsymbol{\beta}^0 - \boldsymbol{\ell}'\boldsymbol{\beta}$. This leads to $\text{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}})$ being expressed as a sum of two variances which we write as

$$\begin{aligned} \text{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}) &= \text{var}(\boldsymbol{\ell}'\boldsymbol{\beta}^0 - \boldsymbol{\ell}'\boldsymbol{\beta}) + \text{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}} - \boldsymbol{\ell}'\boldsymbol{\beta}^0) \\ &\doteq \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\ell} + \boldsymbol{\ell}'\mathbf{T}\boldsymbol{\ell} \end{aligned} \tag{6.28}$$

where, in the words of Kenward and Roger (1997, p. 985), "the component $\mathbf{T}$ [in our notation] represents the amount by which the asymptotic variance-covariance matrix underestimates (in a matrix sense) $\text{var}(\hat{\boldsymbol{\beta}})$." The matrix $\mathbf{T}$ in (6.28) is defined in adapting Kenward and Roger's equation (1) for the variance components model of Section 6.2c to write

$$\begin{aligned} \boldsymbol{\ell}'\mathbf{T}\boldsymbol{\ell} &= \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \\ &\quad \times \left\{ \sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}\left[\mathbf{G}_{ij} - \mathbf{F}_i(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{F}_j\right] \right\} (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\boldsymbol{\ell}, \end{aligned} \tag{6.29}$$

where c_{ij} is an element of the asymptotic variance-covariance matrix of the vector of estimated variance components; i.e.,

$$\mathbf{C} = \left\{ {}_m\, c_{ij} \right\}_{i=0,j=0}^{r\quad r} = \left\{ {}_m\, \text{cov}_\infty(\hat{\sigma}_i^2, \hat{\sigma}_j^2) \right\}_{i=0,j=0}^{r\quad r},$$

which is (6.64) of Section 6.8c. Also in (6.29)

$$\mathbf{G}_{ij} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{X} \text{ and } \mathbf{F}_i = -\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{X}. \tag{6.30}$$

We now use (6.30) in (6.29) together with both

$$\mathbf{h}' = \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1} \tag{6.31}$$

and the general result for matrices $\mathbf{S}_{ij}$ and vectors $\mathbf{t}$ that

$$\mathbf{t}'\sum_i\sum_j c_{ij}\mathbf{S}_{ij}\mathbf{t} = \sum_i\sum_j c_{ij}\mathbf{t}'\mathbf{S}_{ij}\mathbf{t} = \text{tr}\left[\mathbf{C}\left\{ {}_m\, \mathbf{t}'\mathbf{S}_{ij}\mathbf{t} \right\}_{i,j=0}^{r}\right]. \tag{6.32}$$

This gives

$$\boldsymbol{\ell}'\mathbf{T}\boldsymbol{\ell} = \sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}\mathbf{h}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{h} = \operatorname{tr}\left[\mathbf{C}\left\{{}_m\mathbf{h}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{h}\right\}_{i,j=0}^{r}\right]. \tag{6.33}$$

Thus (6.28) becomes

$$\operatorname{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}) \doteq \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\ell} + \operatorname{tr}\left[\mathbf{C}\left\{{}_m\mathbf{h}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{h}\right\}_{i,j=0}^{r}\right]. \tag{6.34}$$

To use (6.34) it does, of course, have to be calculated with $\hat{\mathbf{V}}$ in place of $\mathbf{V}$, meaning also $\hat{\mathbf{P}}$ in place of $\mathbf{P}$—and these replacements also have to be made in (6.31).

c. Bias in the variance

Kenward and Roger (1997) additionally point out that $\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}$ is a biased estimate of $\boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}$, and they investigate that bias for nonsingular $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ with unstructured $\mathbf{V}$. We adapt their methods for the variance components model having $\mathbf{V} = \sum_{i=0}^{r}\mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2$. For investigating the bias a starting point is a two-term Taylor series expansion:

$$\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell} \doteq \boldsymbol{\ell}'\left[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} + (\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2)'\frac{\partial(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}}{\partial\boldsymbol{\sigma}^2}\right.$$

$$\left. + \tfrac{1}{2}\sum_{i=0}^{r}\sum_{j=0}^{r}(\hat{\sigma}_i^2 - \sigma_i^2)(\hat{\sigma}_j^2 - \sigma_j^2)\frac{\partial^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}}{\partial\sigma_i^2\partial\sigma_j^2}\right]\boldsymbol{\ell}. \tag{6.35}$$

This has expected value

$$\begin{aligned}\mathrm{E}[\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}] &\doteq \boldsymbol{\ell}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\boldsymbol{\ell} \\ &+ \tfrac{1}{2}\sum_{i=0}^{r}\sum_{j=0}^{r}\operatorname{cov}(\hat{\sigma}_i^2,\ \hat{\sigma}_j^2)\boldsymbol{\ell}'\frac{\partial^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}}{\partial\sigma_i^2\partial\sigma_j^2}\boldsymbol{\ell}.\end{aligned}$$

On using for the derivative (6.76) from Section 6.12c,

$$\begin{aligned}\mathrm{E}[\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}] &- \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\boldsymbol{\ell} \\ &\doteq -\tfrac{1}{2}\sum_{i=0}^{r}\sum_{j=0}^{r}c_{ij}\boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\end{aligned}$$

$$\times (\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j' + \mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i')\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}$$

$$= -\sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}\mathbf{h}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{h}, \quad \text{on using (6.31)}$$

$$= -\boldsymbol{\ell}'\mathbf{T}\boldsymbol{\ell}, \quad \text{from (6.33).} \tag{6.36}$$

But in (6.28) we want to estimate

$$\text{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}) \doteq \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\boldsymbol{\ell} + \boldsymbol{\ell}'\mathbf{T}\boldsymbol{\ell} \tag{6.37}$$

and from (6.36) we have

$$\text{E}[\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}] \doteq \boldsymbol{\ell}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\boldsymbol{\ell} - \boldsymbol{\ell}'\mathbf{T}\boldsymbol{\ell}. \tag{6.38}$$

Therefore an approximately unbiased estimator for (6.37) based on adjusting $\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell}$ is

$$[\boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\boldsymbol{\ell}]_{adjusted} = \boldsymbol{\ell}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\boldsymbol{\ell} + 2\boldsymbol{\ell}'\mathbf{T}\boldsymbol{\ell} \tag{6.39}$$

with everything calculated using $\hat{\mathbf{V}}$ and $\hat{\mathbf{P}}$ in place of $\mathbf{V}$ and $\mathbf{P}$.

d. Approximate F-statistics

The F-statistic in (6.23) has an $\mathcal{F}$-distribution; it is for the case of $\mathbf{V} = \sigma^2\mathbf{W}^{-1}$ with $\mathbf{W}$ known. But when $\mathbf{V}$ is not of this simple form, it has to be replaced in $\mathbf{S}'\mathbf{X}\hat{\boldsymbol{\beta}}$ and in F by an estimate, $\hat{\mathbf{V}}$, and the resulting value of F, call it $\hat{F}$, has an unknown distribution. If we assume that $\hat{F}$ is distributed approximately as $\mathcal{F}$, one way of making this approximation is to assume that

$$\lambda\hat{F} \sim \mathcal{F}_d^r \tag{6.40}$$

where

$$r = r_{\mathbf{S}} \quad \text{for} \quad \mathbf{S} \text{ of } H_0 : \mathbf{S}'\mathbf{X}\boldsymbol{\beta} = \mathbf{m}.$$

Then, similar to Satterthwaite (1946), λ and d are derived by equating first and second moments of both sides of (6.40). This gives

$$\lambda\text{E}[\hat{F}] = d/(d-2)$$

and

$$\lambda^2 \text{var}(\hat{F}) = 2d^2(r+d-2)/r(d-2)^2(d-4).$$

These lead to

$$\lambda = \frac{d}{(d-2)\text{E}[\hat{F}]}$$

and

$$d = \frac{1+2/r}{\text{var}(\hat{F})/(2\{\text{E}[\hat{F}]\}^2) - 1/r}.$$

Kenward and Roger (1997) derive these results in their equations (7) and (8) and give some extensions. The calculation of $\text{E}[\hat{F}]$ and $\text{var}(\hat{F})$ is tedious.

6.5 PREDICTING RANDOM EFFECTS FOR V KNOWN

The assumptions about random effects differ from those for fixed effects and so treatment of the two kinds of effects is not the same. A fixed effect is considered to be a constant, which we wish to estimate. But a random effect is considered as just an effect coming from a population of effects. It is this population that is an extra assumption, compared to fixed effects, and we would hope it would lead to an estimation method for random effects being an improvement over that for fixed effects. To emphasize this distinction we use the term *prediction* of random effects rather than estimation.

For instance, in Example 4 of Section 1.4a, we treat clinic i as being from a population of clinics, with $\text{E}[y_{ij}|a_i] = \mu + a_i$ being the ith clinic's true response. We may wish to predict the value of a_i to gain information about the performance of that particular clinic. Alternatively, we may want to use the predicted values of the a_i from several clinics in order to rank the clinics, or to select the best ones. Since they are all assumed to be selected from the same distribution it makes sense that their predicted values will have some degree of similarity and be less variable than might be anticipated without such an assumption.

Using the assumption that the realized random effects which determine the data are just a random selection from a conceptual population of such effects, it is not difficult to show that the "best" prediction of a_i (best in the sense of minimized mean squared error of prediction—see Chapter 9) is the conditional mean $\text{E}[a_i|\mathbf{y}]$. In using this as the predictor of a_i we are using the expected value of the random effect in light

of the data. An example of this is in the dairy farming industry where bulls are selected for use in artificial breeding on the basis of their daughters' average milk yield. Suppose the kth bull has a daughter with average milk yield $\bar{y}_k$. It is perfectly reasonable to think that in the population of bulls there will be bulls other than the kth that nevertheless have (or could have) the same daughter average, namely $\bar{y}_k$. Despite this, these bulls will not necessarily all have the same genetic values, let alone all the same as that of bull k. Therefore, since $\bar{y}_k$ is our data, and if a is the random effect representing bull genetic values, the best we can do for estimating bull k's genetic value is the conditional mean $\mathrm{E}[a|\bar{y}_k]$. Not surprisingly, since predictors calculated as $\mathrm{E}[a_i|\mathbf{y}]$ are "best", they have smaller mean squared error than would estimates based on assuming the random effects were fixed effects. They also have less variability and are sometimes called *shrinkage estimators*. This is because, just as in Section 1.4b–iv,

$$\begin{aligned} \mathrm{var}(a) &= \mathrm{var}(\mathrm{E}[a|\mathbf{y}]) + \mathrm{E}[\mathrm{var}(a|\mathbf{y})] \\ &= \mathrm{var}(\tilde{a}) + \text{ a positive value} \end{aligned}$$

where $\tilde{a} = \mathrm{E}[a|\mathbf{y}]$ is the predictor. Thus

$$\mathrm{var}(\tilde{a}) \leq \mathrm{var}(a)$$

and so $\tilde{a}$ is said to be a shrinkage estimator.

From the preceding discussion we now turn to the general case of

$$\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \text{ for } \mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R},$$

for which the conditional expected value $\mathrm{E}[\mathbf{u}|\mathbf{y}]$ is, assuming $\mathbf{y}$ and $\mathbf{u}$ follow a jointly normal distribution

$$\mathrm{E}[\mathbf{u}|\mathbf{y}] = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{6.41}$$

Replacing $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ of (6.19) gives what is called the best linear unbiased predictor (BLUP). Derivation of (6.41) and other results concerning prediction are detailed in Chapter 9.

We write

$$\tilde{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{6.42}$$

and

$$\tilde{\mathbf{u}}^0 = \mathbf{DZ'V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0) = \mathbf{DZ'Py}. \tag{6.43}$$

Then

$$\text{var}(\tilde{\mathbf{u}}) = \mathbf{DZ'V}^{-1}\mathbf{ZD} \quad \text{and} \quad \text{var}(\tilde{\mathbf{u}}^0) = \mathbf{DZ'PZD}, \tag{6.44}$$

the latter result using $\mathbf{PVP} = \mathbf{P}$. Thus with $\mathbf{D}$ and $\mathbf{V}$ known, (6.44) provides opportunities for testing hypotheses or deriving confidence intervals for elements of $\mathbf{u}$. And use can also be made of

$$\text{cov}(\boldsymbol{\beta}^0, \tilde{\mathbf{u}}^{0\prime}) = 0, \qquad \text{cov}(\tilde{\mathbf{u}}^0, \tilde{\mathbf{u}}') = \text{var}(\tilde{\mathbf{u}}^0)$$

and

$$\text{var}(\tilde{\mathbf{u}}^0 - \mathbf{u}) = \mathbf{D} - \mathbf{DZ'PZD}. \tag{6.45}$$

6.6 PREDICTING RANDOM EFFECTS FOR V UNKNOWN

a. Estimation

When $\mathbf{D}$ and $\mathbf{V}$ are unknown, they are typically replaced by $\hat{\mathbf{D}}$ and $\hat{\mathbf{V}}$ in $\tilde{\mathbf{u}}$ of (6.42), giving what could be called the estimated best predictor, to be denoted $\hat{\mathbf{u}}$:

$$\hat{\mathbf{u}} = \hat{\mathbf{D}}\mathbf{Z'}\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{D}}\mathbf{Z'}\hat{\mathbf{P}}\mathbf{y}.$$

b. Sampling variance

Kackar and Harville (1984) give extensive discussion of deriving the mean squared error of $\boldsymbol{\ell}'\hat{\boldsymbol{\beta}}^0 + \mathbf{m}'\hat{\mathbf{u}}$; Prasad and Rao (1990) suggest an alternative approximation and apply it to their three special cases of small-area estimators. And although the Kenward and Roger (1997) form of $\text{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}})$ given by (6.28) and (6.29) is very different looking from Kackar and Harville (1984), they both ultimately reduce to the same thing. Using similar methods of reduction for the variance components model, $\text{var}(\mathbf{m}'\hat{\mathbf{u}} - \mathbf{m}'\mathbf{u})$ comes from Kackar and Harville (1984) as

$$\text{var}(\mathbf{m}'\hat{\mathbf{u}} - \mathbf{m}'\mathbf{u}) = \mathbf{m}'\mathbf{Dm} - \mathbf{m}'\mathbf{DZ'PZDm} \tag{6.46}$$

$$+ \sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}[(\mathbf{m}_i' - \mathbf{m}'\mathbf{DZ'PZ}_i)\mathbf{Z}_i'\mathbf{PZ}_j(\mathbf{m}_j - \mathbf{Z}_j'\mathbf{PZDm})],$$

where $\mathbf{m}' = [\mathbf{m}_0' \ \mathbf{m}_1' \cdots \mathbf{m}_i' \cdots \mathbf{m}_r']$. We write

$$\Delta_{ij} = (\mathbf{m}_i' - \mathbf{m}'\mathbf{DZ'PZ}_i)\mathbf{Z}_i'\mathbf{PZ}_j(\mathbf{m}_j - \mathbf{Z}_j'\mathbf{PZDm})$$

so that

$$\text{var}(\mathbf{m}'\hat{\mathbf{u}} - \mathbf{m}'\mathbf{u}) \doteq \mathbf{m}'\mathbf{D}\mathbf{m} - \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} + \sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}\Delta_{ij}. \quad (6.47)$$

c. Bias in the variance

The procedure used for obtaining the bias of $\widehat{\text{var}}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}})$ in Section 6.4c can be applied similarly to $\widehat{\text{var}}(\mathbf{m}'\hat{\mathbf{u}}^0 - \mathbf{m}'\mathbf{u})$ starting from an expression like that of (6.35), namely

$$\begin{aligned} \text{E}[\mathbf{m}'(\hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{P}}\mathbf{Z}\hat{\mathbf{D}})\mathbf{m}] &\doteq \mathbf{m}'(\mathbf{D} - \mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D})\mathbf{m} \quad (6.48) \\ &\quad - \sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}\mathbf{m}'\frac{\partial^2(\mathbf{D} - \mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D})}{\partial\sigma_i^2\partial\sigma_j^2}\mathbf{m}. \end{aligned}$$

The result, after using (6.79) in Section 6.12, is

$$\text{E}[\mathbf{m}'(\hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{P}}\mathbf{Z}\hat{\mathbf{D}})\mathbf{m}] \doteq \mathbf{m}'(\mathbf{D} - \mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D})\mathbf{m} - \sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}\Delta_{ij}. \quad (6.49)$$

Thus, by comparison with (6.47),

$$\mathbf{m}'(\hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{P}}\mathbf{Z}\hat{\mathbf{D}})\mathbf{m} + 2\sum_{i=0}^{r}\sum_{j=0}^{r} c_{ij}\Delta_{ij} \quad (6.50)$$

is an approximately unbiased estimator of $\text{var}(\mathbf{m}'\hat{\mathbf{u}} - \mathbf{m}'\mathbf{u})$. As with $\text{var}(\boldsymbol{\ell}'\hat{\boldsymbol{\beta}})$, of course, (6.50) must be calculated with $\hat{\mathbf{D}}$, $\hat{\mathbf{V}}$ and $\hat{\mathbf{P}}$ replacing $\mathbf{D}$, $\mathbf{V}$ and $\mathbf{P}$.

6.7 ANOVA ESTIMATION OF VARIANCE COMPONENTS

For random effects we want not only to predict them, as just discussed, but also to estimate their variances. Methods for doing this are detailed in many books and papers, the two main methods being ANOVA and ML. We here briefly outline the main ideas of the ANOVA methodology which, although seldom applicable to GLMMs of later chapters, is important for LMMs, both historically and for its practicality in certain circumstances.

a. Balanced data

Suppose in the 1-way classification random model, with model equation $\mathrm{E}[y_{ij}|a_i] = \mu + a_i$, as in Section 2.1, that we have $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$; i.e., n observations in every one of the m classes, the simplest example of balanced data. Then $\mathbf{V}$ of $\mathbf{V} = \mathbf{ZDZ}' + \mathbf{R}$ turns out to be

$$\mathbf{V} = \left\{ {}_d\, \sigma^2 \mathbf{I}_n + \sigma_a^2 \mathbf{J}_n \right\}_{i=1}^{m}. \tag{6.51}$$

For this, the traditional analysis of variance table contains the following two sums of squares:

$$\mathrm{SSA} = n \sum_{i=1}^{m} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \quad \text{and} \quad \mathrm{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2. \tag{6.52}$$

The expected values of these, based on $\mathbf{V}$ of (6.51), are

$$\mathrm{E}[\mathrm{SSA}] = (m-1)(n\sigma_a^2 + \sigma^2) \quad \text{and} \quad \mathrm{E}[\mathrm{SSE}] = m(n-1)\sigma^2. \tag{6.53}$$

The ANOVA method of estimating variance components is to equate expected values like (6.53) to the corresponding calculated values, (6.52): the solutions for the variance components are taken as the ANOVA estimates thereof. This gives

$$\tilde{\sigma}^2 = \frac{\mathrm{SSE}}{m(n-1)} \tag{6.54}$$

and

$$\tilde{\sigma}_a^2 = \frac{1}{n} \left(\frac{\mathrm{SSA}}{m-1} - \tilde{\sigma}^2 \right). \tag{6.55}$$

This method of estimation extends very directly to analysis of variance of balanced (equal subclass numbers) data where there are more sums of squares than just the two of the preceding example. The resulting estimators are always unbiased, although they can yield negative estimates, as is possible, for example, in (6.55). The estimators are also minimum variance quadratic unbiased. On assuming normality for $\mathbf{y}$ they are minimum variance unbiased and their sampling variances and unbiased estimators thereof are readily available. Searle et al. (1992, Chapter 4) has extensive details for the case of balanced data.

b. Unbalanced data

For unbalanced data (unequal subclass numbers) the ANOVA method can still be applied, but its utility is severely limited. This is because with many cases of unbalanced data there is more than one set of sums of squares that might be laid out as an analysis of variance (see Chapter 5). In such cases there is therefore no unique set of sums of squares as there is with balanced data. Consequently there is no unique set of equations such as (6.54) and (6.55) and no unique estimators.

An extension of this is to use not just sums of squares but quadratic forms of the data, arrayed in a vector $\mathbf{q}$, say, such that $\mathrm{E}[\mathbf{q}] = \mathbf{B}\boldsymbol{\sigma}^2$ where $\boldsymbol{\sigma}^2$ is the vector of variance components in the model under consideration. Then, if $\mathbf{B}^{-1}$ exists, $\tilde{\boldsymbol{\sigma}}^2 = \mathbf{B}^{-1}\mathbf{q}$ provides unbiased ANOVA estimators of the elements of $\boldsymbol{\sigma}^2$. Usually, $\mathbf{q}$ is the same order as $\boldsymbol{\sigma}^2$—as in equations (6.52) and (6.53). It cannot be less. But if it is more, and provided that $\mathbf{B}$ has full column rank, $\tilde{\boldsymbol{\sigma}}^2 = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{q}$ is an unbiased estimator of $\boldsymbol{\sigma}^2$. Note that in using $\mathbf{q}$ there are no rules for choosing its elements other than they be quadratic forms with expected values not involving $\boldsymbol{\beta}$. Hence there are infinite numbers of estimators $(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{q}$; and all of them are ANOVA (or method-of-moments, second moments) estimators. There are also methods of the 1970s such as LaMotte's various quadratic estimators, and Rao's minimum norm quadratic unbiased estimators, some of which utilize *a priori* values of the variance components. Searle et al. (1992, Section 11.3) discuss these methods in some detail and give extensive references.

From a theoretical statistics perspective, ANOVA estimators are not always based on sufficient statistics; and minimal, complete, sufficient statistics do not exist (see E 6.7). As a consequence, there are no uniformly optimal ANOVA estimators. One way out of this dilemma may be to invoke further criteria for choosing quadratic forms for estimating variance components. In Section 10.2b we show that restricted ML (REML) estimation suggests using quadratic forms coming from best linear unbiased predicted (BLUP) values.

A consequence of all this is that ANOVA estimation of variance components is losing some (much) of its popularity. This includes the three well-known methods of Henderson (1953), a landmark paper in its time, which definitively motivated much interest in estimating variance components from unbalanced data and which provided methodology that was pivotal and widely used for some thirty years or more. Extensive details of these three methods, of their application to the 2-way clas-

sification random model, and of ANOVA estimation from unbalanced data are given in the sixty pages of Chapter 6 of Searle et al. (1992).

6.8 MAXIMUM LIKELIHOOD (ML) ESTIMATION

a. Estimators

In place of the numerous forms of ANOVA estimation (with its deficiencies) the method now widely preferred is maximum likelihood (ML) estimation or variations thereof. In applying it to

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2\right) \tag{6.56}$$

we simultaneously seek ML estimators of $\boldsymbol{\beta}$ and $\mathbf{V}$. Section 1.7a–iii describes the need for distinguishing between solutions of ML equations and the ML estimators derived therefrom. In keeping with that, the solution $\dot{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is as in (6.24), only with the solution $\dot{\mathbf{V}}$ replacing $\mathbf{V}$, so that

$$\mathbf{X}\dot{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}\dot{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}\dot{\mathbf{V}}^{-1}\mathbf{y}. \tag{6.57}$$

And for σ^2 we use $\partial \mathbf{V}/\partial \sigma_i^2 = \partial(\sum_j \mathbf{Z}_j \mathbf{Z}_j' \sigma_i^2)/\partial \sigma_i^2 = \mathbf{Z}_i \mathbf{Z}_i'$ in $\partial l/\partial \varphi$ of (6.25). This gives

$$\frac{\partial l}{\partial \sigma_i^2} = -\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i') - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

In equating this to 0 for each $i = 0, 1, \ldots, r$ and continuing with the $\dot{\boldsymbol{\beta}}$ and $\dot{\mathbf{V}}$ notation we have

$$\mathrm{tr}(\dot{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i') = (\mathbf{y} - \mathbf{X}\dot{\boldsymbol{\beta}})'\dot{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\dot{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\dot{\boldsymbol{\beta}}). \tag{6.58}$$

Now (6.57) and (6.58) have to be solved for $\dot{\boldsymbol{\beta}}$ and $\dot{\sigma}^2$, so leading to $\dot{\mathbf{V}}$. But the right-hand side of (6.58) involves $\mathbf{X}\dot{\boldsymbol{\beta}}$ of (6.57); and that equation involves $\dot{\mathbf{V}}$. So somehow these equations must simultaneously be solved numerically (often by iteration).

In point of fact we can reduce the two equations (6.57) and (6.58) by observing from (6.57) that $\dot{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\dot{\boldsymbol{\beta}})$ needed for (6.58) is

$$\dot{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\dot{\boldsymbol{\beta}}) = \left[\dot{\mathbf{V}}^{-1} - \dot{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}\dot{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\dot{\mathbf{V}}^{-1}\right]\mathbf{y} = \dot{\mathbf{P}}\mathbf{y} \tag{6.59}$$

for $\mathbf{P}$ of (6.27) with, of course, $\dot{\mathbf{P}}$ being $\mathbf{P}$ with $\dot{\mathbf{V}}$ in place of $\mathbf{V}$. Therefore (6.58) can be written as

$$\left\{{}_c\,\mathrm{tr}(\dot{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i')\right\}_{i=0}^{r} = \left\{{}_c\,\mathbf{y}'\dot{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{y}\right\}_{i=0}^{r}. \tag{6.60}$$

So now, for obtaining a solution $\dot{\sigma}^2$, we need concentrate only on (6.60) using (6.59). And when a solution is obtained, (6.57) will yield $\mathbf{X}\dot{\boldsymbol{\beta}}$. Of course, solving (6.60) is not simple; for a few experiment designs yielding balanced data it does have straightforward algebraic solutions, as shown in Searle et al. (1992, Sec. 4.7).

At this point we can evaluate the profile likelihood for $\mathbf{V}$, denoted l_P, which is the likelihood for a given value of $\mathbf{V}$ with the maximizing value of $\boldsymbol{\beta}$ for that $\mathbf{V}$ inserted. Using (6.59) in the log likelihood, (6.17), gives the profile log likelihood of

$$\log l_P(\mathbf{V}) = -\tfrac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y} - \tfrac{1}{2}\log|\mathbf{V}| - \tfrac{N}{2}\log(2\pi). \qquad (6.61)$$

Although (6.60) and (6.61) do not appear to involve the σ^2s, they do, of course, because the σ^2s are embedded in $\mathbf{V}$ and $\mathbf{P}$. For unbalanced data and even for some balanced data (e.g., 2-way crossed classification, random model, see Searle et al., 1992, Section 4.7d), solving (6.60) or maximizing (6.61) has to be achieved by arithmetic methods. A prime difficulty is to obtain estimates within the range of the parameters (see Section 2.2b-iii). For the variance components model this means $\hat{\sigma}_i^2 \geq 0$ for $i > 0$ and $\hat{\sigma}_0^2 > 0$. On achieving this, the corresponding $\hat{\mathbf{V}}$ will be the ML estimator $\hat{\mathbf{V}}$, and ML($\mathbf{X}\boldsymbol{\beta}$) will be $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{y}$ as in (6.24). Further discussion of computing techniques for finding the ML estimates will be found in Chapter 10.

b. Information matrix

Asymptotic sampling variances of ML estimators are obtained from the inverse of the information matrix, which is minus the expected value of the matrix of second derivatives (with respect to the parameters) of the likelihood. Expressions for this, for the general model where $\boldsymbol{\mu}$ depends on a parameter $\boldsymbol{\theta}$ and $\mathbf{V}$ depends on a parameter $\boldsymbol{\varphi}$ and where we assume $\mathbf{y} \sim \mathcal{N}[\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{V}(\boldsymbol{\varphi})]$, are given in Section 6.12a–iii. For $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sum_{i=0}^{r}\mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2)$ where for (6.74) $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\theta} = \boldsymbol{\beta}, \mathbf{V} = \sum_{i=0}^{r}\mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2$ and each element of $\boldsymbol{\varphi}$ is a σ_i^2, (6.74) gives the information matrix as

$$\mathbf{I}\begin{bmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \tfrac{1}{2}\left\{ {}_m\,\mathrm{tr}[\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j(\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j)'] \right\}_{i,j=0}^{r} \end{bmatrix}. \qquad (6.62)$$

c. Asymptotic sampling variances

Inverting (6.62) but using a generalized inverse for $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ gives

$$\operatorname{var}_\infty \begin{bmatrix} \mathbf{X}\hat{\boldsymbol{\beta}} \\ \hat{\sigma}^2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}' & \mathbf{0} \\ \mathbf{0} & 2\left[\left\{\operatorname{tr}[\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j(\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j)']\right\}_{i,j=0}^{r}\right]^{-1} \end{bmatrix}$$

so that

$$\operatorname{var}_\infty(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}' \tag{6.63}$$

$$\operatorname{var}_\infty(\hat{\sigma}^2) = 2\left[\left\{\operatorname{tr}[\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j(\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j)']\right\}_{i,j=0}^{r}\right]^{-1} \tag{6.64}$$

and

$$\operatorname{cov}_\infty(\mathbf{X}\hat{\boldsymbol{\beta}},\ \hat{\sigma}^{2'}) = \mathbf{0}. \tag{6.65}$$

6.9 RESTRICTED MAXIMUM LIKELIHOOD (REML)

For estimating variance components an alternative maximum likelihood procedure, known as *restricted* (or *residual*) *maximum likelihood* (REML), maximizes the likelihood of linear combinations of elements of $\mathbf{y}$. They are chosen as $\mathbf{k}'\mathbf{y}$ (for vector $\mathbf{k}$) so that $\mathbf{k}'\mathbf{y}$ contains none of the fixed effects in $\boldsymbol{\beta}$. This means having $\mathbf{k}'$ such that $\mathbf{k}'\mathbf{X} = \mathbf{0}$. For optimality we use the maximum number, $N - r_{\mathbf{X}}$, of linearly independent vectors $\mathbf{k}'$ and write $\mathbf{K} = [\mathbf{k}_1 \ \ \mathbf{k}_2 \ \cdots \ \mathbf{k}_{N-r_{\mathbf{X}}}]$. This results in doing maximum likelihood on $\mathbf{K}'\mathbf{y}$ instead of $\mathbf{y}$, where $\mathbf{K}'\mathbf{X} = \mathbf{0}$ and $\mathbf{K}'$ has full row rank $N - r_{\mathbf{X}}$. (These results are described in Sections M.4f and g)

a. Estimation

For $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with $\mathbf{K}'\mathbf{X} = \mathbf{0}$, we have

$$\mathbf{K}'\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}'\mathbf{V}\mathbf{K}).$$

ML equations for $\mathbf{K}'\mathbf{y}$ can therefore be derived from those for $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, namely (6.58), by replacing

$$\begin{array}{lll} \mathbf{y} & \text{with} & \mathbf{K}'\mathbf{y}; \\ \mathbf{Z} & \text{with} & \mathbf{K}'\mathbf{Z} \end{array} \quad \text{and} \quad \begin{array}{lll} \mathbf{X} & \text{with} & \mathbf{K}'\mathbf{X} = \mathbf{0}; \\ \mathbf{V} & \text{with} & \mathbf{K}'\mathbf{V}\mathbf{K}. \end{array} \tag{6.66}$$

On using

$$\mathbf{P} = \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}',$$

of (6.77) the ML equations for $\mathbf{K}'\mathbf{y}$ reduce to

$$\left\{ {}_c\,\mathrm{tr}(\dot{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i') \right\}_{i=0}^{r} = \left\{ {}_c\,\mathbf{y}'\dot{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{y} \right\}_{i=0}^{r}. \tag{6.67}$$

These are the REML equations, to be solved for $\dot{\sigma}^2$ which occurs in $\dot{\mathbf{P}}$. It is easily seen that they are the same as the ML equations (6.60) except for $\dot{\mathbf{V}}$ on the left-hand side being replaced by $\dot{\mathbf{P}}$ in (6.67).

b. Sampling variances

If the replacements noted in (6.66) are made to $\mathrm{var}_\infty(\hat{\sigma}^2)$ of (6.64) the result is the variance-covariance matrix of the REML estimators:

$$\mathrm{var}_\infty(\hat{\sigma}^2_{\mathrm{REML}}) = 2\left[\left\{ {}_m\,\mathrm{tr}[\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j(\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j)'] \right\}_{i,j=0}^{r} \right]^{-1}.$$

6.10 ML OR REML?

An oft-asked question is: Should one use ML or REML? Searle et al. (1992, Sec. 6.8) definitely prefer each of ML and REML over ANOVA estimation. We firmly endorse that preference, particularly because, as has already been mentioned, ANOVA methods do not apply satisfactorily to generalized linear mixed models.

For addressing "ML versus REML" there are a number of features of the two methods that can easily be stated. Both have the merit of being based on the well-respected maximum likelihood principle. This does have the problem that if any of the maximizing solutions are negative, one has to adjust those solutions to yield estimators in the parameter space (see, for example, Sections 2.2b–ii and –iii). On the other hand, the maximum likelihood principle yields asymptotic sampling variances of the variance components estimators; but it also has the demerit of difficult computability. ML provides estimation of fixed effects, but REML itself does not. Nevertheless, overriding these features there seems to be a growing preference for REML, influenced by its following merits. First, it is sensible for balanced data for which REML *solutions* (not estimators) are the ANOVA estimators—and these, despite their ability to be negative, have the substantive merit of being minimal variance unbiased, under normality – and even minimum variance

quadratic unbiased otherwise. But there is no guarantee that properties of this nature apply to REML solutions from unbalanced data. Second, REML estimators are based on taking into account the degrees of freedom for the fixed effects in the model. Sections 1.7a–ii, 2.1b, 2.2b–vi and 3.2c show examples of this for balanced data. This is particularly important when the rank of $\mathbf{X}$ is large in relation to the sample size. And, although REML for unbalanced data yields no clean algebraic results, presumably this degree-of-freedom feature occurs with unbalanced data too. Third, because $\boldsymbol{\beta}$ is not involved in REML, the resulting estimators (of variance components) are invariant to the value of $\boldsymbol{\beta}$. Change $\boldsymbol{\beta}$ (but with $\mathbf{X}$ unchanged) and one does not alter REML estimators. Finally, REML estimators do not seem to be as sensitive to outliers in the data (see Verbyla, 1993) as are ML estimators.

6.11 OTHER METHODS FOR ESTIMATING VARIANCES

Several other methods for estimating variance components are mentioned briefly in Searle et al. (1992, Chap. 11), and even more briefly here. The methods are referred to mostly by their acronyms, which indicate their primary properties. The best known is MINQUE: minimum norm quadratic unbiased estimation, a method which is based on a pre-assigned value of σ^2. As such, the resulting estimates depend on that value; and for this reason we feel it is of little appeal. Variants of MINQUE are MINQU(0), which takes the pre-assigned value of σ^2 as having every σ_i^2 (except σ^2) as zero; and I-MINQUE, iterated MINQUE, the estimates from which are identical to REML solutions (which, of course, can be negative); and conversely, solutions from the first iteration of REML are a set of MINQUE estimates. There is also MIVQUE, minimum variance quadratic unbiased estimation – and several variants of it.

6.12 APPENDIX

a. Differentiating a log likelihood

– i. *A general likelihood under normality*

For the general model under normality,

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V}) \quad \text{with} \quad \mathrm{E}[\mathbf{y}] = \boldsymbol{\mu} \quad \text{and} \quad \mathrm{var}(\mathbf{y}) = \mathbf{V},$$

the density function is

$$f(\mathbf{y}|\mathbf{u},\mathbf{V}) = \frac{\exp[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})]}{(2\pi)^{\frac{1}{2}N}|\mathbf{V}|^{\frac{1}{2}}}.$$

Thus the log likelihood is

$$l = -\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu}) - \tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}N\log 2\pi. \qquad (6.68)$$

We consider a general parameterization of $\boldsymbol{\mu}$ and $\mathbf{V}$ such that each element of $\boldsymbol{\mu}$ is a function of elements of a parameter vector $\boldsymbol{\theta}$; and, similarly, each element of $\mathbf{V}$ is a function of elements of a parameter vector $\boldsymbol{\varphi}$ which is unrelated to $\boldsymbol{\theta}$. Thus we write

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) \qquad \text{and} \qquad \mathbf{V} = \mathbf{V}(\boldsymbol{\varphi})$$

and so have, after ignoring $N/2\log 2\pi$,

$$l = -\tfrac{1}{2}[\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta})]'[\mathbf{V}(\boldsymbol{\varphi})]^{-1}[\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta})] - \tfrac{1}{2}\log|\mathbf{V}(\boldsymbol{\varphi})|.$$

– ii. *First derivatives*

Direct differentiation of l (which, in application to vectors demands careful consideration of conformability, as seen in Section M.5) gives

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu}). \qquad (6.69)$$

And, for φ_k being an element of $\boldsymbol{\varphi}$ in $\mathbf{V}(\boldsymbol{\varphi})$

$$\frac{\partial l}{\partial \varphi_k} = -\tfrac{1}{2}\left[\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \varphi_k}\right) - (\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \varphi_k}\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]. \qquad (6.70)$$

Using $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ and $\mathbf{V} = \mathbf{V}(\boldsymbol{\varphi})$ in each of these, and equating them to zero gives the ML equations. In the case of (6.70) there will be one such equation for each φ_k of $\boldsymbol{\varphi}$.

– iii. *Information matrix*

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left(\frac{\partial l}{\partial \boldsymbol{\theta}'}\right) = \frac{\partial}{\partial \boldsymbol{\theta}}\left(\left[\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]'\right), \qquad \text{using (6.69)} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}}\left[(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}'}\right] \\
&= -\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\mathbf{V}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}'} + (\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}\frac{\partial^2 \boldsymbol{\mu}}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}
\end{aligned}$$

and so

$$-\mathrm{E}\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}}\right] = \frac{\partial\boldsymbol{\mu}'}{\partial\boldsymbol{\theta}}\mathbf{V}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}}. \tag{6.71}$$

Also

$$\begin{aligned}\frac{\partial^2 l}{\partial\varphi_k\,\partial\boldsymbol{\theta}'} &= \frac{\partial}{\partial\varphi_k}\left[(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}'}\right] = (\mathbf{y}-\boldsymbol{\mu})'\frac{\partial\mathbf{V}^{-1}}{\partial\varphi_k}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}'}\\ &= -(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_k}\mathbf{V}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}'}.\end{aligned}$$

Since $\mathrm{E}[\mathbf{y}-\boldsymbol{\mu}] = \mathbf{0}$

$$-\mathrm{E}\left[\frac{\partial^2 l}{\partial\varphi_k\,\partial\boldsymbol{\theta}'}\right] = \mathbf{0}. \tag{6.72}$$

Next, on differentiating (6.70) with respect to φ_s,

$$\begin{aligned}&\frac{\partial^2 l}{\partial\varphi_s\,\partial\varphi_k} =\\ &-\tfrac{1}{2}\left\{\mathrm{tr}\left(-\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_s}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_k}+\mathbf{V}^{-1}\frac{\partial^2\mathbf{V}}{\partial\varphi_s\,\partial\varphi_k}\right)\right.\\ &+(\mathbf{y}-\boldsymbol{\mu})'\left[(-1)\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_s}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_k}\mathbf{V}^{-1}+\mathbf{V}^{-1}\frac{\partial^2\mathbf{V}}{\partial\varphi_s\,\partial\varphi_k}\mathbf{V}^{-1}\right.\\ &\left.\left.-\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_k}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_s}\mathbf{V}^{-1}\right](\mathbf{y}-\boldsymbol{\mu})\right\}.\end{aligned}$$

Now for any $\mathbf{A}$

$$\mathrm{E}[(\mathbf{y}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{y}-\boldsymbol{\mu})] = \mathrm{tr}\left\{\mathbf{A}\mathrm{E}[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})']\right\} = \mathrm{tr}(\mathbf{AV}).$$

Therefore

$$\begin{aligned}-\mathrm{E}\left[\frac{\partial^2 l}{\partial\varphi_s\,\partial\varphi_k}\right] &= \tfrac{1}{2}\left\{\mathrm{tr}\left(-\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_s}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_k}+\mathbf{V}^{-1}\frac{\partial^2\mathbf{V}}{\partial\varphi_s\,\partial\varphi_k}\right)\right.\\ &\qquad\left.+\mathrm{tr}\left[+\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_s}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\varphi_k}\right.\right.\end{aligned}$$

$$- \mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial\varphi_s\,\partial\varphi_k} + \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial\varphi_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial\varphi_s}\Bigg]\Bigg\}$$

$$= \tfrac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial\varphi_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial\varphi_s}\right). \tag{6.73}$$

Therefore, on assembling (6.71), (6.72) and (6.73), the information matrix is

$$-\mathrm{E}\begin{bmatrix} \dfrac{\partial^2 l}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} & \dfrac{\partial^2 l}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\varphi}'} \\ \left(\dfrac{\partial^2 l}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\varphi}'}\right)' & \dfrac{\partial^2 l}{\partial\boldsymbol{\varphi}\,\partial\boldsymbol{\varphi}'} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{\partial\boldsymbol{\mu}'}{\partial\boldsymbol{\theta}}\mathbf{V}^{-1}\dfrac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}} & \mathbf{0} \\ \mathbf{0} & \tfrac{1}{2}\left\{_m\,\mathrm{tr}\left(\mathbf{V}^{-1}\dfrac{\partial \mathbf{V}}{\partial\varphi_k}\mathbf{V}^{-1}\dfrac{\partial \mathbf{V}}{\partial\varphi_s}\right)\right\} \end{bmatrix}. \tag{6.74}$$

b. Differentiating a generalized inverse

Suppose $\mathbf{A}$ is a function of the scalar x, and that we write $d\mathbf{A}/dx$ as $d\mathbf{A}$. Then for $\mathbf{A}^-$ being a generalized inverse of $\mathbf{A}$ defined by

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$$

we have

$$(d\mathbf{A})\mathbf{A}^-\mathbf{A} + \mathbf{A}(d\mathbf{A}^-)\mathbf{A} + \mathbf{A}\mathbf{A}^-(d\mathbf{A}) = d\mathbf{A}.$$

Post-multiplying by$\mathbf{A}^-\mathbf{A}$ gives

$$(d\mathbf{A})\mathbf{A}^-\mathbf{A} + \mathbf{A}(d\mathbf{A}^-)\mathbf{A} + \mathbf{A}\mathbf{A}^-(d\mathbf{A})\mathbf{A}^-\mathbf{A} = (d\mathbf{A})\mathbf{A}^-\mathbf{A}$$

which is

$$\mathbf{A}(d\mathbf{A}^-)\mathbf{A} = -\mathbf{A}\mathbf{A}^-(d\mathbf{A})\mathbf{A}^-\mathbf{A}. \tag{6.75}$$

Recalling that equations $\mathbf{A}\mathbf{x} = \mathbf{y}$ can be solved as $\mathbf{x} = \mathbf{A}^-\mathbf{y}$ we can "solve" (6.75) as

$$d\mathbf{A}^- = -\mathbf{A}^-\mathbf{A}\mathbf{A}^-(d\mathbf{A})\mathbf{A}^-\mathbf{A}\mathbf{A}^-.$$

When $\mathbf{A}$ is nonsingular this gives $d\mathbf{A}^{-1}$. When $\mathbf{A}$ is singular we can either assume that $\mathbf{A}^-$ is reflexive (i.e. that $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$), or if $\mathbf{A}^-$ is not reflexive, use $\mathbf{A}^* = \mathbf{A}^-\mathbf{A}\mathbf{A}^-$ in its place, and $\mathbf{A}^*$ *is* reflexive. In either case we then get $d\mathbf{A}^- = -\mathbf{A}^-(d\mathbf{A})\mathbf{A}^-$.

c. Differentiation for the variance components model

For (6.35) and the expected value that follows it, we want the partial derivative $\partial^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})/\partial\sigma_i^2\,\partial\sigma_j^2$. Its derivation proceeds as follows. Recall from (6.14) and (6.16) that

$$\mathbf{D}_* = \text{var}(\mathbf{u}) = \text{var}\left\{{}_c\,\mathbf{u}_i\right\}_{i=0}^{r} = \left\{{}_d\,\sigma_i^2\mathbf{I}_{q_i}\right\}_{i=0}^{r}$$

and

$$\mathbf{V} = \mathbf{Z}_*\mathbf{D}_*\mathbf{Z}_*' = \sum_{i=0}^{r}\sigma_i^2\mathbf{Z}_i\mathbf{Z}_i'.$$

As here, and in all that follows, we have $i = 0, 1, \ldots, r$. From $\mathbf{D}_*$ and $\mathbf{V}$ we then have the following results:

$$\frac{\partial\mathbf{D}_*}{\partial\sigma_i^2} = \left\{{}_d\,\delta_{ij}\mathbf{I}_{q_j}\right\}_{j=0}^{r} \text{ with } \delta_{ii} = 1 \text{ and } \delta_{ij} = 0 \text{ for } j \neq i.$$

$$\frac{\partial\mathbf{V}}{\partial\sigma_i^2} = \mathbf{Z}_i\mathbf{Z}_i'$$

$$\frac{\partial\mathbf{V}^{-1}}{\partial\sigma_i^2} = -\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}$$

$$\frac{\partial(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-}{\partial\sigma_i^2} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$$

$$\begin{aligned}\frac{\partial^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-}{\partial\sigma_i^2\,\partial\sigma_j^2} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \\ &\quad\times\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \\ &- (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \\ &- (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \\ &+ (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \\ &\quad\times\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-.\end{aligned}$$

Then for

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1},$$

$$\frac{\partial^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}}{\partial\sigma_i^2\,\partial\sigma_j^2} = -(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j' + \mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') \times \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}. \quad (6.76)$$

Ultimately Δ_{ij} of (6.47) comes from $\partial^2(\mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}'\mathbf{D}\mathbf{m})/\partial\sigma_i^2\,\partial\sigma_j^2$ just as (6.76) was needed for (6.35) to yield (6.36). First recall (e.g., Searle et al., 1992, Sec. M.4f) for $\mathbf{X}$ having N rows and rank $r_{\mathbf{X}}$ that for $\mathbf{K}'$ satisfying $\mathbf{K}'\mathbf{X} = \mathbf{0}$ with $\mathbf{K}'$ having full row rank $N - r_{\mathbf{X}}$ we have

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} = \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'. \quad (6.77)$$

Therefore

$$\frac{\partial\mathbf{P}}{\partial\sigma_i^2} = -\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}' = -\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}. \quad (6.78)$$

Furthermore, for $\mathbf{E}_i = \partial\mathbf{D}_*/\partial\sigma_i^2$, which is a block diagonal matrix with the ith block being $\mathbf{I}_{q_i}$ and the rest $\mathbf{0}$,

$$\begin{aligned}\frac{\partial}{\partial\sigma_i^2}(\mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m}) &= \mathbf{m}'\mathbf{E}_i\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} - \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} + \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{E}_i\mathbf{m} \\ &= 2\mathbf{m}_i'\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} - \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m}\end{aligned}$$

because each term is a scalar and so equals its transpose. Therefore twice the value of Δ_{ij} of (6.47) is

$$\begin{aligned}2\Delta_{ij} &= \frac{\partial^2(\mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m})}{\partial\sigma_i^2\,\partial\sigma_j^2} \\ &= -2\mathbf{m}_i'\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} + 2\mathbf{m}_i'\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{m}_j \\ &\quad - \mathbf{m}_j'\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} + \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m}\end{aligned}$$

$$+ \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} - \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{m}_j$$

$$= 2(-\mathbf{m}_i'\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m} + \mathbf{m}_i'\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{m}_j - \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{m}_j$$

$$+ \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m})$$

$$= 2(\mathbf{m}_i' - \mathbf{m}'\mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{Z}_i)\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j(\mathbf{m}_j - \mathbf{Z}_j'\mathbf{P}\mathbf{Z}\mathbf{D}\mathbf{m}). \tag{6.79}$$

as required.

6.13 EXERCISES

E 6.1 Use (6.2), (6.3), and (1.14) to prove (6.6).

E 6.2 (a) Why does $\mathbf{V}^{-1} = \mathbf{L}'\mathbf{L}$ for some non-singular $\mathbf{L}$?

(b) Through using $\mathbf{L}$ of (a), explain why $\mathbf{X}\boldsymbol{\beta}^0$ of (6.19) is invariant to $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$.

E 6.3 For the linear model $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is of full column rank, show that all linear combinations of $\boldsymbol{\beta}$ are estimable.

E 6.4 In the linear model $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$, with $\mathrm{var}(\mathbf{y}) = \mathbf{V}$, where $\mathbf{V}$ is known and nonsingular invertible, show that the MLE of an estimable function $\mathbf{c}'\boldsymbol{\beta}$ is $\mathbf{c}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. Why is it reasonable to assume that $\mathbf{V}$ is nonsingular whereas it is not reasonable to assume $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is?

E 6.5 Consider the balanced one-way random model:

$$y_{ij}|a_i \sim \text{indep. } \mathcal{N}(\mu + a_i, \sigma^2)\ i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n$$

$$a_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma_a^2).$$

Find a $100(1-\alpha)\%$ confidence interval for the intraclass correlation coefficient.

E 6.6 Write the following models in matrix notation and in each case determine the marginal mean and variance of y. If a factor is not specified, assume it is fixed.

(a) $y_{ij}|a_i \sim \text{indep. } \mathcal{N}(\mu + a_i + \beta_j, \sigma^2); a_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma_a^2)$; $i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n$. (Two-way mixed).

(b) $y_{ij}|a_i, b_j \sim$ indep. $\mathcal{N}(\mu + a_i + b_j, \sigma_a^2)$; $a_i \sim$ i.i.d. $\mathcal{N}(0, \sigma_a^2)$; $b_j \sim$ i.i.d. $\mathcal{N}(0, \sigma_b^2)$; a_i and b_j indep. ; $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$. (Two-way random).

(c) $y_{ijk}|a_i, g_{ij} \sim$ indep. $\mathcal{N}(\mu + a_i + \beta_j + g_{ij}, \sigma^2)$; $a_i \sim$ i.i.d. $\mathcal{N}(0, \sigma_a^2)$; $g_{ij} \sim$ i.i.d. $\mathcal{N}(0, \sigma_g^2)$; a_i and g_{ij} indep.; $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$; $k = 1, 2, \ldots, r$. (Two-way mixed with interaction).

E 6.7 Consider the unbalanced one-way random model:

$$y_{ij}|a_i \sim \text{indep. } \mathcal{N}(\mu + a_i, \sigma^2);\ i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n_i$$

$$a_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma_a^2).$$

Show that the sufficient statistics are not complete. One way to do this is to develop an unbiased estimator of zero, or equivalently two different unbiased estimators of the same parameter based on the sufficient statistics. Do this by constructing two different estimators of σ_a^2. *Hint*: Consider the "usual" sums of squares for treatments in a one-way ANOVA, $\Sigma_i n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, and the unweighted version, $\Sigma_i(\bar{y}_{i\cdot} - \bar{y}_u)^2$, where $\bar{y}_u = (1/m)\Sigma_i \bar{y}_{i\cdot}$. An important implication of this result is that there is no UMVUE for σ_a^2.

E 6.8 (a) For $\mathbf{K}'\mathbf{y}$ of Section 6.9 (REML) write the log likelihood; denote it as l_1.

(b) Kenward and Roger write the log likelihood as

$$l_2 = \text{constant } - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y}$$

for $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$. Show that the quadratic forms in $\mathbf{y}$ are the same in l_1 and l_2.

(c) If $\mathbf{V}$ is a function of t, write $\partial\mathbf{V}/\partial t$ as $\mathbf{V}_t$. Show that $\partial l_1/\partial t = \partial l_2/\partial t$.

E 6.9 For $\mathbf{K}$ of Section 6.9, determine the effect on $\mathbf{Z}'\mathbf{P}\mathbf{y}$ of the replacements listed in (6.66).

E 6.10 The ML and REML equations for estimating σ^2 are (6.60) and (6.67), respectively. Use those equations to derive ML and REML

solutions for the following models. In each case $i = 1, \ldots, m$, and $j = 1, \ldots, n$

(a) $\mathrm{E}[y_{ij}] = \mu, \mathbf{V} = \sigma^2 \mathbf{I}_N$.

(b) $\mathrm{E}[y_{ij}] = \mu + \alpha_i, \mathbf{V} = \sigma^2 \mathbf{I}_N, .$

(c) $\mathrm{E}[y_{ij}] = \mu, \mathbf{V} = \sigma^2 \mathbf{I}_N + \left\{ {}_d \; \sigma_a^2 \mathbf{J}_n \right\}_{i=1}^{m}$.

(d) $\mathrm{E}[y_{ij}] = \mu + bx_i, \mathbf{V} = \sigma^2 \mathbf{I}_N$.

E 6.11 In line with definitions (6.21), define $\mathbf{u}_0$ and $\mathbf{u}_* = \begin{bmatrix} \mathbf{u}_0 \\ \mathbf{u} \end{bmatrix}$ so that

$$\mathbf{Z}_* \mathbf{u}_* = [\mathbf{Z}_0 \quad \mathbf{Z}] \begin{bmatrix} \mathbf{u}_0 \\ \mathbf{u} \end{bmatrix} = \mathbf{Z}_0 \mathbf{u}_0 + \mathbf{Z}\mathbf{u}.$$

For β of (6.29) and $\tilde{\mathbf{u}}^0$ of (6.57) show that $\mathbf{y} - \mathbf{X}\beta^0 - \mathbf{Z}\tilde{\mathbf{u}}^0$ is identical to the predictor of $\mathbf{u}_0^0$ derived from $\mathbf{D}_* \mathbf{Z}_*' \mathbf{P}\mathbf{y}$.

Chapter 7

LONGITUDINAL DATA

7.1 INTRODUCTION

Sections 1.5d and 6.2d each briefly describe the general nature of longitudinal data. Their basic feature consists of successive observations on each of a number of subjects (often people or animals). This can be likened to a randomized complete blocks experiment where the subjects, as blocks, are treated as random, and the successive occasions on which observations are taken are akin to treatments. One big difference is that with longitudinal data the correlation structure among observations on the same subject is often more complicated than that among treatments in the same block.

To begin with, and for most of this chapter, we deal with balanced data, meaning that on each subject there is the same number of observations, to be denoted by n. (Unbalanced data is much more difficult to deal with than balanced data.) For ease of description we refer to the occasions when observations are taken as times; thus each of say, m, subjects provides a datum at n times. For subject i, with $i = 1, 2, \ldots, m$, the datum at time j (for $j = 1, 2, \ldots, n$) is denoted by y_{ij} and the vector of data for subject i is

$$\mathbf{y}_i = [y_{i1} \quad y_{i2} \cdots y_{ij} \cdots y_{in}]' = \left\{ {}_c\, y_{ij} \right\}_{j=1}^{n}. \tag{7.1}$$

And the vector of data on all m subjects is

$$\mathbf{y} = \left\{ {}_c\, \mathbf{y}_i \right\}_{i=1}^{m} = \left\{ {}_c \left\{ {}_c\, y_{ij} \right\}_{j=1}^{n} \right\}_{i=1}^{m}. \tag{7.2}$$

7.2 A MODEL FOR BALANCED DATA

a. Prescription

We define the mean and variance of $\mathbf{y}_i$ as being the same for each i:

$$\mathrm{E}[\mathbf{y}_i] = \boldsymbol{\mu} = \left\{{}_c\,\mu_j\right\}_{i=1}^{n} \quad \text{so that} \quad \mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\mu} \text{ for } \mathbf{X} = \mathbf{1}_m \otimes \mathbf{I}_n.$$

And on taking the $\mathbf{y}_i$s to be independent with $\mathrm{var}(\mathbf{y}_i) = \mathbf{V}_0 \ \forall\ i$ we have

$$\mathbf{V} = \mathrm{var}(\mathbf{y}) = \mathbf{I}_m \otimes \mathbf{V}_0. \tag{7.3}$$

b. Estimating the mean

On assuming normality,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\mu}\ , \mathbf{V}),$$

the ML estimator of $\boldsymbol{\mu}$ is $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ from (6.19), whenever $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ exists. This gives

$$\mathrm{ML}(\boldsymbol{\mu}\ ) = \hat{\boldsymbol{\mu}} = [(\mathbf{1}_m' \otimes \mathbf{I}_n)(\mathbf{I}_m \otimes \mathbf{V}_0^{-1})(\mathbf{1}_m' \otimes \mathbf{I}_n)]^{-1}(\mathbf{1}_m' \otimes \mathbf{I}_n)(\mathbf{I}_m \otimes \mathbf{V}_0^{-1})\mathbf{y}$$

which reduces (see E 7.1) to

$$\hat{\boldsymbol{\mu}} = \left\{{}_c\,\bar{y}_{\cdot j}\right\}_{j=1}, \quad \text{i.e.,} \quad \hat{\mu}_j = \bar{y}_{\cdot j}\ . \tag{7.4}$$

It is to be noticed that this result does not involve $\mathbf{V}_0$; thus it holds no matter what $\mathbf{V}_0$ is. And this is important, because in what follows we consider several forms of $\mathbf{V}_0$, but for all of them with $\mathbf{X} = \mathbf{1}_m \otimes \mathbf{I}_n$ the estimator of $\boldsymbol{\mu}$ is as in (7.4).

c. Estimating $\mathbf{V}_0$

In attributing no structure to $\mathbf{V}_0$, we want to estimate (by ML) its every element. This necessitates differentiating the likelihood with respect to every element of $\mathbf{V}_0$. But since the likelihood

$$L = \frac{\exp\{-\frac{1}{2}[\mathbf{y} - (\mathbf{1} \otimes \mathbf{I})\boldsymbol{\mu}\ ]'(\mathbf{I} \otimes \mathbf{V}_0)^{-1}[\mathbf{y} - (\mathbf{1} \otimes \mathbf{I})\boldsymbol{\mu}\ ]\}}{(2\pi)^{\frac{mn}{2}}|\mathbf{I} \otimes \mathbf{V}_0|^{\frac{1}{2}}} \tag{7.5}$$

involves $\mathbf{V}_0^{-1}$ that differentiating is somewhat cumbersome. It can be circumvented by writing

$$\mathbf{V}_0^{-1} = \mathbf{W}$$

and then, ignoring the 2π term,

$$l = \log L = \tfrac{1}{2}\log|\mathbf{I}\otimes\mathbf{W}| - \tfrac{1}{2}\left\{_r\ (\mathbf{y}_i-\boldsymbol{\mu})'\right\}(\mathbf{I}\otimes\mathbf{W})\left\{_c\ \mathbf{y}_i-\boldsymbol{\mu}\right\}. \quad (7.6)$$

We differentiate this with respect to an element w_{jk} of $\mathbf{W}$, for this purpose treating w_{jk} as different from w_{kj} (even though they are equal because $\mathbf{W}$ is symmetric). Then, on recalling that $\partial(\log|\mathbf{A}|)/\partial x = \mathrm{tr}[\mathbf{A}^{-1}(\partial\mathbf{A}/\partial x)]$, and that $\log|\mathbf{I}_m\otimes\mathbf{W}| = m\log|\mathbf{W}|$, and observing that

$$\frac{\partial\mathbf{W}}{\partial w_{jk}} = \mathbf{E}_{jk},$$

where $\mathbf{E}_{ij}$ is a matrix of all zeros except with element (j,k) being one, we have

$$\begin{aligned}\frac{\partial l}{\partial w_{jk}} &= \tfrac{1}{2}m\,\mathrm{tr}(\mathbf{W}^{-1}\mathbf{E}_{jk}) - \tfrac{1}{2}\left\{_r\ (\mathbf{y}_i-\boldsymbol{\mu})'\right\}(\mathbf{I}\otimes\mathbf{E}_{jk})\left\{_c\ \mathbf{y}_i-\boldsymbol{\mu}\right\}\\ &= \tfrac{1}{2}m\,v_{0,jk} - \tfrac{1}{2}\sum_{i=1}^{m}(y_{ij}-\mu_j)(y_{ik}-\mu_k), \qquad (7.7)\end{aligned}$$

where $v_{0,jk}$ is the (j,k)th element of $\mathbf{V}_0$. We already know from (7.4) that $\mathrm{MLE}(\mu_j) = \hat{\mu}_j = \bar{y}_{\cdot j}$. To get the $\mathrm{MLE}(v_{0,jk})$ we equate (7.7) to zero, with μ_j replaced by $\hat{\mu}_j$. Thus

$$\mathrm{MLE}(v_{0,jk}) = \frac{1}{m}\sum_{i=1}^{m}(y_{\cdot j}-\bar{y}_{\cdot j})(y_{ik}-\bar{y}_{\cdot k}).$$

And since this is true for all j,k we have the matrix result

$$\hat{\mathbf{V}}_0 = \frac{1}{m}\left\{\sum_{i=1}^{m}(y_{ij}-\bar{y}_{\cdot j})(y_{ik}-\bar{y}_{\cdot k})\right\}_{j,k=1}^{n}; \qquad (7.8)$$

i.e., $\hat{\mathbf{V}}_0$ is a Wishart matrix. Since the MLE not constraining $\mathbf{V}_0$ to be symmetric, happens to be symmetric, this is also the constrained MLE.

7.3 A MIXED MODEL APPROACH

Having dealt with general $\mathbf{V}_0$, we now consider some special cases, where $\mathbf{V}_0$ is structured in terms of a few (often just two or three) parameters. We do this by specifying a mixed model for the data.

a. Fixed and random effects

A starting point for a model equation for y_{ij} being the datum on subject i taken at time j is

$$\mathrm{E}[y_{ij}|u_i] = \alpha_j + u_i \tag{7.9}$$

where u_i is a random effect for subject i and α_j is now playing the part of μ_j of Section 7.2. And for $\mathbf{y}$ defined in (7.2) and for

$$\mathbf{u}_{m\times 1} = \left\{ {}_c\, u_i \right\}_{i=1}^{m} \tag{7.10}$$

we write

$$\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{7.11}$$

for

$$\mathbf{X} = \mathbf{1}_m \otimes \mathbf{I}_n, \qquad \mathbf{Z} = \mathbf{I}_m \otimes \mathbf{1}_n \qquad \text{and} \qquad \boldsymbol{\beta} = \left\{ {}_c\, \alpha_j \right\}_{j=1}^{n}. \tag{7.12}$$

b. Variances

To set up a variance-covariance matrix $\mathbf{V}$ for $\mathbf{y}$, we begin by defining

$$\mathrm{var}(\mathbf{y}|\mathbf{u}) = \mathbf{R}. \tag{7.13}$$

Then from (1.14)

$$\begin{aligned} \mathbf{V} &= \mathrm{var}(\mathrm{E}[\mathbf{y}|\mathbf{u}]) + \mathrm{E}[\mathrm{var}(\mathbf{y}|\mathbf{u})] \\ &= \mathrm{var}(\mathbf{Z}\mathbf{u}) + \mathrm{E}[\mathbf{R}] \\ &= \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R} \end{aligned}$$

for $\mathbf{D}$ defined as

$$\mathbf{D} = \mathrm{var}(\mathbf{u}).$$

Thus

$$\mathbf{V} = (\mathbf{I}_m \otimes \mathbf{1}_n)(\mathbf{D} \otimes 1)(\mathbf{I}_m \otimes \mathbf{1}_n') + \mathbf{R} \tag{7.14}$$

$$= \mathbf{D} \otimes \mathbf{J}_n + \mathbf{R}. \tag{7.15}$$

For $\mathbf{D}$ it is not unusual to attribute correlation among the u_is; as in genetics, when subjects can be siblings (such as dairy cows), or even

litter mates (as with pigs or laboratory mice or rats). Thus, along with assuming the same variance for each subject effect we take

$$\text{var}(u_i) = \sigma_u^2 \; \forall \, i \qquad \text{and} \qquad \text{corr}(u_i, u_k) = \rho_u \; \forall \, i \neq k. \tag{7.16}$$

This gives $\mathbf{D}$ as a matrix having diagonal elements σ_u^2 and off-diagonal elements $\rho_u \sigma_u^2$. Thus

$$\mathbf{D} \;=\; \text{var}(\mathbf{u}) = \sigma_u^2[(1-\rho_u)\mathbf{I}_m + \rho_u \mathbf{J}_m]. \tag{7.17}$$

and so $\mathbf{V}$ of (7.15) is

$$\mathbf{V} \;=\; \sigma_u^2[(1-\rho_u)\mathbf{I}_m + \rho_u \mathbf{J}_m] \otimes \mathbf{J}_n + \mathbf{R}. \tag{7.18}$$

A tractable form for $\mathbf{R} = \text{var}(\mathbf{y}|\mathbf{u})$ is to take the $\mathbf{y}_i|\mathbf{u}_i$ variables as being independent and all having the same variance-covariance matrix $\mathbf{R}_0$, so that $\mathbf{R} = \mathbf{I}_m \otimes \mathbf{R}_0$ so giving

$$\mathbf{V} = \sigma_u^2[(1-\rho_u)\mathbf{I}_m + \rho_u \mathbf{J}_m] \otimes \mathbf{J}_n + \mathbf{I}_m \otimes \mathbf{R}_0. \tag{7.19}$$

For balanced data, this is a fairly general form for $\mathbf{V}$. It provides for a uniform variance σ_u^2 of the random effects, for a correlation ρ_u between them, and for the same variance, $\mathbf{R}_0$, of $\mathbf{y}_i|u_i$ for each subject. With $\sigma_u^2 = 0$ it reduces to $\mathbf{V} = \mathbf{I} \otimes \mathbf{R}_0$, which is the same form as $\mathbf{I} \otimes \mathbf{V}_0$ treated earlier; and for $\rho_u = 0$ it is also

$$\mathbf{V} = \mathbf{I}_m \otimes \mathbf{V}_0 \qquad \text{for} \qquad \mathbf{V}_0 = \sigma_u^2 \mathbf{J}_n + \mathbf{R}_0. \tag{7.20}$$

7.4 PREDICTING RANDOM EFFECTS

We start from the general expression for $\tilde{\mathbf{u}}^0$, the best linear unbiased predictor of the random effects

$$\tilde{\mathbf{u}}^0 = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0). \tag{7.21}$$

This result is presented in Section 6.5 and is established in Chapter 9. Clearly this expression assumes that $\mathbf{D} = \text{var}(\mathbf{u})$ and $\mathbf{V} = \text{var}(\mathbf{y})$ are both known. $\mathbf{Z}$ and $\mathbf{X}$ are taken as $\mathbf{Z} = \mathbf{I}_m \otimes \mathbf{1}_n$ and $\mathbf{X} = \mathbf{1}_m \otimes \mathbf{I}_n$ of (7.12), and for our purpose $\boldsymbol{\beta}^0$ is taken as $\hat{\boldsymbol{\mu}}$ of (7.4), $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\beta}} = \left\{ {}_c \; \hat{\alpha}_j \right\} = \left\{ {}_c \; \bar{y}_{\cdot j} \right\}$. Thus (7.21) becomes

$$\tilde{\mathbf{u}}^0 \;=\; \mathbf{D}(\mathbf{I}_m \otimes \mathbf{1}_n')\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{7.22}$$

$$\;=\; \mathbf{D}(\mathbf{I}_m \otimes \mathbf{1}_n')\mathbf{V}^{-1} \left\{ {}_c \; \mathbf{y}_i - \hat{\boldsymbol{\beta}} \right\}_{i=1}^{m}.$$

Simplifications of this for special cases are as follows.

a. Uncorrelated subjects

With $\rho_u = 0$, the correlation between subjects is taken as being zero; and it reduces (7.18) to

$$\mathbf{V} = \mathbf{I}_m \otimes (\sigma_u^2 \mathbf{J}_n + \mathbf{R}_0).$$

The predictor (7.22) then reduces (see Section 7.10a) to

$$\tilde{\mathbf{u}}^0 = \left[\mathbf{I}_m \otimes \frac{\mathbf{1}'\mathbf{R}_0^{-1}}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}\right] \left\{ {}_c\, \mathbf{y}_i - \hat{\boldsymbol{\beta}} \right\}_{i=1}^{m}.$$

This gives

$$\tilde{u}_i^0 = \frac{\mathbf{1}'\mathbf{R}_0^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}})}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}} \qquad \text{for} \qquad \hat{\boldsymbol{\beta}} = \left\{ {}_c\, \bar{\mathbf{y}}_{\cdot j} \right\}. \tag{7.23}$$

This is not particularly tractable unless $\mathbf{R}_0^{-1}$ is analytically manageable; but numerically it will usually offer little difficulty, especially because $\mathbf{R}_0$ has order n, the number of observations on a subject, and this is often not very large.

b. Uncorrelated between, and within, subjects

Here the correlation between subjects, and between observations on each subject, are each taken as zero. This simply involves putting $\mathbf{R}_0 = \sigma^2\mathbf{I}$ in (7.23) which (in Section 7.10b) yields

$$\tilde{u}_i^0 = \frac{n\sigma_u^2}{\sigma^2 + n\sigma_u^2}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}). \tag{7.24}$$

This is a very familiar estimator, known as a *Stein*, or *shrinkage*, *estimator*. It occurs widely in animal genetics when wanting to calculate the estimated genetic value of animals. With genetic definitions

$$\tau \;=\; \text{repeatability} = \frac{\sigma_u^2}{\sigma^2 + \sigma_u^2}$$

and

$$h^2 \;=\; \text{heritability} = \frac{4\sigma_u^2}{\sigma^2 + \sigma_u^2}$$

the fraction multiplying $(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$ then has several forms that are very familiar to animal geneticists:

$$\frac{n\sigma_u^2}{\sigma^2 + n\sigma_u^2} = \frac{\sigma_u^2}{\sigma^2/n + \sigma_u^2} = \frac{n\tau}{1 + (n-1)\tau} = \frac{nh^2}{4 + (n-1)h^2}. \tag{7.25}$$

c. Uncorrelated between, and autocorrelated within, subjects

Another tractable form for $\mathbf{R}_0/\sigma^2$ is a first-order autocorrelation matrix, a 5×5 example of which is

$$\mathbf{A} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

with

$$\mathbf{A}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & -\rho & 1 \end{bmatrix}. \quad (7.26)$$

Using that inverse (generalized to order n) for $\mathbf{R}_0^{-1}\sigma^2$ in (7.23), the value of $\tilde{u}_i^0$ is (see E 7.3)

$$\tilde{u}_i^0 = \frac{\sigma_u^2 \left[(1-\rho)n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + \rho(y_{i1} - \bar{y}_{\cdot 1} + y_{in} - \bar{y}_{\cdot n})\right]}{\sigma^2(1+\rho) + \sigma_u^2[n - (n-2)\rho]}. \quad (7.27)$$

Note that the $\rho(y_{i1} - \bar{y}_{\cdot 1} + y_{in} - \bar{y}_{\cdot n})$ in the numerator represents "end effects" commensurate with the autocorrelation matrix $\mathbf{A}^{-1}$ of (7.26) having first and last diagonal elements different from all other diagonal elements. Also note that for large n, (7.27) reduces to $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$, as it should.

d. Correlated between, but not within, subjects

In each of the preceding cases $\mathbf{V}$ has been of the form $\mathbf{I}_m \otimes \mathbf{V}_0$. Now, though,

$$\mathbf{V} = \sigma_u^2[(1-\rho_u)\mathbf{I}_m + \rho_u \mathbf{J}_m] \otimes \mathbf{J}_n + \sigma^2(\mathbf{I}_m \otimes \mathbf{I}_n) \quad (7.28)$$

$$= (\mathbf{I}_m \otimes \mathbf{V}_0) + \lambda(\mathbf{J}_m \otimes \mathbf{J}_n)$$

for

$$\mathbf{V}_0 = (1-\rho_u)\mathbf{J}_n\sigma_u^2 + \sigma^2\mathbf{I}_n \quad \text{and} \quad \lambda = \rho_u\sigma_u^2. \tag{7.29}$$

We now invoke a theorem (e.g., Puntanen and Styan, 1989) that if $\mathbf{V}$ and $\mathbf{X}$ are such that $\mathbf{VX} = \mathbf{XH}$ for some $\mathbf{H}$ (see E 7.2 for a proof with $\mathbf{H}^{-1}$ existing) then $\hat{\boldsymbol{\beta}}$, the MLE, is the ordinary least squares estimator. This is so here because

$$\begin{aligned}\mathbf{VX} &= (\mathbf{I}\otimes\mathbf{V}_0)(\mathbf{1}_m\otimes\mathbf{I}_n) + \lambda(\mathbf{J}_m\otimes\mathbf{J}_n)(\mathbf{1}_m\otimes\mathbf{I}_n)\\ &= (\mathbf{1}_m\otimes\mathbf{V}_0) + \lambda(m\mathbf{1}_m\otimes\mathbf{J}_n)\\ &= (\mathbf{1}_m\otimes\mathbf{I}_n)[(\mathbf{I}\otimes\mathbf{V}_0) + \lambda(m\mathbf{I}_m\otimes\mathbf{J}_n)] = \mathbf{XH}\end{aligned}$$

for $\mathbf{H}$ being the matrix in the square brackets. Thus we continue to have $\hat{\boldsymbol{\beta}}$ yielding $\hat{\boldsymbol{\beta}} = \left\{{}_c\,\bar{y}_{\cdot j}\right\}$, as in the preceding sections.

To derive the predicted value $\tilde{\mathbf{u}}^0$ of (7.21) using $\mathbf{V}$ of (7.28) does, however, take some considerable algebra—as shown in Section 7.10c. The result is

$$\tilde{u}_i^0 = \frac{n\sigma_u^2(1-\rho_u)}{\sigma^2 + n\sigma_u^2(1-\rho_u)}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}). \tag{7.30}$$

Several features of this result merit comment.

1. For large n, $\tilde{u}_i^0$ tends to $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$.
2. $\rho_u = 0$ gives $\tilde{u}_i^0$ of (7.24).
3. In the $\rho_u = 0$ result of (7.24), replacing σ_u^2 by $\sigma_u^2(1-\rho_u)$ gives the $\rho_u \neq 0$ result in (7.30). This seems reasonable on the grounds that $\rho_u \neq 0$ effectively represents a reduction in the variance of the u_is.
4. The $\rho_u \neq 0$ predictor of (7.30) is always less than the $\rho_u = 0$ predictor of (7.24).
5. Increases in ρ_u lead to decreases in $\tilde{u}_i^0$ of (7.30). This is understandable because for large correlations among the u_is one would expect them to be more alike than for zero (or small) correlations. Indeed, for $\rho_u = 1$, its maximum value, every $\tilde{u}_i^0$ is the same, namely zero.
6. On defining

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

(7.30) becomes

$$\tilde{u}_i^0 = \frac{n\rho(1-\rho_u)}{1-\rho+n\rho(1-\rho_u)}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$$

which increases as ρ increases.

7.5 ESTIMATING PARAMETERS

Having dealt with estimating β and predicting $\mathbf{u}$ in $\mathbf{X}\beta + \mathbf{Zu}$ (for $\mathbf{V}$ assumed known) we now consider estimating the parameters that occur in the forms of $\mathbf{V}$ considered in Section 7.4.

a. The general case

The distributional assumption for the variance components model of Chapter 6, as in (6.56), is

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \quad \mathbf{V} = \sum_i \mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2)$$

and equations (6.60) for ML estimation of the σ^2s are

$$\left\{ {}_c \operatorname{tr}(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i') \right\} = \left\{ {}_c \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} \right\} \tag{7.31}$$

where the appearance of $\mathbf{Z}_i\mathbf{Z}_i'$ comes about because

$$\mathbf{Z}_i\mathbf{Z}_i' = \frac{\partial \mathbf{V}}{\partial \sigma_i^2}.$$

Also

$$\mathbf{P}\mathbf{y} = \left[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\right]\mathbf{y} = \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta^0)$$

for $\beta^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ from (6.29). Hence (7.31) is

$$\left\{ {}_c \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\right) \right\} = \left\{ {}_c (\mathbf{y} - \mathbf{X}\beta^0)'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta^0) \right\}. \tag{7.32}$$

As in Section 6.4, we now think of $\mathbf{V}$ being structured, with elements which are functions of just a few scalar parameters: denote one such

element as φ playing the part of σ_i^2 in (7.32). Its equations are of the form

$$\text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \varphi}\right) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^0)'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \varphi}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^0) \quad (7.33)$$

$$= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^0)'(-1)\frac{\partial \mathbf{V}^{-1}}{\partial \varphi}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^0). \quad (7.34)$$

We use either (7.33) or (7.34) for special cases of $\mathbf{V} = \mathbf{I} \otimes \mathbf{V}_0$, often using the notation

$$\text{LHS}(\varphi) = \text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \varphi}\right) = m\ \text{tr}\left(\mathbf{V}_0^{-1}\frac{\partial \mathbf{V}_0}{\partial \varphi}\right) \quad (7.35)$$

and

$$\text{RHS}(\varphi) = \text{right-hand side of (7.33) and (7.34)},$$

so that the estimating equations are then

$$\text{LHS}(\varphi) = \text{RHS}(\varphi). \quad (7.36)$$

for φ taking in turn each parameter in $\mathbf{V}$, e.g., ρ_u, σ_u^2 and σ^2 in $\mathbf{V}$ of (7.18) when $\mathbf{R} = \sigma^2\mathbf{I}_n$. We do this for the four cases of Section 7.4.

b. Uncorrelated subjects

This has

$$\mathbf{V} = \mathbf{I}_m \otimes \mathbf{V}_0 \qquad \text{for} \qquad \mathbf{V}_0^{-1} = \sigma_u^2\mathbf{J}_n + \mathbf{R}_0$$

with

$$\mathbf{V}_0^{-1} = \mathbf{R}_0^{-1} - \frac{\mathbf{R}_0^{-1}\mathbf{1}\mathbf{1}'\mathbf{R}_0^{-1}}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}.$$

Using this in (7.35) for $\varphi = \sigma_u^2$ gives

$$\text{LHS}(\sigma_u^2) = m\ \text{tr}\left(\mathbf{V}_0^{-1}\mathbf{J}_n\right) = m\mathbf{1}'\mathbf{V}_0^{-1}\mathbf{1} = \frac{m\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}{1+\sigma_u^2\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}. \quad (7.37)$$

Similarly, using (7.34) gives

$$\text{RHS}(\sigma_u^2) = \left\{{}_r\,(\mathbf{y}_i-\hat{\boldsymbol{\beta}})'\right\}_{i=1}^m \left(\mathbf{I}_m \otimes \frac{\mathbf{R}_0^{-1}\mathbf{1}\mathbf{1}'\mathbf{R}_0^{-1}}{1+\sigma_u^2\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1})^2}\right)\left\{{}_c\,\mathbf{y}_i-\hat{\boldsymbol{\beta}}\right\}_{i=1}^m$$

$$= \sum_{i=1}^m \left[\frac{\mathbf{1}'\mathbf{R}_0^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\beta}})}{1+\sigma_u^2\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}\right]^2$$

and so the ML equation is

$$m\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}\left(1+\dot{\sigma}_u^2\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}\right)=\sum_{i=1}^{m}\left[\mathbf{1}'\mathbf{R}_0^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\beta}})\right]^2. \tag{7.38}$$

If $\mathbf{R}_0$ is unspecified, with elements functionally independent of σ_u^2 (which has been assumed in deriving the preceding result), then ML estimation of $\mathbf{R}_0$ will be exactly like that of $\mathbf{V}_0$ in Section 7.2b. We therefore consider having $\mathbf{R}_0=\sigma^2\mathbf{I}_n$.

c. Uncorrelated between, and within, subjects

First, for $\mathbf{R}_0=\sigma^2\mathbf{I}_n$ the estimating equation (7.38) becomes

$$(mn/\dot{\sigma}^2)(1+n\dot{\sigma}_u^2/\dot{\sigma}^2) = \sum_{i=1}^{m}(y_{i\cdot}-\sum_j\bar{y}_{\cdot j})^2/\dot{\sigma}^4;$$

that is,

$$m(\dot{\sigma}^2+n\dot{\sigma}_u^2) = \sum_i\sum_j(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2. \tag{7.39}$$

And, with $\mathbf{V}$ now being

$$\mathbf{V} = \mathbf{I}_m\otimes\left(\sigma_u^2\mathbf{J}_n+\sigma^2\mathbf{I}_n\right),$$

$$\mathbf{V}^{-1} = \mathbf{I}_m\otimes\frac{1}{\sigma^2}\left(\mathbf{I}_n-\frac{\sigma_u^2}{\sigma^2+n\sigma_u^2}\mathbf{J}_n\right)$$

and

$$\text{LHS}(\sigma^2)=\text{tr}\left(\frac{\partial\mathbf{V}}{\partial\sigma^2}\mathbf{V}^{-1}\right) = m\,\text{tr}\left[\frac{1}{\sigma^2}\left(\mathbf{I}_n-\frac{\sigma_u^2}{\sigma^2+n\sigma_u^2}\mathbf{J}_n\right)\right]$$

$$= \frac{mn}{\sigma^2}\left(1-\frac{\sigma_u^2}{\sigma^2+n\sigma_u^2}\right). \tag{7.40}$$

Then, with

$$-\frac{\partial\mathbf{V}^{-1}}{\partial\sigma^2} = \mathbf{I}_m\otimes\left[\frac{1}{\sigma^4}\mathbf{I}_n-\frac{\sigma_u^2(2\sigma^2+n\sigma_u^2)}{\sigma^4(\sigma^2+n\sigma_u^2)^2}\mathbf{J}_n\right]$$

$$\text{RHS}(\sigma^2) = \sum_{i=1}^{m}\left[\frac{1}{\sigma^4}\sum_j(y_{ij}-\bar{y}_{\cdot j})^2-\frac{\sigma_u^2(2\sigma^2+n\sigma_u^2)}{\sigma^4(\sigma^2+n\sigma_u^2)^2}\,n^2(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2\right].$$

Note that

$$\begin{aligned}\frac{n\sigma_u^2(2\sigma^2+n\sigma_u^2)}{\sigma^4(\sigma^2+n\sigma_u^2)^2} &= \frac{2n\sigma^2\sigma_u^2+n^2\sigma_u^4+\sigma^4-\sigma^4}{\sigma^4(\sigma^2+n\sigma_u^2)^2} \\ &= \frac{1}{\sigma^4}-\frac{1}{(\sigma^2+n\sigma_u^2)^2}\end{aligned}$$

so that gives

$$\begin{aligned}\text{RHS}(\sigma^2) &= \frac{1}{\sigma^4}\left[\sum_i\sum_j(y_{ij}-\bar{y}_{\cdot j})^2-\sum_i n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2+\frac{\sum_i n(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}{(\sigma^2+n\sigma_u^2)^2}\right] \\ &= \frac{\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2}{\sigma^4}+\frac{\sum_i\sum_j(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2}{(\sigma^2+n\sigma_u^2)^2}.\end{aligned}$$

Now define

$$\text{SSE}=\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2 \qquad \text{and} \qquad \text{SSB}=\sum_i\sum_j(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2,$$

a between-subjects sum of squares. The estimating equation is

$$\frac{mn}{\dot{\sigma}^2}\left(1-\frac{\dot{\sigma}_u^2}{\dot{\sigma}^2+n\dot{\sigma}_u^2}\right)=\frac{\text{SSE}}{\dot{\sigma}^4}+\frac{\text{SSB}}{(\dot{\sigma}^2+n\dot{\sigma}_u^2)^2}; \tag{7.41}$$

and the earlier equation, (7.39), is

$$m(\dot{\sigma}^2+n\dot{\sigma}_u^2)=\text{SSB}. \tag{7.42}$$

These have to be solved for $\dot{\sigma}^2$ and $\dot{\sigma}_u^2$. To do this, begin by substituting (7.42) into (7.41):

$$\frac{mn}{\dot{\sigma}^2}\left(1-\frac{\dot{\sigma}_u^2}{\dot{\sigma}^2+n\dot{\sigma}_u^2}\right) = \frac{\text{SSE}}{\dot{\sigma}^4}+\frac{m}{(\dot{\sigma}^2+n\dot{\sigma}_u^2)^2}.$$

But

$$\frac{\dot{\sigma}_u^2}{\dot{\sigma}^2+n\dot{\sigma}_u^2} = \frac{1}{n}\left(1-\frac{\dot{\sigma}^2}{\dot{\sigma}^2+n\dot{\sigma}_u^2}\right)$$

and so

$$\frac{mn}{\dot{\sigma}^2}\left(1-\frac{1}{n}+\frac{\dot{\sigma}^2}{n(\dot{\sigma}^2+n\dot{\sigma}_u^2)}\right)=\frac{\text{SSE}}{\dot{\sigma}^4}+\frac{m}{\dot{\sigma}^2+n\dot{\sigma}_u^2},$$

which leads to

$$\dot{\sigma}^2 = \frac{\text{SSE}}{m(n-1)} = \left(1 - \frac{1}{m}\right) \text{MSE} \tag{7.43}$$

because $\text{SSE} = (m-1)(n-1)\text{MSE}$. And then from (7.42)

$$\dot{\sigma}_u^2 = \frac{1}{n}\left(\frac{\text{SSB}}{m} - \dot{\sigma}^2\right). \tag{7.44}$$

These results agree exactly, as they should, with those for $\dot{\sigma}_e^2$ and $\dot{\sigma}_\beta^2$ at the bottom of page 150 of Searle et al. (1992). Confirmation of this demands very careful consideration of the two notations, see E 7.5.

d. Uncorrelated between, and autocorrelated within, subjects

We here have $\mathbf{V} = \mathbf{I}_m \otimes \mathbf{A}\sigma^2$ for $\mathbf{A}_{n\times n}$ being the n-order form of the 5×5 example in (7.26). Therefore in (7.34) and (7.35) $\mathbf{A}\sigma^2$ plays the part of $\mathbf{V}_0$. So from using (7.26) for order n in $\mathbf{V} = \mathbf{I} \otimes \mathbf{V}_0 = \mathbf{I} \otimes \sigma^2\mathbf{A}$ we get from (7.35) for $\varphi = \sigma^2$

$$\text{LHS}(\sigma^2) = m \text{ tr}\left[(\sigma^2\mathbf{A})^{-1}\frac{\partial(\sigma^2\mathbf{A})}{\partial\sigma^2}\right] = m \text{ tr}(\mathbf{A}^{-1}\mathbf{A})/\sigma^2 = mn/\sigma^2.$$

And from $\text{RHS}(\varphi)$ of (7.34)

$$\begin{aligned}
\text{RHS}(\sigma^2) &= -(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\left[\mathbf{I}_m \otimes \frac{\partial(\mathbf{A}\sigma^2)^{-1}}{\partial\sigma^2}\right](\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (7.45)\\
&= \left\{{}_r\,(\mathbf{y}_i - \hat{\boldsymbol{\beta}})'\right\}\left[\mathbf{I}_m \otimes \frac{-\mathbf{A}^{-1}}{\sigma^4}\right]\left\{{}_c\,\mathbf{y}_i - \hat{\boldsymbol{\beta}}\right\}\\
&= \frac{1}{\sigma^4}\sum_{i=1}^{m}(\mathbf{y}_i - \hat{\boldsymbol{\beta}})'\mathbf{A}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}). \qquad (7.46)
\end{aligned}$$

With $\hat{\boldsymbol{\beta}}$ from (7.4), we can write

$$\mathbf{y}_i - \hat{\boldsymbol{\beta}} = \left\{{}_c\,y_{ij} - \bar{y}_{\cdot j}\right\}_{j=1}^{m} \equiv \left\{{}_c\,\delta_{ij}\right\}_{j=1}^{m}, \tag{7.47}$$

so defining δ_{ij}. Then, on using (7.26) for $\mathbf{A}^{-1}$ generalized from order 5 to order n, we get

$$\text{RHS}(\sigma^2) = \frac{1}{\sigma^4}\frac{1}{1-\rho^2}\sum_{i=1}^{m}\left[(1+\rho^2)\sum_{j=1}^{n}\delta_{ij}^2 - \rho^2(\delta_{i1}^2+\delta_{in}^2) - 2\rho\sum_{j=2}^{n}\delta_{ij}\delta_{i,j-1}\right]. \quad (7.48)$$

Therefore $\text{LHS}(\sigma^2) = \text{RHS}(\sigma^2)$ gives the estimating equation

$$mn\hat{\sigma}^2(1-\hat{\rho}^2) = (1+\hat{\rho}^2)\sum_{i=1}^{m}\sum_{j=1}^{n}\delta_{ij}^2 - \hat{\rho}^2\sum_{i=1}^{m}(\delta_{i1}^2+\delta_{in}^2) - 2\hat{\rho}\sum_{i=1}^{m}\sum_{j=2}^{n}\delta_{ij}\delta_{i,j-1}. \quad (7.49)$$

Now doing the same thing for $\varphi = \rho$ gives

$$\text{LHS}(\rho) = m\ \text{tr}\left[(\sigma^2\mathbf{A})^{-1}\frac{\partial(\sigma^2\mathbf{A})}{\partial\rho}\right] = m\ \text{tr}\left(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\rho}\right).$$

On generalizing $\mathbf{A}^{-1}$ and $\mathbf{A}$ of (7.26) to order n it will be found (see E 7.6) that $\text{LHS}(\rho)$ reduces to

$$\text{LHS}(\rho) = \frac{-2m(n-1)\rho}{1-\rho^2}.$$

And $\text{RHS}(\rho)$ will be the same as $\text{RHS}(\sigma^2)$ of (7.45) but with

$$\frac{\partial(\mathbf{A}\sigma^2)^{-1}}{\partial\sigma^2} = \frac{-\mathbf{A}^{-1}}{\sigma^4} \text{ replaced by } \frac{1}{\sigma^2}\frac{\partial(\mathbf{A}^{-1})}{\partial\rho}.$$

Therefore from (7.46)

$$\text{RHS}(\rho) = \frac{1}{\sigma^2}\sum_{i=1}^{m}(\mathbf{y}_i-\hat{\boldsymbol{\beta}})'\frac{\partial\mathbf{A}^{-1}}{\partial\rho}(\mathbf{y}_i-\hat{\boldsymbol{\beta}}).$$

Now in looking at $\mathbf{A}^{-1}$ it will be found that in $\mathbf{A}^{-1}$ the element

$$\left.\begin{array}{lll} 1/(1-\rho^2) & \text{becomes} & 2\rho/(1-\rho^2)^2 \\ -\rho/(1-\rho^2) & \text{becomes} & -(1+\rho^2)/(1-\rho^2)^2 \\ (1+\rho^2)/(1-\rho^2)^2 & \text{becomes} & 4\rho(1-\rho^2)^2 \end{array}\right\} \text{in } \frac{\partial\mathbf{A}^{-1}}{\partial\rho}.$$

Therefore, in making these changes in going from (7.46) to (7.48) for RHS(ρ) gives

$$\text{RHS}(\rho) = -\frac{1}{\sigma^2}\frac{1}{(1-\rho^2)^2}\sum_{i=1}^{m}\left[4\rho\sum_{j=1}^{n}\delta_{ij}^2 - 2\rho(\delta_{i1}^2+\delta_{in}^2)\right.$$

$$\left.-2(1+\rho^2)\sum_{j=2}^{n}\delta_{ij}\delta_{i,j-1}\right].$$

Then equating this to LHS(ρ) gives the second estimating equation as

$$m(n-1)\hat{\rho}\hat{\sigma}^2(1-\hat{\rho}^2) = 2\hat{\rho}\sum_{i=1}^{m}\sum_{j=1}^{n}\delta_{ij}^2 - \hat{\rho}\sum_{i=1}^{m}(\delta_{i1}^2+\delta_{in}^2)$$

$$-(1+\hat{\rho}^2)\sum_{i=1}^{m}\sum_{j=2}^{n}\delta_{ij}\delta_{i,j-1}. \qquad (7.50)$$

Clearly, this and (7.49) have to be solved numerically. They appear to have no algebraic solution for $\hat{\sigma}^2$ and $\hat{\rho}$. And the solutions will be ML estimators only if $1 \leq \hat{\rho} \leq 1$ and $\hat{\sigma}^2 > 0$.

e. Correlated between, but not within, subjects

We have from (7.28)

$$\mathbf{V} = \sigma_u^2[(1-\rho_u)\mathbf{I}_m + \rho_u\mathbf{J}_m] \otimes \mathbf{J}_n + \sigma^2\mathbf{I}_m \otimes \mathbf{I}_n$$

and from Section 7.10c

$$\mathbf{V}^{-1} = \frac{-\sigma_u^2/\sigma^2}{n\sigma_u^2(1-\rho_u)+\sigma^2}\left[(1-\rho_u)\mathbf{I}_m + \frac{\sigma^2\rho_u}{\sigma^2+n\sigma_u^2+(m-1)n\rho_u\sigma_u^2}\mathbf{J}_m\right]\otimes\mathbf{J}_n + \frac{1}{\sigma^2}\mathbf{I}_{mn}.$$

Although differentiating $\mathbf{V}$ with respect to each of the parameters ρ_u, σ_u^2 and σ^2 is easy, the $LHS(\varphi)$ expressions give nothing that is simple. For example

$$LHS(\rho_u) = \frac{m(m-1)n^2\rho_u}{[n\sigma_u^2(1-\rho_u)+\sigma^2][n\sigma_u^2(1-\rho_u)+\sigma^2+mn\rho_u\sigma_u^2]}.$$

We therefore pursue no further attempts at finding analytic forms of the MLEs of ρ_u, σ_u^2 and σ^2.

7.6 UNBALANCED DATA

a. Example and model

Suppose a clinical trial consists of $m = 5$ patients observed on $n = 7$ occasions but where some of the patients from time to time fail to visit the clinic. Table 7.1 is an example of this.

Table 7.1. Patients Visiting a Clinic ($\surd$ indicates a visit)

Patient	Clinic Visit $j = 1, \ldots, n$						
$i = 1, \ldots, m$	1	2	3	4	5	6	7
1	$\surd$	$\surd$		$\surd$		$\surd$	$\surd$
2	$\surd$		$\surd$			$\surd$	$\surd$
3	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$		
4		$\surd$		$\surd$			$\surd$
5	$\surd$			$\surd$	$\surd$		$\surd$

We still write $\mathrm{E}[\mathbf{y}|\boldsymbol{\mu}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ as in (7.11) but now the specification of $\mathbf{X}$ and $\mathbf{Z}$ is more complicated than in (7.12). For Table 7.1 the data for patient 1 have

$$\mathrm{E}[\mathbf{y}_1|u_1] = \begin{bmatrix} \mathrm{E}[y_{11}|u_1] \\ \mathrm{E}[y_{12}|u_1] \\ \mathrm{E}[y_{14}|u_1] \\ \mathrm{E}[y_{16}|u_1] \\ \mathrm{E}[y_{17}|u_1] \end{bmatrix} = \begin{bmatrix} 1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0 \\ 0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&1 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} u_1 \quad (7.51)$$

$$= \mathbf{X}_1\boldsymbol{\beta} + \mathbf{1}_{n_1} u_1 \quad (7.52)$$

where n_i is the number of times patient i visits the clinic using $i = 1$ in (7.52).

Assembling $\mathbf{y} = \left\{ {}_c\, \mathbf{y}_i \right\}_{i=1}^{m}$ as in (7.2) gives

$$\mathrm{E}[\mathbf{y}|\mathbf{u}] = \left\{ {}_c\, \mathbf{X}_i \right\}_{i=1}^{m} \boldsymbol{\beta} + \left\{ {}_d\, \mathbf{1}_{n_i} \right\}_{i=1}^{m} \mathbf{u}. \quad (7.53)$$

where $\boldsymbol{\beta}$ is $n \times 1$ and $\mathbf{u}$ is $m \times 1$. In comparing (7.51) and (7.52), it is apparent that for patient i, $\mathbf{X}_i$ is $n_i \times n$ with one row for each checkmark for that patient in Table 7.1; and that row of $\mathbf{X}_i$ is null except for a one corresponding to the checkmark. As examples, in the

first line of Table 7.1 there are checkmarks for $j = 1$ and $j = 2$. These generate [1 0 0 0 0 0 0] and [0 1 0 0 0 0 0] as rows of $\mathbf{X}_i$.

Using $\text{var}(\mathbf{y}|\mathbf{u}) = \mathbf{R}$, (7.53), and with $\mathbf{D} = \text{var}(\mathbf{u})$ of (7.17), gives

$$\begin{aligned} \text{var}(\mathbf{y}) = \mathbf{V} &= \left\{_d\, \mathbf{1}_{n_i}\right\}_{i=1}^{m} \sigma_u^2 [(1-\rho_u)\mathbf{I}_m + \rho_u \mathbf{J}_m] \left\{_d\, \mathbf{1}'_{n_i}\right\}_{i=1}^{m} + \mathbf{R} \\ &= \sigma_u^2(1-\rho_u)\left\{_d\, \mathbf{J}_{n_i}\right\}_{i=1}^{m} + \sigma_u^2 \rho_u \mathbf{J}_N + \mathbf{R} \end{aligned} \tag{7.54}$$

where

$$N = \sum_{i=1}^{m} n_i . \tag{7.55}$$

We now consider three special cases similar to those dealt with in Section 7.4 for deriving $\tilde{\mathbf{u}}$. Now, though, $\mathbf{X} = \left\{_c\, \mathbf{X}_i\right\}$ of (7.53) is no longer as simple as the $\mathbf{1}_m \otimes \mathbf{I}_n$ of (7.12) for balanced data, as is evident from the description of $\mathbf{X}_i$ following (7.53). As a result, there is no theorem such as that mentioned in Section 7.4. Therefore for each of our three special cases we deal with $\mathbf{V}$, $\mathbf{V}^{-1}$, $\hat{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$.

b. Uncorrelated subjects

For balanced data, every patient had n observations and one could assume the same variance-covariance matrix, $\mathbf{R}_0$, for all patients and so have $\mathbf{R} = \mathbf{I}_m \otimes \mathbf{R}_0$ as in Section 7.4a. But that is not possible for patient i having n_i data with n_i not being the same for all patients. The nearest counterpart is to have $\mathbf{R}$ be block diagonal, of blocks of order n_i: $\mathbf{R}_0 = \left\{_d\, \mathbf{R}_i\right\}_{i=1}^{m}$.

Notation

At this point our curly bracket notation mostly involves i ranging from 1 to m, so we cease indicating that range unless context demands it.

– i. *Matrix* V *and its inverse*

Then with $\mathbf{R} = \left\{_d\, \mathbf{R}_i\right\}$ and $\rho_u = 0$, (7.54) gives

$$\mathbf{V} = \left\{_d\, \mathbf{V}_i\right\} \quad \text{for} \quad \mathbf{V}_i = \sigma_u^2 \mathbf{J}_{n_i} + \mathbf{R}_i \tag{7.56}$$

and so

$$\mathbf{V}^{-1} = \left\{_d\, \mathbf{V}_i^{-1}\right\}. \tag{7.57}$$

– ii. ***Estimating the fixed effects***

With

$$\mathbf{X} = \left\{ {}_{c}\, \mathbf{X}_i \right\}$$

from (7.53), we get

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \sum_{i=1}^{m} \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i$$

and

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \sum \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i \tag{7.58}$$

and so

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \sum \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i. \tag{7.59}$$

This has no simplification such as $\hat{\beta}_j = \hat{\mu}_j = \bar{y}_{\cdot j}$ of (7.4) and (7.23) in the balanced data case.

– iii. ***Predicting the random effects***

We have

$$\mathbf{D}\mathbf{Z}' = \sigma_u^2 \mathbf{I}_m \left\{ {}_{d}\, \mathbf{1}_{n_i}' \right\} = \sigma_u^2 \left\{ {}_{d}\, \mathbf{1}_{n_i}' \right\}$$

and so with $\mathbf{V}^{-1}$ of (7.57)

$$\begin{aligned} \tilde{\mathbf{u}}^0 &= \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \sigma_u^2 \left\{ {}_{d}\, \mathbf{1}_{n_i}' \right\} \left\{ {}_{d}\, \mathbf{V}_i^{-1} \right\} \left\{ {}_{c}\, \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} \right\} \\ &= \sigma_u^2 \left\{ {}_{c}\, \mathbf{1}_{n_i}'\mathbf{V}^{-1}\mathbf{y}_i - \mathbf{1}_{n_i}'\mathbf{V}_i^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}} \right\}. \end{aligned} \tag{7.60}$$

The nature of $\mathbf{X}_i$ discussed following (7.53), together with $\mathbf{V}_i$ not being the same for all i (not even of the same order), makes $\hat{\boldsymbol{\beta}}$ of (7.59) not very tractable; and so (7.60) is not amenable to further simplification. But for known $\mathbf{V}_i$, both it and (7.59) are reasonable computations.

c. Uncorrelated between, and within, subjects

– i. ***Matrix* V *and its inverse***

Using $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$ in (7.56), akin to $\mathbf{R}_0 = \sigma^2\mathbf{I}$ in Section 7.4b, gives

$$\mathbf{V} = \left\{ {}_{d}\, \mathbf{V}_i \right\} \quad \text{for} \quad \mathbf{V}_i = \sigma_u^2\mathbf{J}_{n_i} + \sigma^2\mathbf{I}_{n_i}$$

and

$$\mathbf{V}^{-1} = \left\{ {}_d\, \mathbf{V}_i^{-1} \right\} \quad \text{for} \quad \mathbf{V}_i^{-1} = \frac{1}{\sigma^2}\left(\mathbf{I}_{n_i} - \frac{\sigma_u^2}{\sigma^2 + n_i\sigma_u^2}\mathbf{J}_{n_i}\right).$$

– ii. ***Estimating the fixed effects***

On looking at the example of $\mathbf{X}_i$ in (7.51) we can see that $\mathbf{X}_i\mathbf{X}_i' = \mathbf{I}_{n_i}$ and $\mathbf{X}_i\mathbf{1}_n = \mathbf{1}_n$; but, unfortunately, these are not what are needed in

$$\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i = \frac{1}{\sigma^2}\left(\mathbf{X}_i'\mathbf{X}_i - \frac{\sigma_u^2}{\sigma^2 + n_i\sigma_u^2}\mathbf{X}_i'\mathbf{J}_{n_i}\mathbf{X}_i\right)$$

of order n. This appears to have no attractive simplification; nor does

$$\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i = \frac{1}{\sigma^2}\left(\mathbf{X}_i'\mathbf{y}_i - \frac{\sigma_u^2}{\sigma^2 + n_i\sigma_u^2}\mathbf{X}_i'\mathbf{1}_{n_i}\mathbf{1}_{n_i}'\mathbf{y}_i\right)$$

simplify beyond the following. Define n_{ij} as 1 if y_{ij} exists and 0 if y_{ij} does not exist; i.e., n_{ij} is the number of data on subject i at time j, either 0 or 1. Then

$$\begin{aligned}
\mathbf{X}_i'\mathbf{y}_i &= \left\{ {}_c\, n_{ij}y_{ij} \right\}_{j=1}^{n} \\
\mathbf{X}_i'\mathbf{1}_{n_i} &= \left\{ {}_c\, n_{ij} \right\}_{j=1}^{n} \\
\mathbf{1}_{n_i}'\mathbf{y}_i &= y_{i\cdot} \qquad (7.61) \\
\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i &= \frac{1}{\sigma^2}\left\{ {}_c\, n_{ij}\left[y_{ij} - \frac{\sigma_u^2}{\sigma^2 + n_i\sigma_u^2}y_{i\cdot}\right]\right\}_{j=1}^{n} \\
\hat{\beta} &= \left[\frac{1}{\sigma^2}\sum_{i=1}^{m}\left(\mathbf{X}_i'\mathbf{X}_i - \frac{\sigma_u^2}{\sigma^2 + n_i\sigma_u^2}\mathbf{X}_i'\mathbf{J}_{n_i}\mathbf{X}_i\right)\right]^{-1} \\
&\quad \times \sum_{i=1}^{m}\frac{1}{\sigma^2}\left\{ {}_c\, n_{ij}\left[y_{ij} - \frac{\sigma_u^2}{\sigma^2 + n_i\sigma_u^2}y_{i\cdot}\right]\right\}_{j=1}^{n}.
\end{aligned}$$

– iii. ***Predicting the random effects***

$$\mathbf{D}\mathbf{Z}' = \sigma_u^2\mathbf{I}_m\left\{ {}_d\, \mathbf{1}_{n_i}' \right\} = \sigma_u^2\left\{ {}_d\, \mathbf{1}_{n_i}' \right\}$$

$$\begin{aligned}
\mathbf{DZ'V^{-1}} &= \sigma_u^2 \left\{ {}_d\, \mathbf{1}'_{n_i} \right\} \frac{1}{\sigma^2} \left\{ {}_d \left(\mathbf{I}_{n_i} - \frac{\sigma_u^2}{\sigma^2 + n_i\sigma_u^2} \mathbf{J}_{n_i} \right) \right\} \\
&= \sigma_u^2 \left\{ {}_d\, \frac{\mathbf{1}'_{n_i}}{\sigma^2 + n_i\sigma_u^2} \right\} \qquad (7.62) \\
\mathbf{DZ'V^{-1}(y - X\hat{\beta})} &= \sigma_u^2 \left\{ {}_d\, \frac{\mathbf{1}'_{n_i}}{\sigma^2 + n_i\sigma_u^2} \right\} \left(\mathbf{y} - \left\{ {}_c\, \mathbf{X}_i\hat{\beta} \right\} \right) \\
&= \sigma_u^2 \left\{ {}_c\, \frac{n_i\bar{y}_{i\cdot}}{\sigma^2 + n_i\sigma_u^2} - \frac{\sum_{j=1}^m n_{ij}\hat{\beta}_j}{\sigma^2 + n_i\sigma_u^2} \right\} \\
\tilde{u}_i^0 &= \frac{n_i\sigma_u^2}{\sigma^2 + n_i\sigma_u^2} \left(\bar{y}_{i\cdot} - \frac{\Sigma_j n_{ij}\hat{\beta}_j}{n_i} \right).
\end{aligned}$$

d. Correlated between, but not within, subjects

Going back to (7.54) with $\mathbf{R} = \left\{ {}_d\, \sigma^2\mathbf{I}_{n_i} \right\}$ we have

$$\mathbf{V} = \left\{ {}_d\, \sigma_u^2(1 - \rho_u)\mathbf{J}_{n_i} + \sigma_u^2\mathbf{I}_{n_i} \right\} + \sigma_u^2\rho_u\mathbf{J}_N,$$

which is the analogue for unbalanced data of the $\mathbf{V}$ of the beginning of Section 7.4d and 7.5e (which are for balanced data); see E 7.7. We know of no way to invert this.

7.7 AN EXAMPLE OF SEVERAL TREATMENTS

Instead of having just m subjects (or patients) as described in Section 7.1, suppose we had m subjects in each of T treatments, a total of Tm patients. Then take y_{tij} as the datum of the ith person in the tth treatment at time j. Then for

$$\mathbf{y} = \left\{ {}_c \left\{ {}_c \left\{ {}_c\, y_{tij} \right\}_{j=1}^{n} \right\}_{i=1}^{m} \right\}_{t=1}^{T}$$

with u_{ti} being the (random) effect for that person. We have, instead of (7.9)

$$\mathrm{E}\left[y_{tij} | u_{ti}\right] = \mu + \tau_t + \alpha_j + u_{ti} \qquad (7.63)$$

where τ_t and α_j are the fixed effects for the tth treatment and jth time, respectively. From this we get

$$\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \quad \text{for} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \{_c\, \tau_t\}_{t=1}^{T} \\ \{_c\, \alpha_j\}_{j=1}^{n} \end{bmatrix}$$

with $\mathbf{X}$ being the partitioned matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{Tmn} & \mathbf{I}_T \otimes \mathbf{1}_{mn} & \mathbf{1}_{Tm} \otimes \mathbf{I}_n \end{bmatrix} \tag{7.64}$$

and

$$\mathbf{Z} = \mathbf{I}_{Tm} \otimes \mathbf{1}_n.$$

On assuming that data on different patients are independent we take

$$\mathbf{V} = \mathrm{var}(\mathbf{y}) = \mathbf{I}_{Tm} \otimes \mathbf{V}_0.$$

Then, for the theorem mentioned in Section 7.4d we find that

$$\begin{aligned} \mathbf{VX} &= (\mathbf{I}_T \otimes \mathbf{I}_m \otimes \mathbf{V}_0) \\ &\quad \times \left[(\mathbf{1}_T \otimes \mathbf{1}_m \otimes \mathbf{1}_n) \quad (\mathbf{I}_T \otimes \mathbf{1}_m \otimes \mathbf{1}_n) \quad (\mathbf{1}_T \otimes \mathbf{1}_m \otimes \mathbf{I}_n)\right] \\ &= \left[(\mathbf{1}_T \otimes \mathbf{1}_m \otimes \mathbf{V}_0\mathbf{1}_n) \quad (\mathbf{I}_T \otimes \mathbf{1}_m \otimes \mathbf{V}_0\mathbf{1}_n) \quad (\mathbf{1}_T \otimes \mathbf{1}_m \otimes \mathbf{V}_0)\right] \end{aligned}$$

and so there is no $\mathbf{H}$ for which $\mathbf{VX} = \mathbf{XH}$ for a general $\mathbf{V}_0$. But if

$$\mathbf{V}_0 = \sigma^2[(1-\rho)\mathbf{I}_n + \rho\mathbf{J}_n] \tag{7.65}$$

then for $\lambda = \sigma^2[(1-\rho) + np]$

$$\mathbf{V}_0\mathbf{1}_n = \lambda\mathbf{1}_n$$

and for

$$\mathbf{H} = \begin{bmatrix} \lambda\mathbf{I}_{T+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_0 \end{bmatrix}$$

it will be found that $\mathbf{VX} = \mathbf{XH}$. Therefore for $\mathbf{X}$ of (7.64) and $\mathbf{V}$ based on (7.65) the estimators of μ, τ_t and α_j will be exactly the same, for balanced data, as in the simple additive fixed effects model

$\mathrm{E}[y_{tj}] = \mu + \tau_t + \alpha_j$; for which one such set of estimators is well known to be

$$\begin{aligned}\hat{\mu} &= \bar{y}_{...} \\ \hat{\tau}_t &= \bar{y}_{t\cdot\cdot} - \bar{y}_{...} \\ \hat{\alpha}_j &= \bar{y}_{\cdot\cdot j} - \bar{y}_{...}\,.\end{aligned}$$

For deriving $\tilde{u}_{ti}$ we have $\mathbf{D} = \sigma_u^2 \mathbf{I}_{Tm}$ and $\mathbf{Z} = \mathbf{I}_{Tm} \otimes \mathbf{1}_n$, so that in

$$\begin{aligned}\tilde{\mathbf{u}}^0 &= \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \\ \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1} &= \sigma_u^2(\mathbf{I}_{Tm} \otimes \mathbf{1}_n')[\mathbf{I}_{Tm} \otimes \{(1-\rho)\mathbf{I}_n + \rho\mathbf{I}_n\}]^{-1}\frac{1}{\sigma^2} \\ &= (\sigma_u^2/\sigma^2)\left[\mathbf{I}_{Tm} \otimes \frac{1}{1-\rho}\mathbf{1}_n'\left(\mathbf{I}_n - \frac{\rho}{1-\rho+np}\mathbf{J}_n\right)\right] \\ &= \frac{\sigma_u^2}{\sigma^2(1-\rho+np)}\left(\mathbf{I}_{Tm} \otimes \mathbf{1}_m'\right).\end{aligned}$$

This leads to

$$\begin{aligned}\tilde{u}_{ti}^0 &= \frac{\sigma_u^2}{\sigma^2(1-\rho+np)}(\bar{y}_{ti\cdot} - \hat{\mu} - \hat{\tau}_t - \bar{\hat{\alpha}}_\cdot) \\ &= \frac{\sigma_u^2}{\sigma^2(1-\rho+np)}(\bar{y}_{ti\cdot} - \bar{y}_{t\cdot\cdot}).\end{aligned}$$

7.8 GENERALIZED ESTIMATING EQUATIONS

If we define balanced data as the case where $\mathbf{V}\mathbf{X} = \mathbf{X}\mathbf{H}$ for some $\mathbf{H}$ (some justification of so defining it is given via examples in the exercises) then the ordinary least squares estimator of $\boldsymbol{\beta}$ is equal to the generalized least squares estimator which is the same as the maximum likelihood estimator. What would happen if the ordinary least squares estimator were used for unbalanced data? For this section, for ease of exposition we will assume that $\mathbf{X}$ is of full column rank, and use the notation $\hat{\boldsymbol{\beta}}_{\mathbf{W}^{-1}}$ to denote the weighted least squares estimator $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$. Thus $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$ denotes the ordinary least squares estimator and $\hat{\boldsymbol{\beta}}_{\mathbf{V}}$ denotes the MLE with a known variance-covariance matrix $\mathbf{V} = \mathrm{var}(\mathbf{y})$, i.e., $\hat{\boldsymbol{\beta}}_{\mathbf{V}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

If primary interest lies in estimating $\boldsymbol{\beta}$, a possible estimator is $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$. How well does it perform? First, note that $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$ is unbiased no matter

what the value of $\mathbf{V}$:

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}}_{\mathbf{I}}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{7.66}$$

Furthermore, it is straightforward to calculate the variance of $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$:

$$\begin{aligned} \mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathbf{I}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{7.67}$$

How does this compare with $\hat{\boldsymbol{\beta}}_{\mathbf{V}}$, which is the optimal estimator for known $\mathbf{V}$? We know that $\hat{\boldsymbol{\beta}}_{\mathbf{V}}$ is also unbiased with variance $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ which is smaller than the variance of $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$, but how much smaller? Interestingly, it is often not very much smaller.

For example, consider the extremely simple situation where the mean of all the observations is μ and the data come in m equicorrelated clusters with correlation $\rho > 0$. Further assume that half the clusters are of size n and the other half are of size λn. We thus have

$$\mathbf{y} \sim (\mathbf{1}\mu, \mathbf{V}),$$

where

$$\mathbf{V} = \begin{bmatrix} \mathbf{I}_{m/2} \otimes \mathbf{V}_{0,n} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m/2} \otimes \mathbf{V}_{0,\lambda n} \end{bmatrix} \tag{7.68}$$

and

$$\mathbf{V}_{0,n} = \sigma^2 \left[(1-\rho)\mathbf{I}_n + \rho\mathbf{J}_n\right]. \tag{7.69}$$

It is straightforward to derive an expression (see E 7.10) for the variance of $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{V}}$ and to show that the relative efficiency of the ordinary least squares estimator decreases as a function of increasing ρ. Furthermore, the relative efficiency in the worst case, when $\rho = 1$, is given by $2(1+\lambda^2)/(1+\lambda)^2$. So if the data are balanced with $\lambda = 1$, the relative efficiency is 1, as expected. If the sample size is 50% larger in one group than in the other, then the worst the variance of ordinary least squares estimator can be is 4% larger than the optimal estimator. Even in the case when the sample size is double ($\lambda = 2$) the variance is only $2(5)/9 = 1.1$ times as large (see E 7.10 and E 7.11).

Certainly it is possible to construct examples where the ordinary least squares estimator does arbitrarily badly in relation to the optimal estimator, but the point of the above calculation is that often the ordinary least squares estimator is quite good for moderate degrees of unbalancedness. If the ordinary least squares estimator performs so well for a wide variety of problems and, furthermore, obviates the need to estimate the variance-covariance structure, then why not use standard regression and analysis of variance software whenever the efficiency is high?

The problem is that even though the efficiency is high, the apparent variance can be grossly incorrect. Intuitively, if the observations are positively correlated and we treat them as if they are independent, then the data appear much less variable than they actually are and we can drastically underestimate the variance. This would hold true whether or not the data were balanced.

Again consider a clustered variance scenario with

$$\mathbf{y}_i \sim \text{ i.i.d. } (\mathbf{1}\beta, \mathbf{V}_{0,n}) \tag{7.70}$$

so that

$$\mathbf{V} = \mathbf{I}_m \otimes \mathbf{V}_{0,n}, \tag{7.71}$$

where $\mathbf{V}_{0,n}$ is defined in (7.69) and the (scalar) β is the mean. The variance that would be estimated from a standard regression or ANOVA program would be

$$\begin{aligned} \widehat{\text{var}}(\hat{\beta}_{\mathbf{I}}) &= \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \hat{\sigma}^2/mn \\ &= \hat{\sigma}^2/N \end{aligned} \tag{7.72}$$

with

$$\hat{\sigma}^2 = \mathbf{y}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}/[N - r(\mathbf{X})] \tag{7.73}$$

and using $N = mn$.

On average $\widehat{\text{var}}(\hat{\beta}_{\mathbf{I}})$ estimates

$$\begin{aligned} \mathrm{E}[\widehat{\text{var}}(\hat{\beta}_{\mathbf{I}})] &= \mathrm{E}[\hat{\sigma}^2]/N \\ &= \frac{1}{N - r(\mathbf{X})}\text{tr}[\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{V}]/N \\ &= \frac{1}{N-1}\text{tr}[\mathbf{V} - \frac{1}{N}\mathbf{1}\mathbf{1}'\mathbf{V}]/N \end{aligned}$$

$$= \frac{1}{N-1}\sigma^2[N-(1+\{n-1\}\rho)]/N$$
$$= \frac{\sigma^2}{N}\left[1-\frac{(n-1)\rho}{N-1}\right] < \frac{\sigma^2}{N}. \tag{7.74}$$

On the other hand, the true variance of $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$ is

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}_{\mathbf{I}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{1}{N}\mathbf{1}_N'\mathbf{V}\mathbf{1}_N\frac{1}{N} \\ &= \frac{\sigma^2}{N}[1+(n-1)\rho] > \frac{\sigma^2}{N}. \end{aligned} \tag{7.75}$$

Thus the estimated variance averages less than σ^2/N, the variance if all the observations were independent, while the true variance is larger than σ^2/N; and it can be substantially so. If $\rho = 0.5$, a small to moderate correlation, and if $m = 5$ and $n = 11$, then the underestimation is by a factor of $(1+5)/(1-10[0.5]/54) \doteq 6.6$!

The remedy is relatively straightforward; use $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$ but correctly assess its variance. For example, with model (7.70), we estimate $\hat{\beta} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \bar{y}_{\cdot\cdot}$. Then

$$\hat{\mathbf{V}}_0 = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{1}_n\hat{\mu})(\mathbf{y}_i - \mathbf{1}_n\hat{\mu})' \tag{7.76}$$

is a consistent estimator (see E 7.13) of $\text{var}(\mathbf{y}_i)$, no matter what its form. When the mean structure takes a more complicated form, namely $\text{E}[\mathbf{y}_i] = \mathbf{X}_i\boldsymbol{\beta}$, we would generalize accordingly:

$$\hat{\mathbf{V}}_0 = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})'. \tag{7.77}$$

This is the basic idea behind *generalized estimating equations* (GEEs), which are developed more fully in Chapter 8. Another aspect of GEEs is the specification of a *working* variance-covariance structure. That is, in some cases we might suspect that a specific form of covariance might hold for $\mathbf{y}_i$ of (7.70). Also, the use of the independence assumption can lead to inefficient estimators (Fitzmaurice, 1995). In such cases it is straightforward to make adjustments. Let $\mathbf{W}$ represent the *inverse* of the assumed variance-covariance structure of each of the $\mathbf{y}_i$s:

$$\mathbf{W} = [\text{var}_W(\mathbf{y}_i)]^{-1}, \tag{7.78}$$

where $\text{var}_W(\cdot)$ denotes a temporarily assumed variance-covariance structure. If we believed this structure, our estimator of β would be $\hat{\beta}_{\mathbf{W}^{-1}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ with variance

$$\text{var}(\hat{\beta}_{\mathbf{W}^{-1}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}, \tag{7.79}$$

which is sometimes called the *sandwich variance formula* since $\mathbf{X}'\mathbf{V}\mathbf{X}$ is "sandwiched" between two $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$s. As before, a consistent estimator of $\mathbf{V}$ is formed from the independent "replicates," $\mathbf{y}_i$. For more details see Diggle et al. (1994, Sec. 4.6).

7.9 A SUMMARY OF RESULTS

The results developed in Sections 7.4 and 7.5 are sequenced by model within each estimation situation. Here we give a summary of those results and those of Section 7.2 sequenced by estimation situation within each model.

Note: Equation numbers in square brackets refer to equations occurring earlier in the chapter.

a. Balanced data

– i. *With some generality*

$$\mathbf{y} \sim \mathcal{N}\left[(\mathbf{1}_m \otimes \mathbf{I}_n)\mu, \mathbf{V} = \mathbf{I}_m \otimes \mathbf{V}_0\right],$$

and

$$\mu = \left\{{}_c\,\mu_j\right\}_{j=1}^{n}$$

the MLE of μ_j is

$$\hat{\mu}_j = \bar{y}_{\cdot j} \tag{7.4}$$

and that of $\mathbf{V}_0$ is

$$\hat{\mathbf{V}}_0 = \frac{1}{m}\left\{\sum_{i=1}^{m}(y_{ij} - \bar{y}_{\cdot j})(y_{ik} - \bar{y}_{\cdot k})\right\}_{j,k=1}^{n}, \tag{7.8}$$

a Wishart matrix. The estimator $\hat{\mu}_j = \bar{y}_{\cdot j}$ applies for the special cases of balanced data we now list; it is not repeated, and it is denoted $\hat{\beta}$.

– ii. ***Uncorrelated subjects***

With $\mathbf{R}_0$ defined just prior to (7.19)

$$\tilde{u}_i^0 = \frac{\mathbf{1}'\mathbf{R}_0^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}})}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}} \qquad [7.23]$$

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^m \mathbf{1}'\mathbf{R}_0^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}) - m'\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}{m(\mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1})^2}. \qquad [7.38]$$

– iii. ***Uncorrelated between, and within, subjects***

$$\tilde{u}_i^0 = \frac{n\sigma_u^2}{\sigma^2 + n\sigma_u^2}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \qquad [7.24]$$

$$\dot{\sigma}^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2}{m(n-1)} \qquad [7.43]$$

$$\dot{\sigma}_u^2 = \frac{1}{n}\left[\left\{\sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right\}/m - \dot{\sigma}^2\right]. \qquad [7.44]$$

– iv. ***Uncorrelated between, and autocorrelated within, subjects***

$$\tilde{u}_i^0 = \frac{\sigma_u^2[(1-\rho)n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + \rho(y_{i1} - \bar{y}_{\cdot 1} + y_{in} - \bar{y}_{\cdot n}]}{\sigma^2(1+\rho) + \sigma_u^2[n - (n-2)\rho]}. \qquad [7.27]$$

For $\delta_{ij} \equiv y_{ij} - \bar{y}_{\cdot j}$

$$mn\hat{\sigma}^2(1-\hat{\rho}^2) = (1+\hat{\rho}^2)\sum_i \sum_j \delta_{ij}^2 - \hat{\rho}^2 \sum_{i=1}^m (\delta_{i1}^2 + \delta_{in}^2)$$
$$-2\hat{\rho}\sum_{i=1}^m \sum_{j=2}^n \delta_{ij}\delta_{i,j-1} \qquad [7.49]$$

$$m(n-1)\hat{\rho}\hat{\sigma}^2(1-\hat{\rho}^2) = 2\hat{\rho}\sum_i \sum_j \delta_{ij}^2 - \hat{\rho}\sum_i (\delta_{i1}^2 + \delta_{in}^2)$$
$$-(1+\hat{\rho}^2)\sum_{i=1}^m \sum_{j=2}^n \delta_{ii}\delta_{i,j-1}. \qquad [7.50]$$

– v. ***Correlated between, but not within, subjects***

$$\tilde{u}_i^0 = \frac{n\sigma_u^2(1-\rho_u)}{\sigma^2 + n\sigma_u^2(1-\rho_u)}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \,. \qquad [7.30]$$

Estimators for σ_u^2, ρ_u and σ^2: no results (see the end of Section 7.5).

b. Unbalanced data

– i. ***Uncorrelated subjects***

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{m} (\mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \sum_{i=1}^{m} \mathbf{V}_i^{-1} \mathbf{y}_i \qquad [7.59]$$

$$\tilde{u}_i^0 = \sigma_u^2 \left(\mathbf{1}_{n_i}' \mathbf{V}_i^{-1} \mathbf{y}_i - \mathbf{1}_{n_i}' \mathbf{V}_i^{-1} \mathbf{X}_i \hat{\boldsymbol{\beta}} \right) . \qquad [7.60]$$

– ii. ***Uncorrelated between, and within, subjects***

Special case: $n_{ij} = 1$ if y_{ij} exists
$= 0$ if y_{ij} does not exist.

$$\hat{\boldsymbol{\beta}} = \left[\sum_i \left(\mathbf{X}_i' \mathbf{X}_i - \frac{\sigma_u^2}{\sigma^2 + n_i \sigma_u^2} \mathbf{X}_i' \mathbf{J}_{n_i} \mathbf{X}_i \right) \right]^{-1} \times \sum_{i=1}^{m} \left\{ {}_c\, n_{ij} \left(y_{ij} - \frac{\sigma_u^2}{\sigma^2 + n_i \sigma_u^2} y_{i\cdot} \right) \right\} \qquad [7.61]$$

$$\tilde{u}_i^0 = \frac{n_i \sigma_u^2}{\sigma^2 + n_i \sigma_u^2} \left(\bar{y}_{i\cdot} - \frac{\sum_i n_{ij} \hat{\beta}_j}{\sum_j n_{ij}} \right) . \qquad [7.62]$$

– iii. ***Correlated between, but not within, subjects***

No results could be obtained (see Section 7.6d).

7.10 APPENDIX

a. For Section 7.4a

In $\mathbf{V} = \mathbf{I}_m \otimes (\sigma_u^2 \mathbf{J}_n + \mathbf{R}_0)$ write $\sigma_u^2 \mathbf{J}_n = \sigma_u^2 \mathbf{1}_n \mathbf{1}_n'$ and use the standard result for any nonsingular $\mathbf{A}$:

$$(\mathbf{A} + \lambda \mathbf{t}\mathbf{t}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{t}\mathbf{t}'\mathbf{A}^{-1}}{1/\lambda + \mathbf{t}'\mathbf{A}^{-1}\mathbf{t}}$$

to get

$$(\mathbf{R}_0 + \sigma_u^2 \mathbf{1}\mathbf{1}')^{-1} = \mathbf{R}_0^{-1} - \frac{\mathbf{R}_0^{-1}\mathbf{1}\mathbf{1}'\mathbf{R}_0^{-1}}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}.$$

Then, since $\rho_u = 0$ gives $\mathbf{D} = \sigma_u^2 \mathbf{I}_m$, with $\mathbf{Z} = \mathbf{I}_m \otimes \mathbf{1}_n$ we get for $\tilde{\mathbf{u}}$

$$\begin{aligned}
\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1} &= \sigma_u^2 \mathbf{I}_m (\mathbf{I}_m \otimes \mathbf{1}_n')\mathbf{V}^{-1} = \sigma_u^2 (\mathbf{I}_m \otimes \mathbf{1}_n')\mathbf{V}^{-1} \\
&= \sigma_u^2 \mathbf{I}_m \otimes \left(\mathbf{1}_n' \left[\mathbf{R}_0^{-1} - \frac{\mathbf{R}_0^{-1}\mathbf{1}\mathbf{1}'\mathbf{R}_0^{-1}}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}} \right] \right) \\
&= \mathbf{I}_m \otimes \frac{\mathbf{1}'\mathbf{R}_0^{-1}}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}} = \left\{ {}_d \frac{\mathbf{1}'\mathbf{R}_0^{-1}}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}} \right\}_{i=1}^{m}. \\
\tilde{\mathbf{u}}^0 &= \left\{ {}_d \frac{\mathbf{1}'\mathbf{R}_0^{-1}}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}} \right\}_{i=1}^{m} \left\{ {}_c \, \mathbf{y}_i - \hat{\boldsymbol{\beta}} \right\}_{i=1}^{m}. \\
\tilde{u}_i^0 &= \frac{\mathbf{1}'\mathbf{R}_0^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}})}{1/\sigma_u^2 + \mathbf{1}'\mathbf{R}_0^{-1}\mathbf{1}}. \qquad [7.23]
\end{aligned}$$

b. For Section 7.4b

Put $\mathbf{R}_0^{-1} = (1/\sigma^2)\mathbf{I}_n$ in (7.23), to get

$$\tilde{u}_i^0 = \frac{(1/\sigma^2)\left(y_{i\cdot} - \sum_{j=1}^{n} \bar{y}_{\cdot j}\right)}{1/\sigma_u^2 + n/\sigma^2} = \frac{n\sigma_u^2(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}{\sigma^2 + n\sigma_u^2}.$$

c. For Section 7.4d

$$\begin{aligned}
\mathbf{V} &= \sigma_u^2 \left[(1 - \rho_u)\mathbf{I}_m + \rho_u \mathbf{J}_m \right] \otimes \mathbf{J}_n + \sigma^2 (\mathbf{I}_m \otimes \mathbf{I}_n) \\
\mathbf{M} &= \left[\theta \mathbf{A} \otimes \mathbf{J}_n + \lambda (\mathbf{I}_m \otimes \mathbf{I}_n) \right] \text{ is a generalization.}
\end{aligned}$$

Suppose $\mathbf{M}^* = \mathbf{B} \otimes \mathbf{J}_n + (1/\lambda)(\mathbf{I}_m + \mathbf{I}_n)$ is $\mathbf{M}^{-1}$. Then in $\mathbf{M}\mathbf{M}^* = \mathbf{I}$ we want the coefficient of $\mathbf{J}_n$ to be $\mathbf{0}$. Thus

$$\begin{aligned} 0 &= n\theta\mathbf{A}\mathbf{B} + (\theta/\lambda)\mathbf{A}) + \lambda\mathbf{B} \\ \Rightarrow \mathbf{B} &= -(n\theta\mathbf{A} + \lambda\mathbf{I})^{-1}\theta\mathbf{A}/\lambda. \end{aligned}$$

Therefore

$$[\theta\mathbf{A}\otimes\mathbf{J}_n+\lambda(\mathbf{I}_m\otimes\mathbf{I}_n)]^{-1} = -\left[(n\theta\mathbf{A} + \lambda\mathbf{I})^{-1}\theta\mathbf{A}/\lambda\right]\otimes\mathbf{J}_n+(1/\lambda)(\mathbf{I}_m\otimes\mathbf{I}_n)$$

and so for $\mathbf{V}$

$$\theta = \sigma_u^2, \quad \mathbf{A} = (1-\rho_u)\mathbf{I}_m + \rho_u\mathbf{J}_m \quad \text{and} \quad \lambda = \sigma^2.$$

Therefore

$$\begin{aligned} \mathbf{V}^{-1} &= -\left\{n\sigma_u^2\left[(1-\rho_u)\mathbf{I}_m + \rho_u\mathbf{J}_m\right] + \sigma^2\mathbf{I}_m\right\}^{-1} \\ &\qquad \times \theta\mathbf{A}/\lambda \otimes \mathbf{J}_n + (1/\lambda)(\mathbf{I}_m \otimes \mathbf{I}_n) \\ &= -\Bigg(\left\{\left[n\sigma_u^2(1-\rho_u) + \sigma^2\right]\mathbf{I}_m + n\sigma_u^2\rho_u\mathbf{J}_m\right\}^{-1} \\ &\qquad \theta\mathbf{A}/\lambda\Bigg) \otimes \mathbf{J}_n + \tfrac{1}{\lambda}(\mathbf{I}_m \otimes \mathbf{I}_n) \\ &= \left\{\frac{-1}{n\sigma_u^2(1-\rho_u)+\sigma^2}\left[\mathbf{I}_m - \frac{n\sigma_u^2\rho_u}{n\sigma_u^2(1-\rho_u)+\sigma^2+mn\sigma_u^2\rho_u}\mathbf{J}_m\right]\right. \\ &\qquad \left.\times \frac{\theta\mathbf{A}}{\lambda}\right\} \otimes \mathbf{J}_n + \tfrac{1}{\lambda}(\mathbf{I}_m \otimes \mathbf{I}_n) \\ &= \frac{-\sigma_u^2/\sigma^2}{n\sigma_u^2(1-\rho_u)+\sigma^2}\Bigg[(1-\rho_u)\mathbf{I}_m + \rho_u\mathbf{J}_m \\ &\qquad - \frac{n\sigma_u^2\rho_u\mathbf{J}_m\{(1-\rho_u)\mathbf{I}_m + \rho_u\mathbf{J}_m\}}{n\sigma_u^2(1-\rho_u)+\sigma^2+mn\sigma_u^2\rho_u}\Bigg] \otimes \mathbf{J}_n + \frac{\mathbf{I}_{mn}}{\sigma^2} \\ &= \frac{-\sigma_u^2/\sigma^2}{n\sigma_u^2(1-\rho_u)+\sigma^2}\left[(1-\rho_u)\mathbf{I}_m + \frac{\rho_u\sigma^2\mathbf{J}_m}{\sigma^2+n\sigma_u^2+(m-1)n\rho_u\sigma_u^2}\right] \\ &\qquad \otimes \mathbf{J}_n + \frac{\mathbf{I}_{mn}}{\sigma^2}. \end{aligned}$$

Then with $\mathbf{Z} = \mathbf{I}_m \otimes \mathbf{1}_n$ and $\mathbf{D}$ of (7.17)

$$\mathbf{D}\mathbf{Z}' = \sigma_u^2\left[(1-\rho_u)\mathbf{I}_m + \rho_u\mathbf{J}_m\right]\left[\mathbf{I}_m \otimes \mathbf{1}_n'\right]$$

$$= \sigma_u^2 [(1-\rho_u)\mathbf{I}_m + \rho_u \mathbf{J}_m] \otimes \mathbf{1}_n',$$

$$\mathbf{DZ'V}^{-1} = \frac{(-\sigma_u^2/\sigma^2)\sigma_u^2}{n\sigma_u^2(1-\rho_u)+\sigma^2}\left[(1-\rho_u)^2\mathbf{I}_m + \rho_u(1-\rho_u)\mathbf{J}_m + \frac{\{(1-\rho_u)\rho_u\sigma^2 + m\rho_u^2\sigma^2\}\mathbf{J}_m}{\sigma^2 + n\sigma_u^2 + (m-1)n\rho_u\sigma_u^2}\right] \otimes n\mathbf{1}_n' + \frac{\sigma_u^2}{\sigma^2}[(1-\rho_u)\mathbf{I}_m + \rho_u\mathbf{J}_m] \otimes \mathbf{1}_n'$$

$$= \frac{\sigma_u^2}{\sigma^2}(1-\rho_u)\left[1 - \frac{(1-\rho_u)n\sigma_u^2}{n\sigma_u^2(1-\rho_u)+\sigma^2}\right]\mathbf{I}_m \otimes \mathbf{1}_n' + \frac{\sigma_u^2}{\sigma^2}\rho_u\left[1 - \frac{n\sigma_u^2}{n\sigma_u^2(1-\rho_u)+\sigma^2} \times \left\{1 - \rho_u + \frac{\sigma^2[1+(m-1)\rho_u]}{\sigma^2 + n\sigma_u^2[1+(m-1)\rho_u]}\right\}\right]\mathbf{J}_m \otimes \mathbf{1}_n'.$$

Write $\tau \equiv 1 + (m-1)\rho_u$ and then

$$\mathbf{DZ'V}^{-1} = \frac{\sigma_u^2(1-\rho_u)}{n\sigma_u^2(1-\rho_u)+\sigma^2}\mathbf{I}_m \otimes \mathbf{1}_n' + \frac{\sigma_u^2\rho_u}{\sigma^2} \times \left[\frac{\alpha-\beta}{\gamma}\right]\mathbf{J}_m \otimes \mathbf{1}_n'$$

where

$$\alpha = [n\sigma_u^2(1-\rho_u)+\sigma^2][\sigma^2 + n\sigma_u^2\tau],$$
$$\beta = n\sigma_u^2[(1-\rho_u)(\sigma^2 + n\sigma_u^2\tau) + \sigma^2\tau],$$

and

$$\gamma = [n\sigma_u^2(1-\rho_u)+\sigma^2][\sigma^2 + n\sigma_u^2\tau].$$

Then $\alpha - \beta$ reduces very simply to σ^4, so that

$$\mathbf{DZ'V}^{-1} = \frac{\sigma_u^2}{n\sigma_u^2(1-\rho_u)+\sigma^2}\left[(1-\rho_u)\mathbf{I}_m + \frac{\sigma^2\rho_u}{\sigma^2 + n\sigma_u^2[1+(m-1)\rho_u]}\mathbf{J}_m\right] \otimes \mathbf{1}_n'.$$

Now

$$\mathbf{I}_m \otimes \mathbf{1}_n' = \left\{{}_d\,\mathbf{1}_n'\right\}_{i=1}^{m} \quad \text{and} \quad \mathbf{J}_m \otimes \mathbf{1}_n' = \mathbf{J}_{m\times mn}.$$

Also

$$\begin{aligned}(\mathbf{J}_{m\times mn})(\mathbf{1}_m \otimes \mathbf{I}_n) &= (\mathbf{J}_{m\times m} \otimes \mathbf{1}_n')(\mathbf{1}_m \otimes \mathbf{I}_n) \\ &= m\mathbf{1}_m \otimes \mathbf{1}_n' \\ &= m\mathbf{J}_{mn}.\end{aligned}$$

Using these in post-multiplying $\mathbf{DZ'V}^{-1}$ by $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ where $\mathbf{X} = \mathbf{1}_m \otimes \mathbf{I}_n$ of (7.12), and where $\hat{\boldsymbol{\beta}} = \left\{ {}_c\, \bar{y}_{\cdot j} \right\}$ as mentioned at the end of the paragraph preceding (7.30) gives

$$\begin{aligned}\tilde{u}_i^0 &= \frac{\sigma_u^2}{n\sigma_u^2(1-\rho_u)+\sigma^2}\left[(1-\rho_u)\Big(y_{i\cdot} - \sum_{j=1}^{n}\bar{y}_{\cdot j}\Big)\right. \\ &\qquad \left. + \frac{\sigma_u^2\rho_u}{\sigma^2 + n\sigma_u^2[1+(m-1)\rho_u]}(y_{\cdot\cdot} - m\sum_j \bar{y}_{\cdot j})\right] \\ &= \frac{n\sigma_u^2(1-\rho_u)}{n\sigma_u^2(1-\rho_u)+\sigma^2}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\,, \qquad [7.30]\end{aligned}$$

since $y_{\cdot\cdot} - m\sum_{j=1}^{n}\bar{y}_{\cdot j} = mn(\bar{y}_{\cdot\cdot} - \bar{y}_{\cdot\cdot}) = 0$.

7.11 EXERCISES

E 7.1 Reduce ML($\boldsymbol{\mu}$) in Section 7.2b to $\hat{\boldsymbol{\mu}}$ of (7.4).

E 7.2 If $\mathbf{H}^{-1}$ exists and for $\mathbf{X}$ of full column rank, show that $\mathbf{VX} = \mathbf{XH}$ leads to $(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{y} = (\mathbf{X'X})^{-1}\mathbf{X'y}$.

E 7.3 In Section 7.4c show that $\mathbf{AA}^{-1} = \mathbf{I}$ and derive (7.27).

E 7.4 In Section 7.5b confirm the final form of LHS(σ_u^2) and RHS(σ_u^2). Also, verify the simplification of (7.38) to (7.39).

E 7.5 Results (7.43) and (7.44) are said (at the end of Section 7.5c) to be the same as in Searle et al.(1992, p.150). Those results are

$$\dot{\sigma}_e^2 = \left[1 - \frac{a-1}{b(an-1)}\right]\text{MSE} \quad \text{and} \quad \dot{\sigma}_\beta^2 = \frac{\text{SSB}/b - \dot{\sigma}_e^2}{an}. \qquad (7.80)$$

At first sight these are somewhat different from (7.43) and (7.44). Reconcile the two sets of results. *Hint*: Do not match the two methods by the criterion of rows-and-columns layout, but match them by random effects.

E 7.6 Derive (7.49) and (7.50) from their respective LHS(·) = RHS(·) equations.

E 7.7 Explain why $\mathbf{V}$ of Section 7.6d is the unbalanced data analogue of (7.28).

E 7.8 In Section 7.7 show that $\mathbf{VX} = \mathbf{XH}$.

E 7.9 For the setting of Section 7.7, explain why $\mathrm{E}[y_{tij}|u_{ti}] = \theta_{tj}$ leads to $\hat{\theta}_{tj} = \bar{y}_{t\cdot j}$. Describe how $\hat{\theta}_{tj}$ leads to $\hat{\mu}$, $\hat{\tau}$ and $\hat{\alpha}_j$ of that Section.

E 7.10 Derive the relative efficiency of $\hat{\boldsymbol{\beta}}_{\mathbf{I}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{V}}$ as described in the paragraph immediately following (7.69).

E 7.11 Again following the discussion after (7.69), show that the relative efficiency of the ordinary least squares estimator decreases as a function of increasing ρ. Furthermore, show that the relative efficiency in the worst case, when $\rho = 1$, is given by $2(1+\lambda^2)/(1+\lambda)^2$.

E 7.12 Show the calculation of $\mathrm{tr}(\mathbf{V} - \mathbf{1}\mathbf{1}'\mathbf{V}/N)$ involved in deriving (7.74).

E 7.13 Show that $\hat{\mathbf{V}}_0$ of (7.76) is consistent for $\mathbf{V}_0$ as m tends to ∞.

E 7.14 In E 6.6(c) write

$$\mathbf{y} = \left\{ _c \left\{ _c \left\{ _c \, y_{ijk} \right\}_{k=1}^{r} \right\}_{j=1}^{n} \right\}_{i=1}^{m}.$$

Then

$$\begin{aligned} \mathbf{V} &= \mathrm{var}(\mathbf{y}) \\ &= (\mathbf{I_m} \otimes \mathbf{J}_n \otimes \mathbf{J}_r)\sigma_a^2 + (\mathbf{I_m} \otimes \mathbf{I}_n \otimes \mathbf{J}_r)\sigma_g^2 + (\mathbf{I_m} \otimes \mathbf{I}_n \otimes \mathbf{I}_r)\sigma^2 \end{aligned}$$

and

$$\mathbf{X} = [(\mathbf{1_m} \otimes \mathbf{1}_n \otimes \mathbf{1}_r) \ (\mathbf{1_m} \otimes \mathbf{I}_n \otimes \mathbf{I}_r) \ (\mathbf{1_m} \otimes \mathbf{1}_n \otimes \mathbf{I}_r)].$$

(a) For $m = 2$, $n = 3$, and $r = 2$ write out $\mathbf{y}$, $\mathbf{V}$, and $\mathbf{X}$.

(b) Find $\mathbf{H}$ such that $\mathbf{VX} = \mathbf{XH}$.

(c) With $r = 1$ (i.e., effectively deleting k from $\mathbf{y}$), what are $\mathbf{V}$ and $\mathbf{X}$ for E 6.6(a)? And what is $\mathbf{H}$ such that $\mathbf{VX} = \mathbf{XH}$?

(d) Repeat (c) for E 6.6(b).

Chapter 8

GENERALIZED LINEAR MIXED MODELS (GLMMs)

8.1 INTRODUCTION

The use of random factors is not restricted to linear mixed models, the topic of Chapter 6. For many of the same reasons as seen there, we may want to incorporate random factors into nonlinear models. That is, we may wish to build a model that accommodates correlated data, or to consider the levels of a factor as selected from a population of levels in order to make inference to that population.

For example, suppose we wish to study factors affecting cost of hospitalization by taking a random sample of patient records from each of 15 teaching hospitals. The costs within a hospital almost certainly must be regarded as correlated. They will be similar because of the general costs of running the hospital, billing practices, costs of nearby, competing hospitals, and so on. Also, a goal may be to make inferences to a larger population of research hospitals. Both of these could be accommodated by incorporating random hospital effects into the model. And a potential benefit could be gained by using best prediction technology to improve the predictions for individual hospitals.

How should the random effects be incorporated? For many problems the decision between treating a factor as fixed or random is a subtle one. As illustration, if we change the example of hospitalization costs slightly and study only three hospitals, all of which are unique and whose effects

cannot easily be regarded as a random sample, we must treat them as fixed. But this would not fundamentally change the way in which they would be incorporated into the mean of the response. This line of argument suggests that random factors should be incorporated in the same manner and in the same portion of the model as the fixed factors. This is exactly the approach of Chapter 6. Our basic linear model there had mean $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$. We incorporated random effects by enlarging the model as $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$. If we write a combined model matrix $\mathbf{X}^* = [\mathbf{X} \quad \mathbf{Z}]$ and an enlarged "parameter" vector $\boldsymbol{\beta}^* = [\boldsymbol{\beta}' \quad \mathbf{u}']'$ it is easy to see that $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}^*\boldsymbol{\beta}^*$.

This suggests a straightforward extension of the generalized linear models of Chapter 5: Append the random effects in the form $\mathbf{Zu}$ to the linear predictor $\mathbf{X}\boldsymbol{\beta}$. This will achieve the two main goals of incorporating correlation and allowing broader inference. However, the nonlinear nature of the model creates complications not encountered in Chapter 6.

In the remainder of the chapter we define the generalized linear mixed model (GLMM), explore the consequences of adding random factors and discuss a variety of inferential methods. The issue of prediction of random effects we leave to Chapter 9. Models in which the random effects cannot be incorporated in a linear predictor are dealt with briefly in Chapter 11.

8.2 STRUCTURE OF THE MODEL

a. Conditional distribution of y

To specify the model we start with the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$. As in (5.5) and (5.6), the response vector $\mathbf{y}$ is typically, but not necessarily, assumed to consist of conditionally independent elements, each with a distribution with density from the exponential family or similar to the exponential family:

$$\begin{aligned} y_i|\mathbf{u} &\sim \text{indep. } f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) \\ f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) &= \exp\{[y_i\gamma_i - b(\gamma_i)]/\tau^2 - c(y_i,\tau)\}. \end{aligned} \tag{8.1}$$

From (5.12) we know that the conditional mean of y_i is related to γ_i in (8.1) via the identity $\mu_i = \partial b(\gamma_i)/\partial\gamma_i$. It is a transformation of this mean that we wish to model as a linear model in both the fixed and

random factors:

$$\begin{aligned} \mathrm{E}[y_i|\mathbf{u}] &= \mu_i \\ g(\mu_i) &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}. \end{aligned} \tag{8.2}$$

As in Chapter 5, $g(\cdot)$ is a known function, called the *link function* (since it links together the conditional mean of y_i and the linear form of predictors), $\mathbf{x}_i'$ is the ith row of the model matrix for the fixed effects, and $\boldsymbol{\beta}$ is the fixed effects parameter vector. To that specification we have added $\mathbf{z}_i'$, which is the ith row of the model matrix for the random effects, and $\mathbf{u}$, the random effects vector. Note that we are using μ_i here to denote the conditional mean of y_i given $\mathbf{u}$, not the unconditional mean. To complete the specification we assign a distribution to the random effects:

$$\mathbf{u} \sim f_{\mathbf{U}}(\mathbf{u}). \tag{8.3}$$

In light of the fact that the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$ is just a notational extension of the generalized linear model of Chapter 5 (i.e., μ_i represents the conditional mean rather than the marginal or unconditional mean; otherwise, all is the same), many of the relationships derived there will hold. Correspondingly, as below (5.14), we denote the conditional variance of y_i given $\mathbf{u}$ as $\tau^2 v(\mu_i)$ in order to display its dependence on the conditional mean μ_i.

8.3 CONSEQUENCES OF HAVING RANDOM EFFECTS

a. Marginal versus conditional distribution

Since the model specification in (8.1) and (8.2) is made conditional on the value of $\mathbf{u}$, we now derive aspects of the marginal distribution of $\mathbf{y}$ in order to understand what has been assumed for the observed data.

b. Mean of y

The mean of $\mathbf{y}$ can be derived by the usual device of iterated expectation:

$$\begin{aligned} \mathrm{E}[y_i] &= \mathrm{E}\left[\mathrm{E}[y_i|\mathbf{u}]\right] \\ &= \mathrm{E}[\mu_i] \\ &= \mathrm{E}[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})]. \end{aligned} \tag{8.4}$$

This cannot, in general, be simplified, due to the nonlinear function $g^{-1}(\cdot)$.

To illustrate for a particular $g(\cdot)$, suppose we have a log link so that $g(\mu) = \log \mu$ and $g^{-1}(x) = \exp\{x\}$. Then we have

$$\begin{aligned} \mathrm{E}[y_i] &= \mathrm{E}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] \\ &= \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathrm{E}[\exp\{\mathbf{z}_i'\mathbf{u}\}] \\ &= \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}M_{\mathbf{u}}(\mathbf{z}_i), \end{aligned} \tag{8.5}$$

where $M_{\mathbf{u}}(\mathbf{z}_i)$ is the moment generating function of $\mathbf{u}$ evaluated at $\mathbf{z}_i$ (see Section S.1c).

Suppose further that $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and that each row of $\mathbf{Z}$ has a single entry equal to 1 with all the rest being zero. Then $M_u(\mathbf{z}_i) = \exp\{\sigma_u^2/2\}$ and

$$\mathrm{E}[y_i] = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\exp\{\sigma_u^2/2\}$$

or

$$\log \mathrm{E}[y_i] = \mathbf{x}_i'\boldsymbol{\beta} + \sigma_u^2/2. \tag{8.6}$$

c. Variances

To derive the marginal variance of $\mathbf{y}$ we use formula (1.14):

$$\begin{aligned} \mathrm{var}(y_i) &= \mathrm{var}(\mathrm{E}[y_i|\mathbf{u}]) + \mathrm{E}[\mathrm{var}(y_i|\mathbf{u})] \\ &= \mathrm{var}(\mu_i) + \mathrm{E}[\tau^2 v(\mu_i)] \\ &= \mathrm{var}(g^{-1}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}]) + \mathrm{E}\left[\tau^2 v(g^{-1}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}])\right], \end{aligned} \tag{8.7}$$

which again cannot be simplified appreciably without making specific assumptions about the form of $g(\cdot)$ and/or the conditional distribution of $\mathbf{y}$.

To illustrate the derivation assume, as before, that we have a log link and now further assume that the elements of $\mathbf{y}$, given $\mathbf{u}$, are independent with a Poisson distribution. Hence the conditional variance of y_i given $\mathbf{u}$ is $\tau^2 v(\mu_i) = \mu_i$. Using these facts in (8.7) gives

$$\begin{aligned} \mathrm{var}(y_i) &= \mathrm{var}(\mu_i) + \mathrm{E}[\mu_i] \\ &= \mathrm{var}(\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}) + \mathrm{E}\left[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}\right] \end{aligned}$$

$$\begin{aligned} &= \mathrm{E}[(\exp\{2(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})\})] - [\mathrm{E}(\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\})]^2 \\ &\qquad + \mathrm{E}\left[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}\right] \qquad (8.8) \\ &= \exp\{2\mathbf{x}_i'\boldsymbol{\beta}\}\left(M_u(2\mathbf{z}_i) - [M_u(\mathbf{z}_i)]^2 + \exp\{-\mathbf{x}_i'\boldsymbol{\beta}\}M_u(\mathbf{z}_i)\right). \end{aligned}$$

If we make the further assumption that $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and that each row of $\mathbf{Z}$ has a single entry equal to 1 with all the rest being zero, then

$$\begin{aligned} \mathrm{var}(y_i) &= \exp\{2\mathbf{x}_i'\boldsymbol{\beta}\}\left(\exp\{2\sigma_u^2\} - \exp\{\sigma_u^2\}\right) + \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\exp\{\sigma_u^2/2\} \\ &= \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \sigma_u^2/2\}\left(\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\left[\exp\{3\sigma_u^2/2\} - \exp\{\sigma_u^2/2\}\right] + 1\right) \\ &= \mathrm{E}[y_i]\left(\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\left[\exp\{3\sigma_u^2/2\} - \exp\{\sigma_u^2/2\}\right] + 1\right). \qquad (8.9) \end{aligned}$$

Since the term in parentheses in (8.9) is greater than 1, we see that the variance is larger than the mean. Therefore, although the conditional distribution of y_i given $\mathbf{u}$ is Poisson, the marginal distribution cannot be. In fact, under these assumptions, it will always be overdispersed (see Section 2.6b-ii) compared to the Poisson distribution. In this sense we can think of random effects as a way to model or attribute overdispersion to a particular source.

d. Covariances and correlations

As noted before, the use of random effects introduces a correlation among observations which have any random effect in common. The same is true for generalized linear mixed models. Assuming conditional independence of the elements of $\mathbf{y}$ and using (1.16), we have

$$\begin{aligned} \mathrm{cov}(y_i, y_j) &= \mathrm{cov}(\mathrm{E}[y_i|\mathbf{u}], \mathrm{E}[y_j|\mathbf{u}]) + \mathrm{E}[\mathrm{cov}(y_i, y_j|\mathbf{u})] \\ &= \mathrm{cov}(\mu_i, \mu_j) + \mathrm{E}[0] \\ &= \mathrm{cov}(g^{-1}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}], g^{-1}[\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u}]). \qquad (8.10) \end{aligned}$$

If we have a log link, this can be evaluated as

$$\begin{aligned} \mathrm{cov}(y_i, y_j) &= \mathrm{cov}(\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}, \exp\{\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u}\}) \\ &= \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_j'\boldsymbol{\beta}\}\mathrm{cov}(\exp\{\mathbf{z}_i'\mathbf{u}\}, \exp\{\mathbf{z}_j'\mathbf{u}\}) \qquad (8.11) \\ &= \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_j'\boldsymbol{\beta}\}\left[M_u(\mathbf{z}_i + \mathbf{z}_j) - M_u(\mathbf{z}_i)M_u(\mathbf{z}_j)\right]. \end{aligned}$$

Again we make further assumptions, namely that $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_u^2)$ and that each row of $\mathbf{Z}$ has a single entry equal to 1 with all the rest being zero. Then

$$\operatorname{cov}(y_i, y_j) = \exp\{\mathbf{x}_i'\beta + \mathbf{x}_j'\beta\}\left[\exp\{\sigma_u^2\}(\exp\{\mathbf{z}_i'\mathbf{z}_j\sigma_u^2\} - 1)\right] \tag{8.12}$$

which is equal to zero if $\mathbf{z}_i'\mathbf{z}_j = 0$ (i.e., if the two observations do not share a random effect) and is positive otherwise (in which case $\mathbf{z}_i'\mathbf{z}_j = 1$).

From (8.12) and (8.9), when $\mathbf{z}_i'\mathbf{z}_j = 1$, we can calculate the correlation (after canceling $\exp\{\mathbf{x}_i'\beta + \mathbf{x}_j'\beta\}$ in the numerator and denominator) as:

$$\begin{aligned}
&\operatorname{corr}(y_i, y_j) \\
&= \frac{e^{2\sigma_u^2} - e^{\sigma_u^2}}{\sqrt{\left(e^{2\sigma_u^2} - e^{\sigma_u^2} + e^{-\mathbf{x}_i'\beta + \sigma_u^2/2}\right)\left(e^{2\sigma_u^2} - e^{\sigma_u^2} + e^{-\mathbf{x}_j'\beta + \sigma_u^2/2}\right)}} \\
&= \frac{1}{\sqrt{\left(1 + \eta e^{-\mathbf{x}_i'\beta}\right)\left(1 + \eta e^{-\mathbf{x}_j'\beta}\right)}}, \qquad (8.13)
\end{aligned}$$

where η is given by $1/(e^{3\sigma_u^2/2} - e^{\sigma_u^2/2})$.

8.4 ESTIMATION BY MAXIMUM LIKELIHOOD

a. Likelihood

From (8.1), (8.2), and (8.3) it is straightforward to write down a formula for the likelihood:

$$L = \int \prod_i f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u}, \tag{8.14}$$

where, as before, the integration is over the q-dimensional distribution of $\mathbf{u}$.

As an example, consider modeling data in correlated clusters thought to come from a Poisson distribution. An example of such a situation is described in Diggle et al. (1994) in which they consider the analysis of the number of epileptic seizures in patients on a drug or placebo. In this context, the clusters would be repeated measurements taken

on the same patients. Let y_{ij} denote the jth count taken on the ith cluster. We might therefore create a model as:

$$\begin{aligned} y_{ij}|\mathbf{u} &\sim \text{indep. Poisson}(\mu_{ij}); \quad i = 1, 2, \ldots, m; j = 1, 2, \ldots n_i; \\ \log \mu_{ij} &= \mathbf{x}'_{ij}\beta + u_i \\ u_i &\sim \text{i.i.d. } \mathcal{N}(0, \sigma_u^2). \end{aligned} \tag{8.15}$$

This uses a log link and a normal distribution for the random cluster (patient) effects. The normal distribution assumption for the random effects is viable since the log link carries the range of the parameter space for μ_{ij} into the entire real line. The random effects u_i are shared among observations within the same cluster and hence those observations are being modeled as correlated.

The log likelihood can be simplified as follows (see E 8.3)

$$\begin{aligned} l &= \log\left(\prod_{i=1}^{m}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\frac{\mu_{ij}^{y_{ij}}e^{-\mu_{ij}}}{y_{ij}!}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}u_i^2}du_i\right) \\ &= \mathbf{y}'\mathbf{X}\beta - \sum_{i,j}\log y_{ij}! \\ &\quad + \sum_i \log\int_{-\infty}^{\infty}\exp\left\{y_{i\cdot}u_i - \Sigma_j e^{\mathbf{x}'_{ij}\beta+u_i}\right\}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}u_i^2}du_i. \end{aligned} \tag{8.16}$$

Unfortunately, (8.16) cannot be simplified further or evaluated in closed form and hence maximizing values cannot be expressed in closed form either.

In the simplest cases, numerical integration for calculating the likelihood is straightforward and hence numerical maximization of the likelihood is not too difficult. For example, for (8.15), as seen in (8.16), the log likelihood is the sum of independent contributions from each cluster, each of which involves just a single-dimensional integral. This integral can be evaluated accurately using standard quadrature techniques, for example, Gauss–Hermite quadrature (see Chapter 10).

This "brute force" approach to maximum likelihood works relatively well in simple situations: a single random effect, two or perhaps three nested random effects, and random effects which come in clusters (e.g., longitudinal data with subjects having random intercepts and slopes). However, for more complicated structures (e.g., crossed random factors) it fails.

b. Likelihood equations

– i. *For the fixed effects parameters*

Even though the likelihood equations are numerically difficult, we can write them in a simpler form. From (8.14)

$$l = \log \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} = \log f_{\mathbf{Y}}(\mathbf{y}), \tag{8.17}$$

so that

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \frac{\partial}{\partial \beta} \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} \Big/ f_{\mathbf{Y}}(\mathbf{y}) \\ &= \int \left[\frac{\partial}{\partial \beta} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) \right] f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} \Big/ f_{\mathbf{Y}}(\mathbf{y}), \end{aligned} \tag{8.18}$$

since $f_{\mathbf{U}}(\mathbf{u})$ does not involve β. Noting that

$$\begin{aligned} \frac{\partial}{\partial \beta} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) &= \left(\frac{1}{f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})} \frac{\partial f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})}{\partial \beta} \right) f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) \\ &= \frac{\partial \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})}{\partial \beta} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) \end{aligned} \tag{8.19}$$

we can rewrite (8.18) as

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \int \frac{\partial \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})}{\partial \beta} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} \Big/ f_{\mathbf{Y}}(\mathbf{y}) \\ &= \int \frac{\partial \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u})}{\partial \beta} f_{\mathbf{U}|\mathbf{y}}(\mathbf{u}|\mathbf{y})\, d\mathbf{u}. \end{aligned} \tag{8.20}$$

Using (5.18), which gives the derivative of the log likelihood for a GLM, in (8.20) gives

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \int \mathbf{X}'\mathbf{W}^*(\mathbf{y} - \boldsymbol{\mu}) f_{\mathbf{U}|\mathbf{y}}(\mathbf{u}|\mathbf{y})\, d\mathbf{u} \\ &= \mathbf{X}'\mathrm{E}[\mathbf{W}^*|\mathbf{y}] - \mathbf{X}'\mathrm{E}[\mathbf{W}^*\boldsymbol{\mu}|\mathbf{y}], \end{aligned} \tag{8.21}$$

where $\mathbf{W}^* = \left\{ {}_d[a(\phi)v(\mu_i)g_\mu(\mu_i)]^{-1} \right\}$.

The likelihood equation for β is therefore

$$\mathbf{X}'\mathrm{E}[\mathbf{W}^*|\mathbf{y}] = \mathbf{X}'\mathrm{E}[\mathbf{W}^*\boldsymbol{\mu}|\mathbf{y}], \tag{8.22}$$

which is similar to (5.19), the difference being that $\mathbf{W}^*$ and $\mathbf{W}^*\mu$ are replaced by their conditional expected values given $\mathbf{y}$.

In cases like the Poisson example of (8.15), $\mathbf{W}^* = \mathbf{I}$ and the equations simplify to

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathrm{E}[\mu|\mathbf{y}]. \tag{8.23}$$

Computing issues related to solving these equations are described in Chapter 10.

– ii. *For the random effects parameters*

A result similar to (8.20) can be derived for the ML equations for the parameters in the distribution of $f_{\mathbf{U}}(\mathbf{u})$. Let φ denote those parameters so that

$$\begin{aligned} \frac{\partial l}{\partial \varphi} &= \int \frac{\partial \log f_{\mathbf{U}}(\mathbf{u})}{\partial \varphi} f_{\mathbf{U}|\mathbf{y}}(\mathbf{u}|\mathbf{y})\, d\mathbf{u} \\ &= \mathrm{E}\left[\frac{\partial \log f_{\mathbf{U}}(\mathbf{u})}{\partial \varphi}\middle|\mathbf{y}\right]. \end{aligned} \tag{8.24}$$

Further simplifications are not possible without specifying a form for the random effects distribution.

8.5 MARGINAL VERSUS CONDITIONAL MODELS

Instead of starting from the conditional specification as in (8.1), (8.2), and (8.3), we might directly hypothesize a model for the mean of $\mathbf{y}$. As an example, suppose y_{ij} is equal to 1 if the jth child of woman i is born prematurely and is zero otherwise and assume we have a single predictor x_{ij} = number of drinks of alcohol per day. The marginal approach would model the marginal mean of y_{ij} directly by, for example, assuming that a logistic regression model, as in (3.109), fits the data:

$$\mathrm{logit}(\mathrm{E}[y_{ij}]) = \mathrm{logit}(\mathrm{P}\{y_{ij} = 1\}) = \alpha + \beta x_{ij}.$$

In words, the model would be for logit of the probability of premature birth, averaged over a population of women. Of course, if the model was for correlated data, we would not be able to assume the observations were independent.

On the other hand, our typical conditional approach corresponds to hypothesizing the existence of a random factor for women and specifying a conditional model such as

$$\mathrm{logit}(\mathrm{E}[y_{ij}|\mathbf{u}]) = \alpha + \beta x_{ij} + u_i,$$

where u_i represents the random woman effect. This corresponds to modeling the conditional probability of a premature birth for each woman.

From a probabilistic perspective, we can calculate the marginal distribution of $\mathbf{y}$ (at least conceptually—it might be computationally difficult) from the distribution of $\mathbf{u}$ and the conditional distribution of $\mathbf{y}|\mathbf{u}$. It is not possible to recover the marginal of $\mathbf{u}$ and the conditional distribution of $\mathbf{y}|\mathbf{u}$ from the marginal distribution of $\mathbf{y}$. This would seem to favor the conditional specification of the model.

However, in some cases, the marginal distribution (or perhaps only the marginal mean) may be adequate for answering questions of interest. For example, in the alcohol consumption example, a natural question of interest is how much could the incidence of premature birth be reduced by lowering, on average, women's alcohol consumption. In such cases, the potentially difficult problem of specifying the conditional distribution of $\mathbf{y}|\mathbf{u}$ and the marginal distribution of $\mathbf{u}$ can be avoided. This is an advantage of marginal modeling and the basis of the generalized estimating equations approach, which is described in Section 8.6a.

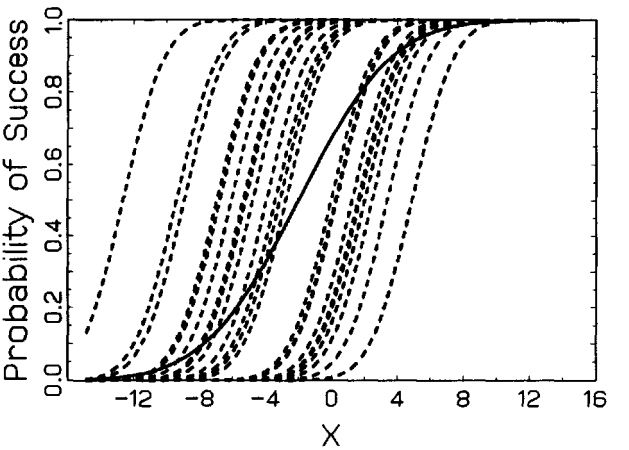
Distinguishing conditional from marginal models is straightforward probabilistically, but it is often difficult in practice. For example, a researcher might be interested in "the influence of alcohol consumption on premature birth", which would not specify which type of model to build. In our experience, researchers often think about building models in a mechanistic way, which seems more compatible with the conditional approach. Again considering the premature birth example, a researcher might think about the influence of alcohol consumption by trying to understand how alcohol influenced individual women's physiology.

The distinction between conditional and marginal models is an important one to keep in mind in practice. The reason is perhaps easiest to see in a binary data, probit-normal model:

$$\begin{aligned} \mathrm{E}[y_i] = \mathrm{P}\{y_i = 1\} &= \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}), \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}). \end{aligned} \tag{8.25}$$

It is not hard to show, as we do later in (8.52), that if the conditional

Figure 8.1: Probability of success versus a predictor for the marginal and conditional versions of a probit-normal model. Solid line, marginal model; Dashed lines, realizations of the conditional model.



mean is given as in (8.25), then the unconditional mean is

$$\mathrm{E}[y_i] = \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sqrt{\mathbf{z}_i'\mathbf{D}\mathbf{z}_i + 1}}\right) \equiv \Phi\left(\mathbf{x}_i'\boldsymbol{\beta}^*\right) \tag{8.26}$$

so that $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\sqrt{\mathbf{z}_i'\mathbf{D}\mathbf{z}_i + 1}$. Hence $\boldsymbol{\beta}$ represents the magnitude of the effect of the predictors on the conditional distribution, while $\boldsymbol{\beta}^*$ represents the magnitude of the effect on the marginal distribution.

Clearly, $\boldsymbol{\beta}$ is always larger than $\boldsymbol{\beta}^*$ in absolute value. Why this is so is perhaps easiest to understand graphically, as shown in Figure 8.1. Each of the realized values of the conditional model [i.e., equation (8.25) plotted for various realized u_i], shown by the dashed lines in the figure, have a large value of $\boldsymbol{\beta}$. However, the variance of the random effect is quite large and when all the curves are averaged, the resulting unconditional mean has a much smaller slope and much smaller value for $\boldsymbol{\beta}^*$.

There are other advantages to the conditional approach. For example, if two studies are performed in different populations with different variances the marginal models will be different even though the conditional models are the same. Returning to the alcohol consumption example, suppose a small, preliminary study is conducted with a pur-

posefully homogeneous study population (and hence a small variance for the subject random effect). Later, a larger scale study is conducted on a wider scale with a more heterogeneous study population. Even if the effect on every person in both studies is the same, the marginal models will differ because of the different variances.

8.6 OTHER METHODS OF ESTIMATION

The difficulty in evaluating the likelihood for models such as (8.15) and the fact that numerical maximization must be resorted to has led to both alternative approaches and to a body of research for effective ways to compute and maximize the likelihoods. The latter topic is treated in Chapter 10; here we introduce some of the alternative methods of estimation.

a. Generalized estimating equations

The generalized estimating equations (GEEs) approach begins by positing a marginal generalized linear model for the mean of $\mathbf{y}$ as a function of the predictors. For example, for binary data we might hypothesize a logistic regression for the mean:

$$\text{logit}(\text{E}[\mathbf{y}]) = \mathbf{X}\boldsymbol{\beta}.$$

If we used a working assumption of independence (as in Section 7.8) of all the elements of $\mathbf{y}$, the ML estimating equations for $\boldsymbol{\beta}$ would be, from E 5.9,

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\text{E}[\mathbf{y}]. \tag{8.27}$$

These are *unbiased estimating equations*, meaning that the difference between the right-hand side and the left-hand side is zero, namely $\text{E}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\text{E}[\mathbf{y}]) = \mathbf{0}$. It is not surprising and is true under regularity conditions (Heyde, 1997, Sec. 12.2) that solutions to unbiased estimating equations give consistent estimators.

Operationally, this estimator could be calculated by pretending that all the data were independent and conducting a standard logistic regression analysis. Just as described in Section 7.8, this is often a nearly fully efficient estimator but the standard errors, variance estimates, tests and confidence intervals would often be highly misleading. Again, as in Section 7.8, this can be dealt with by using the *estimator* that naively assumes independence but properly calculating its (large-sample) variance.

For longitudinal data with m subjects and with $\mathbf{y}_i$ denoting the data for the ith subject, we have

$$\mathbf{y} = \left\{ {}_c\, \mathbf{y}_i \right\}_{i=1}^{m} \text{ and } \mathbf{X} = \left\{ {}_c\, \mathbf{X}_i \right\}_{i=1}^{m}$$

and the estimating equation, (8.27), for binary data becomes

$$\sum \mathbf{X}_i \mathbf{y}_i = \sum \mathbf{X}_i \mathrm{E}[\mathbf{y}_i]. \tag{8.28}$$

From this it can be shown that (Heyde, 1997, Secs. 4.2 and 12.4) the large-sample variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \mathrm{var}_\infty(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \,\mathrm{var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\textstyle\sum \mathbf{X}_i'\mathbf{X}_i)^{-1}(\sum \mathbf{X}_i' \,\mathrm{var}(\mathbf{y}_i)\mathbf{X}_i)(\sum \mathbf{X}_i'\mathbf{X}_i)^{-1}, \end{aligned} \tag{8.29}$$

upon the assumption of independence among the $\mathbf{y}_i$. This can be consistently estimated as $m \to \infty$ with

$$\widehat{\mathrm{var}}_\infty(\hat{\boldsymbol{\beta}}) = (\textstyle\sum \mathbf{X}_i'\mathbf{X}_i)^{-1}(\sum \mathbf{X}_i'(\mathbf{y}_i - \hat{\mathrm{E}}[\mathbf{y}_i])(\mathbf{y}_i - \hat{\mathrm{E}}[\mathbf{y}_i])'\mathbf{X}_i)(\sum \mathbf{X}_i'\mathbf{X}_i)^{-1}, \tag{8.30}$$

where $\hat{\mathrm{E}}[\mathbf{y}_i] = 1/(1 + \exp\{-\mathbf{X}_i\hat{\boldsymbol{\beta}}\})$.

The working assumption of independence may lead to inefficient estimators (Fitzmaurice, 1995) and other working variance-covariance structures can be entertained. In that case, for (8.28) we have

$$\sum \mathbf{X}_i \mathbf{W}_i \mathbf{y}_i = \sum \mathbf{X}_i \mathbf{W}_i \mathrm{E}[\mathbf{y}_i], \tag{8.31}$$

where $\mathbf{W}_i^{-1} = \mathrm{var}_W(\mathbf{y}_i)$ is the working variance for $\mathbf{y}_i$. This engenders corresponding changes in (8.29) and (8.30). See Diggle et al. (1994, Secs. 8.2.3 and 8.4.2) for more details.

b. Penalized quasi–likelihood

For the generalized linear models (GLMs) of Chapter 5, the use of working variates, as in (5.4) and (5.28), and the principle of quasi–likelihood (Section 5.6) are highly useful concepts. Quasi–likelihood is attractive because of its ability to generate highly efficient estimators without making precise distributional assumptions. Working variates form the basis of efficient computing algorithms for both maximum likelihood and maximum quasi–likelihood. A natural question is whether they can be adapted for use in GLMMs.

Working variates for GLMs begin with a Taylor expansion of the link function around the mean of y_i :

$$t_i \equiv \mathbf{x}_i'\boldsymbol{\beta} + g_\mu(\mu_i)(y_i - \mu_i).$$

The working variate thus follows a linear model and can be used to form a provisional estimate of $\boldsymbol{\beta}$. This local approximation is repeated at each update of an iterative algorithm.

The direct analog of working variates for the GLMM specification in (8.1) and (8.2) would be an expansion around the conditional mean of y_i :

$$t_i \equiv \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u} + g_\mu(\mu_i)(y_i - \mu_i)$$

or

$$\mathbf{t} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}), \tag{8.32}$$

where $\boldsymbol{\Delta} = \left\{ {}_d\, g_\mu(\mu_i) \right\}$. To derive a local approximation, the next step would be to calculate the variance of $\mathbf{t}$. But this approach quickly becomes complicated since $\boldsymbol{\Delta}$ (through its dependence on $\boldsymbol{\mu}$) and $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}|\mathbf{u}]$ itself are random functions of $\mathbf{u}$ and their variances are not easily calculated.

A possible simplification is to set $\mathbf{u}$ in $\boldsymbol{\Delta}$ equal to its mean, $\mathbf{0}$, simplifying (8.32) to be

$$\mathbf{t} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\Delta}^*(\mathbf{y} - \boldsymbol{\mu}), \tag{8.33}$$

where $\boldsymbol{\Delta}^* = \left\{ {}_c\, g_\mu[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})] \right\}$.

Under this simplification

$$\begin{aligned} \mathrm{var}(\mathbf{t}) &= \mathbf{Z}\mathbf{D}\mathbf{Z}' + \boldsymbol{\Delta}^*\mathrm{var}(\mathbf{y} - \boldsymbol{\mu})\boldsymbol{\Delta}^* \\ &\equiv \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}. \end{aligned} \tag{8.34}$$

That is, the working variate $\mathbf{t}$ approximately follows a linear mixed model (LMM) as in Chapter 6. This suggests an iterative algorithm (Schall, 1991) in which an LMM is fitted to get estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$. These are then used to recalculate the working variate, and so on.

A completely different justification of this approach is via what is called *penalized quasi-likelihood* (PQL). Recall that quasi-likelihood

does not specify a distribution, only the mean-to-variance relationship. This is not a sufficient basis on which to estimate the variance-covariance structure. One suggestion (Green and Silverman, 1994) to remedy this defect is to add a penalty function to the quasi–likelihood of the form $\frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}$, that is

$$\text{PQL} = \sum Q_i - \frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}, \tag{8.35}$$

where Q_i is defined in (5.50).

The maximum quasi–likelihood equations would come from differentiating (8.35) with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ and would be [compare (5.53)]

$$\frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}$$

and

$$\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) - \mathbf{D}^{-1}\mathbf{u} = \mathbf{0}. \tag{8.36}$$

These lead (Breslow and Clayton, 1993) to a computational algorithm similar to that of Schall(1991). Yet another justification for this approach is via Laplace approximations (see Chapter 10 and Wolfinger, 1994).

Despite the number of ways in which basically the same approach has been justified, it has not been found to work well in practice, especially for binary data in small clusters (Breslow and Clayton, 1993; Breslow and Lin, 1995; Lin and Breslow, 1996). We therefore recommend that unmodified penalized quasi–likelihood not be used in practice. More detail is given in Chapter 10.

c. Conditional likelihood

An approach very different in nature to integrating random effects out of the distribution and working with the marginal distribution is to consider a conditional likelihood and construct conditional estimators and tests as in Section 3.8e. In the conditional approach we start with the conditional distribution of the data given the random effects, but instead of hypothesizing a distribution for these effects and integrating them out, they are treated as fixed parameters. The sufficient statistics for them are derived and the conditional distribution given the sufficient statistics (which, by definition, is free of these effects) is used for inferential purposes.

A classic application of the conditional approach is to the case of matched pairs binary data. For example, suppose we wish to ask whether cancer patients get more effective treatment in major cancer centers than in community hospitals. We cannot compare remission rates directly since patient populations might be drastically different. For example, major cancer centers might appear to provide poorer treatment merely because they treat the most difficult cases.

A possible solution is to employ a matched pairs design: A patient from a cancer center is matched with a patient from a community hospital on the basis of treatment date, type of treatment received, and patient's age. Suppose that the response variable is whether or not there is a sizable shrinkage in tumor size within 90 days and let $y_{ij} = 1$ for shrinkage and 0 otherwise. Here i indexes pairs $i = 1, 2, \ldots, n$ and j indexes type of hospital (with $j = 1$ representing a cancer center and $j = 2$ representing a community hospital). Also, let x_{ij} be 0 when $j = 1$ and 1 when $j = 2$.

A possible model for y_{ij} is:

$$\begin{aligned} y_{ij}|u_i &\sim \text{indep. Bernoulli}[\pi(x_{ij})] \\ \text{logit}[\pi(x_{ij})] &= \alpha + u_i + \beta x_{ij} \end{aligned}$$

Primary interest focuses on β, which represents the log odds of tumor shrinkage for community hospitals as compared to cancer centers, which is assumed constant within each pair. The u_i represent the pair-to-pair differences in the probability of tumor shrinkage.

What happens if we treat the u_i as fixed effects and estimate them, along with β? Maximum likelihood gives (see E 8.6).

$$\hat{\beta} = 2 \log \frac{N_{01}}{N_{10}}, \tag{8.37}$$

where N_{10} is the number of pairs with $y_{i1} = 1$ and $y_{i2} = 0$ and where N_{01} is the number of pairs with $y_{i1} = 0$ and $y_{i2} = 1$. This is perhaps easiest to visualize in a 2×2 format as in Table 8.1.

Table 8.1: Fate of Matched Pairs

Cancer Center	Community Hospital	
	Success	Failure
Success	N_{11}	N_{10}
Failure	N_{01}	N_{00}

It is not hard to show that this ML estimator is twice what it "should" be in the sense that it converges to 2β (see E 8.7).

What is the remedy? A commonly-used approach is that of conditional likelihood, in which we derive the sufficient statistics for the u_i and work with the conditional distribution given those sufficient statistics.

We follow the development in (3.134) through (3.138). If interest focuses on β, then we will want to base inferences on the conditional distribution of $T = \sum_{i,j} y_{ij}x_{ij}$ given $S_1 = y_{1\cdot}, S_2 = y_{2\cdot}, \ldots, S_m = y_{m\cdot}$. In our example,

$$\begin{aligned} T &= \sum_{i,j} y_{ij}x_{ij} \\ &= \sum_i y_{i2} \\ &= y_{\cdot 2} \\ &= \text{number of successes in community hospitals,} \end{aligned} \tag{8.38}$$

so we want the conditional distribution of the total number of successes in the community hospitals conditional on the number of successes in each pair. From Table 8.1, $T = N_{11} + N_{01}$. Now N_{11} is just the number of pairs that have two successes, so conditional on S_i, it is known and fixed. We therefore focus on the conditional distribution of N_{01} given the S_i. We build it up in two steps. First consider a pair for which $S_i = 1$.

If $S_i = 1$, there are two possibilities: $\{y_{i1} = 0, y_{i2} = 1\}$ or $\{y_{i1} = 1, y_{i2} = 0\}$. The conditional probability of the first event is

$$\begin{aligned} \mathrm{P}\{y_{i1} = 1, y_{i2} = 0 | S_i = 1\} &= \frac{\mathrm{P}\{y_{i1} = 1, y_{i2} = 0\}}{\mathrm{P}\{y_{i1} = 0, y_{i2} = 1\} + \mathrm{P}\{y_{i1} = 1, y_{i2} = 0\}} \\ &= \frac{\left(1 - \frac{1}{1+e^{-(\alpha+u_i)}}\right)\frac{1}{1+e^{-(\alpha+u_i+\beta)}}}{\left(1 - \frac{1}{1+e^{-(\alpha+u_i)}}\right)\frac{1}{1+e^{-(\alpha+u_i+\beta)}} + \frac{1}{1+e^{-(\alpha+u_i)}}\left(1 - \frac{1}{1+e^{-(\alpha+u_i+\beta)}}\right)} \\ &= \frac{e^{-(\alpha+u_i+\beta)}}{e^{-(\alpha+u_i)} + e^{-(\alpha+u_i+\beta)}} \\ &= \frac{1}{1+e^{-\beta}}. \end{aligned} \tag{8.39}$$

So the conditional distribution of y_{i2} given $S_i = 1$ is Bernoulli with a probability of success having a logit of β.

Writing T as

$$\begin{aligned} T &= N_{11} + N_{01} \\ &= \sum_{S_i=2} 1 + \sum_{S_i=1} I_{\{y_{i2}=1\}}, \end{aligned} \tag{8.40}$$

shows that the conditional distribution of $N_{01} = T - N_{11}$ given the S_i is the sum of independent Bernoullis, each with conditional probability of success of $(1 + e^{-\beta})^{-1}$, that is,

$$N_{01}|S \sim \text{binomial}\left(N_{01} + N_{10}, \frac{1}{1+e^{-\beta}}\right). \tag{8.41}$$

It is therefore straightforward to show that the maximizing value of the conditional distribution (the conditional MLE) is

$$\hat{\beta} = \log \frac{N_{01}}{N_{10}}, \tag{8.42}$$

which is a consistent estimator of β as $N_{11} + N_{01}$ increases.

Also, under $H_0 : \beta = 0$, the distribution is binomial$(N_{11} + N_{01}, \frac{1}{2})$, from which exact tests or p-values can easily be derived. To do so, we use as our test statistic N_{01}, the number of successes in the community hospitals out of the pairs for which $S_i = 1$. If we were testing against the alternative that $H_A : \beta > 0$, we would reject for large values of N_{01}. Therefore the p-value for the one-tailed test would be:

$$p\text{-value} = \mathrm{P}\{X \geq N_{01}\}, \tag{8.43}$$

where $X \sim$ binomial$(N_{01} + N_{10}, \frac{1}{2})$.

As a numerical illustration, consider the hypothetical data of Table 8.2.

Table 8.2: Matched Pairs Data

Cancer Center	Community Hospital		Total
	Success	Failure	
Success	501	157	658
Failure	146	132	278
Total	647	289	936

The conditional analysis discards the 501 + 132 = 633 responses for which the response in the cancer center and community hospital are

the same and bases the analysis on the 303 remaining. The p-value for the one-tailed test is

$$\begin{aligned} p\text{-value} &= \mathrm{P}\{X \geq 157\} \\ \text{where } X &\sim \text{binomial}(303, 1/2), \end{aligned} \tag{8.44}$$

which is approximately 0.283. A usual convention is to multiply the one-sided p-value by 2 to get the two-tailed p-value, so the answer is $2(0.283) = 0.566$, suggesting no difference between cancer centers and community hospitals.

By its nature the conditional approach has three potentially serious drawbacks. First, because it treats the random effects as unknown parameters to be conditioned away, it is incapable of making inference to quantities involving the random effects: for example, their variances or predicted values. Second, it discards information that might be available by making only weak assumptions about the form of the distribution from which the random effects are chosen. Third, it removes any information that would be gained by comparing across levels of the random effects. In some situations, virtually *all* the information of interest is garnered from comparisons across levels of a random effect. Hence use of a conditional approach would be disastrous in such a context: It would eliminate all the information of interest due to the extreme manner in which the effects are handled.

Basically, the conditional approach is effective and attractive when interest centers almost exclusively on effects that can be measured *within* levels of a random factor. When extensive information exists across levels of a random factor or when interest focuses on the random factor itself, the conditional approach cannot be used.

d. Simpler models

To avoid the computational difficulties of GLMMs, other models have been considered, mostly on the basis of computational convenience. For example, the beta-binomial model of Section 2.6b has long been used for modeling correlated binary data. The most basic use of the beta-binomial is for a context in which there are m groups, and within the ith group we have n_i observations, $y_{ij} \sim \text{binomial}(k_{ij}, p_{ij})$, with j running from 1 to n_i. This is conditional on the values of the p_{ij}. To complete the specification we assume that $p_{ij} \sim \text{beta}(\alpha_i, \beta_i)$. Given these distributional assumptions it is straightforward to show that the

likelihood is given by

$$L = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{B(\alpha_i + y_{ij}, \beta_i + k_{ij} - y_{ij})}{B(\alpha_i, \beta_i)}, \tag{8.45}$$

where $B(\alpha, \beta)$ represents the beta function.

Working with the log likelihood, some simplifications occur as in (2.79). However, the likelihood still cannot be maximized in closed form, and numerical maximization must be resorted to. Likelihood ratio tests can be performed to test for dispersion or to compare the means of the various groups.

For data having Poisson distributions, a natural distribution to incorporate correlation is the gamma distribution. As with the beta-binomial model, we consider a situation with m groups, and within group i we have n_i observations, $y_{ij} \sim \text{Poisson}(\lambda_{ij})$ with j running from 1 to n_i. This is conditional on the values of the λ_{ij}. To complete the specification we assume that $\lambda_{ij} \sim \text{gamma}(r_i, \beta_i)$. This allows for easy integration over the distribution of the parameters across groups, so that the likelihood takes a simple closed form.

However, like the beta-binomial model, it is limited in its application. It cannot handle models for the fixed effects (e.g., a regression situation), it cannot separate sources of variation in a crossed design, and it generally does not have the flexibility to tackle a wide variety of practical problems. Generalizations to handle more complicated covariate patterns and more complicated random effects structures have been considered by Lee and Nelder (1996), at the cost of additional computational complexity.

8.7 TESTS OF HYPOTHESES

The usual large-sample tools (see Chapter 2 and Sections S.4 and S.5) are about the only techniques currently available for statistical inference.

a. Likelihood ratio tests

The likelihood ratio test for nested models can be performed in the usual way by comparing $-2 \log \Lambda$ to a chi-square distribution. Testing whether a variance component is zero leads to the same boundary-of-the-parameter-space problem noted before in Chapter 2 [see (2.88)]. In

the simple case where we are testing the null hypothesis that a single variance component is equal to zero, the large-sample distribution is a 50/50 mixture of the constant 0 and a χ^2_1 distribution. The critical values are thus given by (see E 8.4) $\chi^2_{1,1-2\alpha}$ for an α-level test.

Since the likelihood cannot, in general, be evaluated analytically the same is true of the likelihood ratio test statistic. It can be calculated only numerically for a given data set. In many cases it is a challenge even to perform the numerical maximization and calculation.

b. Asymptotic variances

Again, with the difficulty of calculating the likelihood, even large-sample variances and standard errors can be a computational burden. Numerical methods must be resorted to to calculate even the observed Fisher information (i.e., the negative of the second derivative matrix of the log likelihood).

c. Wald tests

For large samples, when construction of the observed or expected information is possible, Wald tests can be formed by utilizing the large–sample normality of estimators. This can be for an individual parameter:

$$\frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\widehat{\text{var}}_\infty(\hat{\beta}_i)}} \sim \mathcal{AN}(0,1) \tag{8.46}$$

or for a set of linear combinations of the parameters,

$$\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{K}'\boldsymbol{\beta}_0 \sim \mathcal{AN}(0, \mathbf{K}'\mathbf{I}^{-1}\mathbf{K}), \tag{8.47}$$

where $\mathbf{I}$ represents the observed or expected information.

d. Score tests

For testing the presence of a single random effect or multiple random effects, score tests have also been proposed (Commenges et al., 1994; Jacqmin-Gadda and Commenges, 1995; Lin, 1997; Commenges and Jacqmin-Gadda, 1997). These have the advantage of not requiring the maximum likelihood estimators under the GLMM. However, they often have less power than the tests based on the random effects models.

8.8 ILLUSTRATION: CHESTNUT LEAF BLIGHT

The American chestnut tree used to be a predominant hardwood in the forests of the eastern United States, reaching 80 to 100 feet in height at maturity and providing timber and low-fat, high-protein nutrition for animals and humans in the form of chestnuts. In the early 1900s an imported fungal pathogen, which causes chestnut leaf blight, was introduced into the United States. The disease spread from infected trees in the New York City area and by 1950 had killed more than 3 billion trees and virtually eliminated the chestnut tree in the United States. Economic losses in both timber and nut production have been estimated in the hundreds of billions of dollars. As well, there are ecological impacts in eliminating a dominant species.

To try to bring this tree back to the U.S. forests, several methods have been explored, including the development of blight-resistant varieties. We focus instead on attempts to weaken the fungus by infecting it with a virus that reduces the fungus' virulence. The basic idea is to release these hypovirulent isolates of chestnut blight fungus and let the viruses infect the natural populations of the fungus, thereby allowing chestnuts trees to survive.

Michael Milgroom from the Department of Plant Pathology at Cornell University, and his colleague, Paolo Cortesi from the University of Milan, have studied this system (Cortesi et al., 1996; Cortesi and Milgroom, 1998). Viruses can spread between fungal individuals only when they come in contact and fuse together. A major obstacle in spreading this virus and thus controlling the disease is that different isolates of the fungus cannot necessarily transfer the virus to one another. They have worked with six incompatibility genes, which may block the transmission of this virus between isolates of the fungus. By developing lab isolates that are compatible with a wide variety of naturally occurring isolates (and thus able to transmit the hypovirulence) an avenue may be opened to biocontrol of this fungal disease.

To estimate the effects of these genes Milgroom and Cortesi have made extensive attempts to pair isolates which differ on the first gene only, the second gene only, the first and the second gene, and so on. For each combination of isolates they attempt transmission an average of 30 times and record a binary response of whether or not the attempt succeeded in transmitting the virus.

Questions of interest include whether pre-identified genes actually do have an influence on transmission of the virus (and if so, to what

degree), whether there are other, as yet unidentified, genes that might affect transmission, and whether transmission is symmetric. By symmetry of transmission we mean the following: Suppose the infected fungus is type b at the locus for the first gene and the non-infected isolate (that we are trying to infect) is type B. The two isolates are the same at the other five loci. Is the probability of transmission the same as when using type B to try to infect type b?

a. A random effects probit model

A common model in genetics for describing the presence or absence of a trait is the threshold model. This arises from assuming that a large number of genes each have a small and additive effect and that when the cumulative effect exceeds a threshold of zero, the trait is present in an individual. Letting $y = 1$ denote the presence of the trait, ϵ represent the additive genetic effect and $\mathbf{x}'\boldsymbol{\beta}$ represent either genetic or non-genetic fixed effects, we can appeal to the central limit theorem to give the probit model:

$$\begin{aligned} \mathrm{P}\{y=1\} &= \mathrm{P}\{\mathbf{x}'\boldsymbol{\beta}+\epsilon>0\} \\ \epsilon &\sim \mathcal{N}(0,1), \end{aligned} \tag{8.48}$$

so that we have

$$\mathrm{P}\{y=1\} = \mathrm{P}\{-\epsilon<\mathbf{x}'\boldsymbol{\beta}\} = \Phi(\mathbf{x}'\boldsymbol{\beta})$$

or

$$\Phi^{-1}(\pi) = \mathbf{x}'\boldsymbol{\beta},$$

where $\pi = \mathrm{E}[y] = \mathrm{P}\{y=1\}$.

We use this model by letting y_i be equal to 1 if the attempt succeeds in transmitting the virus and 0 otherwise.

– i. *The fixed effects*

We concentrate first on building the fixed effects portion of the model. With $\mathbf{x}_i'$ the ith row of the model matrix for the fixed effects, our model is

$$\mathbf{x}_i'\boldsymbol{\beta} = \mu + \sum_{j=1}^{6}\alpha_j MCH_j + \sum_{j=1}^{6}\gamma_j ASY_j, \tag{8.49}$$

where MCH_j is 1 if there is a mismatch at locus j and zero otherwise and ASY_j is $1/2$ if there is a mismatch at locus j with a b donor, $-1/2$ if it is a B donor and 0 if there is no mismatch. The effect of a mismatch on gene j (averaged over donor types b and B) is thus measured by α_j, and γ_j measures the difference between a mismatch with a donor type b and a donor type B.

– ii. *The random effects*

The fact that different isolates of the fungus are used which may differ with regard to genes other than the six pre-identified suggests that we might model their effects as being selected from a normal distribution. Let $\mathbf{Z}_D$ be the model matrix for the donor effects, i.e., an incidence matrix identifying which donor isolates are used for which attempted transmissions. Similarly define $\mathbf{Z}_R$ for the recipient isolates, with the ith row of the matrices denoted respectively as $\mathbf{z}'_{iD}$ and $\mathbf{z}'_{iR}$. With $\mathbf{x}'_i$ defined by (8.49), $\mathbf{Z} = [\mathbf{Z}_D \quad \mathbf{Z}_R]$, and $\mathbf{u}' = [\mathbf{u}'_D \quad \mathbf{u}'_R]$, a reasonable model might then be:

$$\begin{aligned} \mathrm{P}\{y_i = 1|\mathbf{u}\} &= \Phi(\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_{iD}\mathbf{u}_D + \mathbf{z}'_{iR}\mathbf{u}_R) \\ &= \Phi(\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u}) \\ \mathbf{u}_D &\sim \mathcal{N}(0, \mathbf{I}\sigma^2_D) \\ \mathbf{u}_R &\sim \mathcal{N}(0, \mathbf{I}\sigma^2_R) \\ \mathbf{u} &\sim \mathcal{N}(0, \mathbf{D}). \end{aligned} \tag{8.50}$$

In this model, $\mathbf{u}_D$ represents the (random) effects of the donor isolate and $\mathbf{u}_R$ represents the (random) effects of the recipient isolate.

– iii. *Consequences of having random effects*

The unconditional mean is given by

$$\begin{aligned} \mathrm{E}[y_i] &= \mathrm{E}\left[\mathrm{E}[y_i|\mathbf{u}]\right] \\ &= \mathrm{E}[\Phi(\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u})]. \end{aligned}$$

This last quantity can most easily be calculated by appeal to the threshold model. We do not necessarily need to believe that the threshold model holds, but can merely use it as a probabilistic identity.

$$\mathrm{E}[y_i] = \mathrm{E}\left[\mathrm{P}\{\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u} + \epsilon_i > 0|\mathbf{u}\}\right]$$

$$\begin{aligned} &= \mathrm{P}\{\mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u} + \epsilon_i > 0\} \\ &= \mathrm{P}\{-(\mathbf{z}_i'\mathbf{u} + \epsilon_i) < \mathbf{x}_i'\beta\} \\ &= \mathrm{P}\{W < \mathbf{x}_i'\beta\}, \end{aligned} \tag{8.51}$$

where $W \sim \mathcal{N}(0, \mathbf{z}_i'\mathbf{D}\mathbf{z}_i + 1)$. The marginal probability is thus

$$\begin{aligned} \mathrm{E}[y_i] &= \Phi\left(\frac{\mathbf{x}_i'\beta}{\sqrt{\mathbf{z}_i'\mathbf{D}\mathbf{z}_i + 1}}\right) \\ &= \Phi(\mathbf{x}_i'\beta^*), \end{aligned} \tag{8.52}$$

where β^* is equal to $\beta/\sqrt{\mathbf{z}_i'\mathbf{D}\mathbf{z}_i + 1}$.

This result is interesting in two ways. First, it is somewhat surprising (and, as it turns out, special) that the form of the relationship of the mean of $\mathbf{y}$ to the fixed effects is probit either conditionally or unconditionally. Second, it shows that the marginal coefficients on the probit scale are always *attenuated* as compared to the conditional coefficients. Thus it clearly is important to keep in mind when considering any of these models whether they represent the response conditional on the random effects or are, instead, marginal calculations.

Since y_i is binary, it has a marginal Bernoulli distribution with mean, $\mathrm{E}[y_i]$, given by (8.52). Its variance is therefore $\mathrm{E}[y_i]\,(1 - \mathrm{E}[y_i])$.

A typical consequence of including random effects is that they induce a correlation between observations sharing the random effects and this model is no exception. From first principles, the covariance of two observations with the same donor and recipient isolates would be given by $\mathrm{cov}(y_i, y_j) = \mathrm{E}[y_i y_j] - \mathrm{E}[y_i]\mathrm{E}[y_j]$. The second part of this can be evaluated using (8.52) and the first part calculated as

$$\mathrm{E}[y_i y_j] = \int_{-\infty}^{\infty} \Phi(\mathbf{x}_i'\beta + \sigma z)\Phi(\mathbf{x}_j'\beta + \sigma z)\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\,dz, \tag{8.53}$$

where $\sigma = \sqrt{\mathbf{z}_i'\mathbf{D}\mathbf{z}_j}$.

– iv. *Likelihood analysis*

Again, with a conditional specification, the likelihood is most naturally calculated by first writing out the conditional distribution and then integrating out the random factors. The conditional distribution given

$\mathbf{u}$ is the product of Bernoulli densities:

$$f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) = \prod_i \Phi(\mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u})^{y_i}[1 - \Phi(\mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u})]^{1-y_i}. \qquad (8.54)$$

The likelihood would then be given by

$$L(\beta, \mathbf{D}|\mathbf{y}) = \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u}, \qquad (8.55)$$

where the integral is of order equal to the dimension of $\mathbf{u}$, which in this example is 259. Furthermore, for the design of this experiment, the likelihood does *not* break down into smaller-dimensional pieces, as it might with longitudinal data. This poses a serious computational problem.

– v. *Results*

Given the difficulty of calculating the likelihood, the techniques of Section 10.3c were used to fit the model. A logistic version of (8.54) was fitted using the Monte Carlo Newton–Raphson technique. This was done since the computations were somewhat faster than for (8.54). The maximized value of the likelihood was estimated by importance sampling.

The variance components were estimated to be

$$\begin{aligned} \hat{\sigma}_D^2 &= 1.6 \\ \hat{\sigma}_R^2 &= 0.5 \end{aligned}$$

indicating a small to moderate correlation among observations taken on the same donor isolate and a somewhat smaller correlation among observations taken on the same recipient isolate. A likelihood ratio test of H_0: *all* $\gamma_i = 0$ gives a value for $-2 \log \Lambda$ of about 160 (since the value is determined by simulation it is not known exactly), which is highly statistically significant when referred to a chi-square distribution with 6 *df*. This indicates that, unfortunately, transmission is asymmetric: it depends on the value at that locus, not just on whether or not there is a match.

Further analysis shows that the fourth gene (tentatively identified from previous research) does not have an effect on transmission. Neither its direct effect nor its asymmetry effect is statistically significant.

8.9 EXERCISES

E 8.1 Show that, if $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and that each row of $\mathbf{Z}$ has a single entry equal to 1 with all the rest being zero, then $M_u(\mathbf{z}) = \exp\{\sigma_u^2/2\}$.

E 8.2 Prove (8.12).

E 8.3 Show that the log likelihood for (8.15) can be written as (8.16).

E 8.4 Suppose that $Y = \delta X$, where $\mathrm{P}\{\delta = 1\} = \mathrm{P}\{\delta = 0\} = \frac{1}{2}$ and $X \sim \chi_1^2$ independent of δ. Show that

$$\mathrm{P}\{Y > \chi^2_{1,1-2\alpha}\} = \alpha.$$

E 8.5 Derive the log likelihood for the Poisson-gamma model described in Section 8.6d.

E 8.6 Show that for (8.37), where the u_i are treated as fixed, unknown parameters to be estimated, that the MLE of β is given by (8.37). *Hints*: Consider separately pairs in which there are zero, one and two successes and first maximize with respect to the u_i, then β.

E 8.7 Show that $\hat{\beta}$ of (8.37) converges to 2β. *Hint*: Calculate $\mathrm{P}\{y_{i1} = 1, y_{i2} = 0\}$ and $\mathrm{P}\{y_{i1} = 0, y_{i2} = 1\}$ and hence the expected value of N_{01} and N_{10}.

Chapter 9

PREDICTION

9.1 INTRODUCTION

Earlier chapters contain results for estimation known as predicting random effects. Section 1.2 describes what is meant by a random effect; in being a random variable, it has mean and second moments, properties of which are shown in Section 1.4 for traditional LMMs. Some real-life examples of random effects are described in Section 1.5. How one decides whether effects are fixed or random is discussed in Section 1.6, a decision tree in Section 1.7 helps with doing this, and Section 1.7c has a brief discussion of predicting random effects. It is this brevity which we now expand upon. In doing so we provide underlying methodology for expressions given earlier for predicted values of random effects in a variety of models. These predictors can be found in equations (2.56), (2.90), (2.103), (3.92), and (6.42), and in Section 6.6a.

We begin by presenting three different but interrelated methods of prediction. In doing so we make numerous references to Searle et al. (1992) wherein Chapter 7 has the extensive mathematical detail supporting much of what we present here. For brevity's sake the Searle et al. book is denoted by VC, from its title, *Variance Components*.

We deal with the general case of $\mathbf{y}$ being available data, in the model for which $\mathbf{u}$ is a vector of random effects. First and second moments of $\mathbf{u}$ and $\mathbf{y}$ are defined by

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_y \end{bmatrix} \quad \begin{bmatrix} \mathbf{D} & \mathbf{C} \\ \mathbf{C}' & \mathbf{V} \end{bmatrix} \right) \tag{9.1}$$

so that $\mathrm{E}[\mathbf{u}] = \boldsymbol{\mu}_u$, $\mathrm{E}[\mathbf{y}] = \boldsymbol{\mu}_y$ and

$$\mathbf{D} = \mathrm{var}(\mathbf{u}), \quad \mathbf{C} = \mathrm{cov}(\mathbf{u}, \mathbf{y}') \quad \text{and} \quad \mathbf{V} = \mathrm{var}(\mathbf{y}). \tag{9.2}$$

9.2 BEST PREDICTION (BP)

When $f(u, \mathbf{y})$ is the joint density function of $\mathbf{y}$ and scalar u then, with the predictor of u being denoted by $\tilde{u}$, the mean squared error of prediction is

$$\mathrm{E}[\tilde{u} - u]^2 = \int\int (\tilde{u} - u)^2 f(u, \mathbf{y})\, d\mathbf{y}\, du. \tag{9.3}$$

A generalization of this to a vector of random variables $\mathbf{u}$ is

$$\mathrm{E}[(\tilde{\mathbf{u}} - \mathbf{u})'\mathbf{A}(\tilde{\mathbf{u}} - \mathbf{u})] = \int\int (\tilde{\mathbf{u}} - \mathbf{u})'\mathbf{A}(\tilde{\mathbf{u}} - \mathbf{u}) f(\mathbf{u}, \mathbf{y})\, d\mathbf{y}\, d\mathbf{u}, \tag{9.4}$$

where $\mathbf{A}$ is a positive definite symmetric matrix. Clearly, for $\mathbf{A}$ being scalar and unity, (9.4) is identical to (9.3).

a. The best predictor

Our criterion for deriving a predictor is minimum mean square, i.e., we minimize (9.4). The result is what we call the *best predictor*. Note that "best" here means minimum mean squared error of prediction, which is different from a common meaning of "best" being minimum variance. Because variance is variability around a fixed value and because u in (9.3) is a random variable, (9.3) is not the definition of the variance of u. Thus, for estimating a parameter we use the criterion of minimum variance unbiased, while for predicting the realized value of a random variable we use the criterion of minimum mean square error. Thus, as shown in VC p. 262, from minimizing (9.4) we get

$$\text{best predictor: } \tilde{\mathbf{u}} = \mathrm{BP}(\mathbf{u}) = \mathrm{E}[\mathbf{u}|\mathbf{y}], \tag{9.5}$$

i.e., the best predictor of $\mathbf{u}$ is the conditional mean of $\mathbf{u}$ given $\mathbf{y}$. Details are given in Section 9.8.

Noteworthy features of this result are: (i) it holds for all probability density functions $f(\mathbf{u}, \mathbf{y})$ and (ii) it does not depend on the positive definite symmetric matrix $\mathbf{A}$.

b. Mean and variance properties

First and second moments of the best predictor are important. They are discussed in Cochran (1951) and in Rao (1965, pp. 79 and 220–222) for the case of scalar u. First, the best predictor is unbiased for sampling over $\mathbf{y}$:

$$\mathrm{E}_{\mathbf{y}}[\tilde{\mathbf{u}}] = \mathrm{E}_{\mathbf{y}}\left[\mathrm{E}_{\mathbf{u}|\mathbf{y}}[\mathbf{u}|\mathbf{y}]\right] = \mathrm{E}[\mathbf{u}], \tag{9.6}$$

as detailed in Section S.1 of VC. Note that the meaning of the unbiasedness here is that the expected value of the predictor equals that of the random variable for which it is a predictor. This differs from the usual meaning of unbiasedness when estimating a parameter. In that case unbiasedness means that the expected value of (estimator minus parameter) is zero; e.g., $\mathrm{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathbf{0}$, where $\boldsymbol{\beta}$ is a constant. With prediction, unbiasedness means that the expected value of (predictor minus random variable) is zero; e.g., $\mathrm{E}[\tilde{\mathbf{u}} - \mathbf{u}] = \mathbf{0}$ where $\mathbf{u}$ is a random variable. The former gives $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, whereas the latter gives $\mathrm{E}[\tilde{\mathbf{u}}] = \mathrm{E}[\mathbf{u}]$.

Second, prediction errors $\tilde{\mathbf{u}} - \mathbf{u}$ have a variance-covariance matrix that is the mean value, over sampling on $\mathbf{y}$, of that of $\mathbf{u}|\mathbf{y}$:

$$\mathrm{var}(\tilde{\mathbf{u}} - \mathbf{u}) = \mathrm{E}_{\mathbf{y}}\left[\mathrm{var}(\mathbf{u}|\mathbf{y})\right]. \tag{9.7}$$

Also,

$$\mathrm{cov}(\tilde{\mathbf{u}}, \mathbf{u}') = \mathrm{var}(\tilde{\mathbf{u}}) \qquad \text{and} \qquad \mathrm{cov}(\tilde{\mathbf{u}}, \mathbf{y}') = \mathrm{cov}(\mathbf{u}, \mathbf{y}'). \tag{9.8}$$

c. A correlation property

For scalar u there are two further properties of interest. The first is that the correlation between u and any predictor of it that is a function of $\mathbf{y}$ is maximum for the best predictor, that maximum value being

$$\rho(\tilde{u}, u) = \sigma_{\tilde{u}}/\sigma_u. \tag{9.9}$$

A proof of (9.9) following that of Rao (1973, pp. 265–266) is given in VC, p. 263.

d. Maximizing a mean

A second property of interest concerns the mean of a selected upper fraction of a population of random effects. Making this selection on

the basis of values of $\tilde{u}$ ensures that

$$\text{for that selected fraction, } \mathrm{E}[u] \text{ is maximized.} \tag{9.10}$$

This, too, has a proof available in VC at pp. 264–265.

e. Normality

It is to be emphasized that $\tilde{\mathbf{u}} = \mathrm{E}[\mathbf{u}|\mathbf{y}]$ is a random variable, being a function of $\mathbf{y}$ and unknown parameters. Thus the problem of estimating the best predictor $\tilde{\mathbf{u}}$ remains, and demands some knowledge of the joint density $f(\mathbf{u}|\mathbf{y})$. Should this be normal,

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{C} \\ \mathbf{C}' & \mathbf{V} \end{bmatrix} \right), \tag{9.11}$$

using Section S.2b gives

$$\tilde{\mathbf{u}} = \mathrm{E}[\mathbf{u}|\mathbf{y}] = \boldsymbol{\mu}_u + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y). \tag{9.12}$$

Properties (9.7) through (9.10) of $\tilde{\mathbf{u}}$ still hold. In (9.7) we now have from (9.11) that $\mathrm{var}(\mathbf{u}|\mathbf{y}) = \mathbf{D} - \mathbf{C}\mathbf{V}^{-1}\mathbf{C}'$, so that in (9.7)

$$\mathrm{var}(\tilde{\mathbf{u}} - \mathbf{u}) = \mathbf{D} - \mathbf{C}\mathbf{V}^{-1}\mathbf{C}'. \tag{9.13}$$

And using (9.12) in (9.8) gives

$$\mathrm{cov}(\tilde{\mathbf{u}}, \mathbf{u}') = \mathrm{var}(\tilde{\mathbf{u}}) = \mathbf{C}\mathbf{V}^{-1}\mathbf{C}', \text{ and hence } \rho(\tilde{u}_i, u_i) = \sqrt{\frac{\mathbf{c}_i'\mathbf{V}^{-1}\mathbf{c}_i}{\sigma_{u_i}^2}} \tag{9.14}$$

where $\mathbf{c}_i'$ is the ith row of $\mathbf{C}$.

An estimation problem is clearly visible in these results. The predictor is given in (9.12) but it and its succeeding properties cannot be elucidated without having values for, or estimating, the four parameters $\boldsymbol{\mu}_u$, $\boldsymbol{\mu}_y$, $\mathbf{C}$ and $\mathbf{V}$.

9.3 BEST LINEAR PREDICTION (BLP)

a. BLP(u)

The best predictor (9.5) is not necessarily linear in $\mathbf{y}$. Suppose attention is now confined to predictors of $\mathbf{u}$ that *are* linear in $\mathbf{y}$, of the form

$$\tilde{\mathbf{u}} = \mathbf{a} + \mathbf{B}\mathbf{y} \tag{9.15}$$

for some vector $\mathbf{a}$ and matrix $\mathbf{B}$. Minimizing (9.4) for $\tilde{\mathbf{u}}$ of (9.15), in order to obtain the best linear predictor, leads (without any assumption of normality) to

$$\text{BLP}(\mathbf{u}) = \tilde{\mathbf{u}} = \boldsymbol{\mu}_u + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \tag{9.16}$$

where $\boldsymbol{\mu}_u$, $\boldsymbol{\mu}_y$, $\mathbf{C}$ and $\mathbf{V}$ are as defined in (9.11) but without assuming normality as there. Not only do (9.5) and (9.16) demand no assumption of normality, but additionally important is the fact that they also apply no matter what form $\boldsymbol{\mu}_u$ and $\boldsymbol{\mu}_y$ have. Equations (9.5) and (9.16) apply for all forms of those means.

An immediate observation on (9.16) is that it is identical to (9.12). This shows that the best linear predictor (9.16), derivation of which demands no knowledge of the form of $f(\mathbf{u}, \mathbf{y})$, is *identical* to the best predictor under normality, (9.12). Properties (9.13) and (9.14) therefore apply equally to (9.16) as to (9.12). And, of course, $\text{BLP}(\mathbf{u})$ is unbiased, in the sense described following (9.6), namely that $\text{E}[\tilde{\mathbf{u}}] = \text{E}[\mathbf{u}]$. Problems of estimation of the unknown parameters in (9.16) remain.

b. Example

To illustrate (9.16) we use the beta-binomial model of Section 2.6b wherein we have

$$\begin{aligned} \text{E}[y_{ij}|p_i] &= p_i \\ y_{ij}|p_i &\sim \text{Bernoulli}(p_i) \end{aligned}$$

and

$$\begin{aligned} \text{E}[p_i] &= \frac{\alpha}{\alpha+\beta} \\ \text{var}(p_i) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}; \end{aligned}$$

also

$$\text{E}[y_{ij}] = \frac{\alpha}{\alpha+\beta}$$

and

$$\text{var}(y_{ij}) = \frac{\alpha\beta}{(\alpha+\beta)^2}.$$

Therefore using (9.16)

$$\text{BLP}(p_i) = \text{E}[p_i] + \text{cov}(\bar{y}_{i\cdot}, p_i)\,[\text{var}(\bar{y}_{i\cdot})]^{-1}\,(\bar{y}_{i\cdot} - \text{E}[\bar{y}_{i\cdot}])\,.$$

To derive $\text{cov}(\bar{y}_{i\cdot}, p_i)$ we adapt (2.75) as follows:

$$\begin{aligned}\text{cov}(\bar{y}_{i\cdot}, p_i) &= \text{cov}\left(\text{E}[\bar{y}_{i\cdot}|p_i],\, \text{E}[p_i|p_i]\right) + \text{E}\left[\text{cov}(\bar{y}_{i\cdot}, p_i|p_i)\right] \\ &= \text{cov}(p_i, p_i) + 0 \\ &= \text{var}(p_i).\end{aligned}$$

Therefore

$$\begin{aligned}\text{BLP}(p_i) &= \frac{\alpha}{\alpha+\beta} \\ &\quad + \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\left[\frac{\alpha\beta}{(\alpha+\beta)^2}\right]^{-1}\left(\bar{y}_{i\cdot} - \frac{\alpha}{\alpha+\beta}\right) \\ &= \frac{\alpha+\bar{y}_{i\cdot}}{\alpha+\beta+1}.\end{aligned}$$

c. Derivation

Since we want a predictor $\tilde{\mathbf{u}}$ to be linear (in $\mathbf{y}$) we take $\tilde{\mathbf{u}} = \mathbf{a} + \mathbf{B}\mathbf{y}$ and proceed to derive $\mathbf{a}$ and $\mathbf{B}$ so that $\tilde{\mathbf{u}} = \mathbf{a} + \mathbf{B}\mathbf{y}$ is best, meaning that it has minimum mean squared error of prediction. Thus we want to minimize the left-hand side of (9.4), which now gets written as

$$\begin{aligned}q &= \text{E}\left[(\mathbf{a}+\mathbf{B}\mathbf{y}-\mathbf{u})'\mathbf{A}(\mathbf{a}+\mathbf{B}\mathbf{y}-\mathbf{u})\right] \\ &= \text{E}\left[\mathbf{a}'\mathbf{A}\mathbf{a} + 2\mathbf{a}'\mathbf{A}(\mathbf{B}\mathbf{y}-\mathbf{u}) + (\mathbf{B}\mathbf{y}-\mathbf{u})'\mathbf{A}(\mathbf{B}\mathbf{y}-\mathbf{u})\right]. \quad (9.17)\end{aligned}$$

Equating $\partial q/\partial \mathbf{a}$ to $\mathbf{0}$ gives

$$2\mathbf{A}(\mathbf{a} + \text{E}[\mathbf{B}\mathbf{y} - \mathbf{u}]) = \mathbf{0}$$

and so

$$\mathbf{a} = -\text{E}[\mathbf{B}\mathbf{y} - \mathbf{u}] = -(\mathbf{B}\boldsymbol{\mu}_y - \boldsymbol{\mu}_u).$$

Substituting $\mathbf{a}$ into (9.17) gives

$$\begin{aligned}q &= -\text{E}[\mathbf{B}\mathbf{y}-\mathbf{u}]'\mathbf{A}\text{E}[\mathbf{B}\mathbf{y}-\mathbf{u}] + \text{E}\left[(\mathbf{B}\mathbf{y}-\mathbf{u})'\mathbf{A}(\mathbf{B}\mathbf{y}-\mathbf{u})\right] \\ &= \text{tr}\{\mathbf{A}\,\text{var}(\mathbf{B}\mathbf{y}-\mathbf{u})\}, \qquad \text{based on the normality,} \\ &= \text{tr}\{\mathbf{A}(\mathbf{B}\mathbf{V}\mathbf{B}' + \mathbf{D} - \mathbf{B}\mathbf{C}' - \mathbf{C}\mathbf{B}')\}. \quad (9.18)\end{aligned}$$

We wish to minimize this with respect to $\mathbf{B}$. To do this, ignore $\mathbf{A}$ and $\mathbf{D}$ (because they do not involve $\mathbf{B}$), and define $\mathbf{b}_i'$ and $\mathbf{c}_j'$ as the ith

and jth rows of $\mathbf{B}$ and $\mathbf{C}$, respectively. Then the (i,j)th element of $\mathbf{BVB}' - \mathbf{BC}' - \mathbf{CB}'$ is

$$\varphi_{ij} = \mathbf{b}_i'\mathbf{V}\mathbf{b}_j - \mathbf{b}_i'\mathbf{c}_j - \mathbf{c}_i'\mathbf{b}_j.$$

Thus to minimize this element with respect to $\mathbf{b}_i'$ and $\mathbf{b}_j'$

$$\frac{\partial\varphi_{ij}}{\partial\mathbf{b}_i} = \mathbf{V}\mathbf{b}_j - \mathbf{c}_j \qquad \text{and} \qquad \frac{\partial\varphi_{ij}}{\partial\mathbf{b}_j} = \mathbf{V}\mathbf{b}_i - \mathbf{c}_i.$$

Therefore we take the minimizing form of $\mathbf{B}$ as $\mathbf{VB}' = \mathbf{C}'$ and so $\mathbf{B} = \mathbf{CV}^{-1}$. Thus

$$\begin{aligned} \text{BLP}(\mathbf{u}) &= \mathbf{a} + \mathbf{B}\mathbf{y} = \boldsymbol{\mu}_u + \mathbf{B}(\mathbf{y} - \boldsymbol{\mu}_y) \\ &= \boldsymbol{\mu}_u + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y). \end{aligned} \tag{9.19}$$

d. Ranking

In establishing, as observed in (9.10), that selection on the basis of the best predictor $\tilde{u}$ maximizes $\mathrm{E}[u]$ of the selected proportion of the population, Cochran's (1951) development implicitly relies on each scalar $\tilde{u}$ having the same variance and being derived from a $\mathbf{y}$ that is independent of other $\mathbf{y}$s. Sampling is over repeated samples of u (scalar) and $\mathbf{y}$. However, these conditions are not met for the elements of $\tilde{\mathbf{u}}$ derived in (9.12). Each such element is derived from the whole vector $\mathbf{y}$, their variances are not equal, and the elements of $\mathbf{y}$ used in one element of $\tilde{\mathbf{u}}$ are not necessarily independent of those used for another element of $\tilde{\mathbf{u}}$. Maximizing the probability of correctly ranking individuals on the basis of elements in $\tilde{\mathbf{u}}$ is therefore not assured. In place of this there is a property about pairwise ranking.

Having predicted the (unobservable) realized values of the random variables in the data, a salient problem that is often of great importance is this: How does the ranking on predicted values compare with the ranking on the true (realized but unobservable) values? Henderson (1963) and Searle (1974) show, under certain conditions (including normality), that the probability that predictors of u_i and u_j have the same pairwise ranking as u_i and u_j is maximized when those predictors are elements of BLP($\mathbf{u}$) of (23); i.e., the probability $\mathrm{P}\{\tilde{u}_i - \tilde{u}_j \gtrless 0 | u_i - u_j \gtrless 0\}$ is maximized. Portnoy (1982) extends this to a special components-of-variance model, for which he shows that ranking all the u_is of $\mathbf{u}$ in the same order as the $\tilde{u}_j$ (the best linear predictors) rank themselves

does maximize the probability of ranking the u_is correctly. He does, however, go on to show that in models more general than variance components models, there can be predictors that lead to higher values of this probability than do the best linear predictors, which are elements of the vector $\text{BLP}(\mathbf{u}) = \boldsymbol{\mu}_u + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$.

9.4 LINEAR MIXED MODEL PREDICTION (BLUP)

The preceding discussion is concerned with the prediction of random variables. Through maximizing the probability of correct ranking, the predictors are appropriate values upon which to base selection; e.g., in genetics, selecting the animals with highest predictions to be parents of the next generation. Consideration is now given to linear mixed model prediction, corresponding to mixed models in which some factors are fixed and others are random.

a. BLUE($\mathbf{X}\boldsymbol{\beta}$)

In Section 6.3 we derived at equation (6.20)

$$\text{ML}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{9.20}$$

By concentrating attention on estimating $\mathbf{X}\boldsymbol{\beta}$ rather than $\boldsymbol{\beta}$ we achieved the invariance of $\mathbf{X}\boldsymbol{\beta}^0$ to $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$ compared to the non-invariance of $\boldsymbol{\beta}^0$. Deriving $\text{ML}(\mathbf{X}\boldsymbol{\beta})$ of (9.20) relied upon assuming normality of the data vector $\mathbf{y}$. But that same estimator $\mathbf{X}\boldsymbol{\beta}^0$ can also be derived, without requiring normality, as the *best linear unbiased estimator* (BLUE) of $\mathbf{X}\boldsymbol{\beta}$:

$$\text{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}^0. \tag{9.21}$$

Best in this context means that of all linear (in $\mathbf{y}$) unbiased estimators of $\mathbf{X}\boldsymbol{\beta}$, the "best", i.e., the $\text{BLUE}(\mathbf{X}\boldsymbol{\beta})$, is the one with the smallest variance. This is established by taking the estimator to have the form $\boldsymbol{\lambda}'\mathbf{y}$ and deriving $\boldsymbol{\lambda}$ so that $\boldsymbol{\lambda}'\mathbf{y}$ is both unbiased for $\mathbf{t}'\mathbf{X}\boldsymbol{\beta}$ (for given $\mathbf{t}'$) and also so that $\text{var}(\boldsymbol{\lambda}'\mathbf{y})$ is minimized. With $2\mathbf{m}$ being a vector of Lagrange multipliers, this leads to wanting

$$\boldsymbol{\lambda}'\mathbf{y} = \mathbf{X}'\mathbf{t} \tag{9.22}$$

and needing to minimize

$$\boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda} + 2\mathbf{m}'(\mathbf{X}'\boldsymbol{\lambda} - \mathbf{X}'\mathbf{t}) \tag{9.23}$$

with respect to $\boldsymbol{\lambda}$. The resulting equations for $\boldsymbol{\lambda}$ and $\mathbf{m}$ (which is of little interest) are

$$\mathbf{V}\boldsymbol{\lambda} + \mathbf{X}\mathbf{m} = \mathbf{0} \tag{9.24}$$

$$\mathbf{X}'\boldsymbol{\lambda} = \mathbf{X}'\mathbf{t} \tag{9.25}$$

or, in matrix form

$$\begin{bmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ (\mathbf{t}'\mathbf{X})' \end{bmatrix}, \tag{9.26}$$

an example of equation (1) of Hayes and Haslett (1999). The solution for $\boldsymbol{\lambda}$ is what leads to BLUE$(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}^0$ of (9.21).

b. BLUP$(\mathbf{t}'\mathbf{X}\boldsymbol{\beta} + \mathbf{s}'\mathbf{u})$

The counterpart of BLUE$(\mathbf{t}'\mathbf{X}\boldsymbol{\beta})$ in an LM or GLM is BLUP$(\mathbf{t}'\mathbf{X}\boldsymbol{\beta} + \mathbf{s}'\mathbf{u})$ in an LMM or GLMM where $\mathbf{u}$ is a vector of random effects and $\mathbf{t}'$ and $\mathbf{s}'$ are known vectors. And BLUP is *best linear unbiased predictor.*

Akin to the derivation of BLUE, we seek $\boldsymbol{\lambda}$ for $\boldsymbol{\lambda}'\mathbf{y}$ to be unbiased for $\mathbf{t}'\mathbf{X}\boldsymbol{\beta} + \mathbf{s}'\mathbf{u}$; and in taking $\mathrm{E}[\mathbf{u}] = \mathbf{0}$, which is customary, this unbiasedness leads, just as in (9.22), to

$$\mathbf{X}'\boldsymbol{\lambda} = \mathbf{X}'\mathbf{t}.$$

We additionally seek $\boldsymbol{\lambda}$ so as to minimize the variance of the prediction error of $[\boldsymbol{\lambda}'\mathbf{y} - (\mathbf{t}'\mathbf{X}\boldsymbol{\beta} + \mathbf{s}'\mathbf{u})]$. This is done by minimizing

$$\begin{aligned} \theta &= \mathrm{var}\,[\boldsymbol{\lambda}'\mathbf{y} - (\mathbf{t}'\mathbf{X}\boldsymbol{\beta} + \mathbf{s}'\mathbf{u})] + 2\mathbf{m}'(\mathbf{X}'\boldsymbol{\lambda} - \mathbf{X}'\mathbf{t}) \\ &= \boldsymbol{\lambda}'\mathbf{V}\boldsymbol{\lambda} + \mathbf{s}'\mathbf{D}\mathbf{s} - 2\boldsymbol{\lambda}'\mathbf{C}\mathbf{s} + 2\mathbf{m}'(\mathbf{X}'\boldsymbol{\lambda} - \mathbf{X}'\mathbf{t}) \end{aligned}$$

with respect to $\boldsymbol{\lambda}$ and $\mathbf{m}$. This gives equations

$$\begin{bmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{C}\mathbf{s} \\ (\mathbf{t}'\mathbf{X})' \end{bmatrix}. \tag{9.27}$$

Solving for $\boldsymbol{\lambda}$ yields $\boldsymbol{\lambda}'\mathbf{y}$ as

$$\mathrm{BLUP}(\mathbf{t}'\mathbf{X}\boldsymbol{\beta} + \mathbf{s}'\mathbf{u}) = \mathbf{t}'\mathbf{X}\boldsymbol{\beta}^0 + \mathbf{s}'\mathbf{C}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0) \tag{9.28}$$

where $\mathbf{X}\boldsymbol{\beta}^0$ is the same BLUE$(\mathbf{X}\boldsymbol{\beta})$ of (9.21).

Special cases of (9.28) are (i) for $\mathbf{s} = \mathbf{0}$ and $\mathbf{t}'$ taking successive rows of $\mathbf{I}$, then BLUP$(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}^0$; and (ii) for $\mathbf{t}' = \mathbf{0}$ and $\mathbf{s}'$ being successive rows of $\mathbf{I}$, BLUP$(\mathbf{u}) = \mathbf{C}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)$. And a very familiar special case is that of $\mathbf{C} = \mathbf{Z}\mathbf{D}$, giving BLUP$(\mathbf{u}) = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)$ of (6.43).

c. Two variances

The variance and covariances of some eight variants of (9.28) are given in VC p. 272. We show just two of them here.

Let

$$w = \mathbf{t}'\mathbf{X}\boldsymbol{\beta} + \mathbf{s}'\mathbf{u}$$

with

$$\tilde{w} = \mathbf{t}'\mathbf{X}\boldsymbol{\beta}^0 + \mathbf{s}'\mathbf{C}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0).$$

Then for

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \tag{9.29}$$

$$\text{var}(\tilde{w}) = \mathbf{t}'\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{t} + \mathbf{s}'\mathbf{C}\mathbf{P}\mathbf{C}'\mathbf{s} \tag{9.30}$$

$$\text{var}(\tilde{w} - w) = \text{var}(\tilde{\mathbf{w}}) - 2\mathbf{t}'\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{C}'\mathbf{s}. \tag{9.31}$$

d. Other derivations

Although Henderson had developed and used (9.28) in dairy cow selection programs prior to publication in Henderson (1963), in the statistical literature an early version can be found in Goldberger (1962). His equation (3.12) has $\mathbf{X}$ of full column rank, and with its $\boldsymbol{\Omega} = \mathbf{V}$, $\mathbf{x}'_* = \mathbf{t}'\mathbf{X}$ and its $\mathbf{w}' = \mathbf{s}'\mathbf{C}$ (3.12) is a special case of (9.28).

Of the numerous ways for deriving (9.28), five are detailed in VC, pp. 271–275. One is primarily algebraic as in the derivation of (9.18); another simplifies $\mathrm{E}[\mathbf{u}|\mathbf{y}_c]$ for $\mathbf{y}_c$ being $\mathbf{y}$ corrected for fixed effects $\boldsymbol{\beta}$; a third starts with assuming that $\mathbf{w}$ is linear in $\mathbf{y}$, such as $\mathbf{a} + \mathbf{B}\mathbf{y}$, another procedure is based on partitioning $\mathbf{y}$ as $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0)$ and $\mathbf{X}\boldsymbol{\beta}^0$; and finally there is a Bayes method.

9.5 REQUIRED ASSUMPTIONS

It is interesting to note that BP, BLP and BLUP do not all require the same assumptions. In some general sense BP requires more assumptions than BLP which in turn requires more than BLUP. For $\mathrm{BP}(\mathbf{u}) = \mathrm{E}[\mathbf{u}|\mathbf{y}]$ one needs to know the distribution of $\mathbf{u}|\mathbf{y}$. But BLP demands knowing only first and second moments of $\mathbf{u}$ and $\mathbf{y}$. BLUP requires knowing only $\mathbf{V} = \text{var}(\mathbf{y})$ and $\mathbf{C} = \text{cov}(\mathbf{u}, \mathbf{y}')$, but first moments are not needed, with fixed effects $\boldsymbol{\beta}$ being estimated through using $\mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. In not one of BP, BLP or BLUP is normality needed.

9.6 ESTIMATED BEST PREDICTION

An expression for BP($\mathbf{u}$) more usable in practice than E[$\mathbf{u}|\mathbf{y}$] demands knowing the distribution of $\mathbf{y}$ and $\mathbf{u}$. And even then, numerical values are needed for the parameters of that distribution. This need is clearly demonstrated by considering

$$\text{BLP}(\mathbf{u}) = \boldsymbol{\mu}_u + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y). \tag{9.32}$$

In many situations of having data on which one wants to use (9.32), numerical values for $\boldsymbol{\mu}_u$, $\mathbf{C}$, $\mathbf{V}$ and $\boldsymbol{\mu}_y$ may not be available. So in order to use (9.32) estimates of these parameters need to be found. Often this entails deriving such estimates from the data being used to obtain estimates of (9.32).

Just as one can simultaneously and optimally estimate μ and σ^2 from data $\mathbf{x} \sim \mathcal{N}(\mu\mathbf{1},\ \sigma^2\mathbf{I})$, one would ideally like to optimally estimate $\tilde{\mathbf{u}}$ of (9.32) along with $\boldsymbol{\mu}_u$, $\mathbf{C}$, $\mathbf{V}$ and $\boldsymbol{\mu}_y$. But this is seldom (if ever) feasible. The usual procedure, therefore, is to estimate $\boldsymbol{\mu}_u$, $\mathbf{C}$, $\mathbf{V}$ and $\boldsymbol{\mu}_y$ and replace those parameters in (9.32) by their estimates. No matter what estimation methods are used for the parameters, denote the resulting estimates by $\hat{\boldsymbol{\mu}}_u$, $\hat{\mathbf{C}}$, $\hat{\mathbf{V}}$ and $\hat{\boldsymbol{\mu}}_y$. Then in denoting BLP($\mathbf{u}$) as $\tilde{\mathbf{u}}$,

$$\text{BLP}(\mathbf{u}) \equiv \tilde{\mathbf{u}} = \boldsymbol{\mu}_u + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$$

we can have a calculated value

$$\hat{\mathbf{u}} = \hat{\boldsymbol{\mu}}_u + \hat{\mathbf{C}}\hat{\mathbf{V}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}_y)$$

as an estimated BLP($\mathbf{u}$). Similarly for

$$\begin{aligned} \text{BLUP}(\mathbf{u}) &= \mathbf{u}^0 = \boldsymbol{\mu}_u + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0), \\ \text{EBLUP}(\mathbf{u}) &= \hat{\mathbf{u}}^0 = \hat{\boldsymbol{\mu}}_u + \hat{\mathbf{C}}\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^0) \end{aligned} \tag{9.33}$$

is an estimated BLUP, with

$$\mathbf{X}\hat{\boldsymbol{\beta}}^0 = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}.$$

Note that these estimated predictors have been derived from a purely practical viewpoint: Estimate parameters of the distribution of $\mathbf{u}$ and $\mathbf{y}$ and in the predictors simply replace the parameters by those estimates. In doing this no statistical rationale such as minimum variance has

been invoked. The estimated predictors have just been set up in what seems like an "obvious" manner. Nevertheless, $\hat{\mathbf{u}}^0$ and $\mathbf{X}\hat{\boldsymbol{\beta}}^0$ are (with $\hat{\mathbf{V}}$ being the MLE of $\mathbf{V}$) ML estimators of $\mathbf{u}^0$ and $\mathbf{X}\boldsymbol{\beta}^0$, respectively. Their properties, such as mean and variance (let alone distribution) are largely intractable. This is so if for no other reason that if the estimated parameters are based on $\mathbf{y}$, then the parameter estimates are correlated, not only with each other but also with $\mathbf{y}$. And these correlations will have to be taken into account when seeking moments of the estimated predicted values. And doing that is not easy. The best that has been done in the research literature so far is various attempts at developing approximations. Some of those results are dealt with in Chapter 6, Linear Mixed Models. Analogous results for generalized linear models, or nonlinear models would be even more complicated than what is in Chapter 6. Kackar and Harville (1984) and Prasad and Rao (1990) consider some of these complications for the estimated BLUP($\mathbf{u}$), namely $\hat{\mathbf{u}}^0$ of (9.33).

9.7 HENDERSON'S MIXED MODEL EQUATIONS

a. Origin

A set of equations developed by Henderson in Henderson et al. (1959), which simultaneously yield BLUE($\mathbf{X}\boldsymbol{\beta}$) and BLUP($\mathbf{u}$) in the LMM with model equation $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, have come to be known as the *mixed model equations* (MMEs). For the joint density of $\mathbf{u}$ and $\mathbf{y}$ being normal, as in (9.11), with $\mathbf{C} = \mathbf{D}\mathbf{Z}'$, $\mathrm{var}(\mathbf{y}|\mathbf{u}) = \mathbf{R}$, and $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$, this density is

$$\begin{aligned} f(\mathbf{y},\mathbf{u}) &= f(\mathbf{y}|\mathbf{u})f(\mathbf{u}) \qquad (9.34) \\ &= \frac{\exp\{-\frac{1}{2}[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) + \mathbf{u}'\mathbf{D}^{-1}\mathbf{u}]\}}{(2\pi)^{\frac{1}{2}(N+q.)}|\mathbf{R}|^{\frac{1}{2}}|\mathbf{D}|^{\frac{1}{2}}}, \end{aligned}$$

where $q.$ is the number of columns in $\mathbf{Z}$ (i.e., the number of random effects in the model for the data).

Henderson's approach was to maximize (9.34) with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$, which results in

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \qquad (9.35)$$

These are the MMEs. Their form is worthy of note: Without the $\mathbf{D}^{-1}$ in the lower right-hand submatrix of the matrix on the left, they would be the ML equations for the model treated as if $\mathbf{u}$ represented fixed effects, rather than random effects.

b. Solutions

After a minor amount of algebra (see E 9.7) it will be found that the solutions to (9.35) are

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

and

$$\tilde{\mathbf{u}} = \mathbf{u}^0 = \text{BLUP}(\mathbf{u}) = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0).$$

The MMEs not only represent a procedure for calculating a $\boldsymbol{\beta}^0$ and $\tilde{\mathbf{u}}$, but are also computationally more economical than the ML equations which lead to $\mathbf{X}\boldsymbol{\beta}^0$. Those equations require inversion of $\mathbf{V}$ of order N. But the MMEs need inversion of a matrix of order only $p + q.$, the total number of levels of fixed and random effects in the data. And this number is usually much smaller than the N, the number of observations. True, the MMEs do require inversion of both $\mathbf{R}$ and $\mathbf{D}$, but for variance components linear models these are often diagonal, which makes these inversions easy.

c. Use in ML estimation of variance components

An interesting feature of the MMEs is that parts of them can be used for setting up iterative procedures for calculating ML and REML estimates of variance components in variance components models. Derivation of these iterative procedures is shown in great detail in VC, pp. 277–285. Unfortunately the detailed algebra for these derivations is tedious and lengthy, and so is not given here. The interested reader can go to the VC reference. Presented here are the main results from that reference.

– i. *ML estimation*

Define

$$\mathbf{W} = (\mathbf{I} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{D})^{-1} = \{\mathbf{W}_{ij}\}_{i,j=1}^{r};$$

q_i = number of levels of the ith random effect;

and $\sigma_i^{2(m)}$ is the calculated value of σ_i^2 after the mth round of iteration. The superscript parenthesized m is used throughout to indicate iteration number. Then for $i = 1, \ldots, r$ the ML equations (6.60) can be reduced to the iterations

$$\sigma_i^{2(m+1)} = \frac{\tilde{\mathbf{u}}_i^{\prime(m)}\tilde{\mathbf{u}}_i^{(m)} + \sigma_i^{2(m)}\mathrm{tr}\left[\mathbf{W}_{ii}^{(m)}\right]}{q_i} \tag{9.36}$$

or

$$\sigma_i^{2(m+1)} = \frac{\tilde{\mathbf{u}}_i^{\prime(m)}\tilde{\mathbf{u}}_i^{\prime(m)}}{q_i - \mathrm{tr}\left[\mathbf{W}_{ii}^{(m)}\right]}. \tag{9.37}$$

each along with

$$\sigma_i^{2(m+1)} = \mathbf{y}'\left[(\mathbf{y} - \mathbf{X}\beta^{0(m)} - \mathbf{Z}\tilde{\mathbf{u}}^{(m)})\right]/N. \tag{9.38}$$

– ii. *REML estimation*

The iterative procedure is essentially the same for REML as it is for ML but with the following changes. Define

$$\mathbf{S} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{R}^{-1} \tag{9.39}$$

and use $\mathbf{S}$ instead of $\mathbf{R}^{-1}$ in $\mathbf{W}$. Then, with $N - \mathrm{rank}(\mathbf{X})$ replacing N as the denominator of (9.38), (9.36) through (9.38) are an iterative procedure for REML estimation of variance components. Making these changes in (9.36), for example, gives

$$\sigma_{i(\mathrm{REML})}^{2(m+1)} = \frac{\tilde{\mathbf{u}}_i^{\prime(m)}\tilde{\mathbf{u}}_i^{(m)} + \sigma_{i(\mathrm{REML})}^{2(m)}\mathrm{tr}\left[\mathbf{W}_{ii(\mathrm{REML})}^{(m)}\right]}{q_i}$$

for $\mathbf{W}_{ii(\mathrm{REML})}$ being the ith diagonal submatrix of

$$\mathbf{W}_{(\mathrm{REML})} = (\mathbf{I} + \mathbf{Z}'\mathbf{S}\mathbf{Z}\mathbf{D})^{-1} = \left\{\mathbf{W}_{ij(\mathrm{REML})}\right\}_{i,j=1}^{r}.$$

VC, pp. 277–285 not only gives details of deriving the preceding results, but also derives the information matrix for both the ML and REML procedures.

9.8 APPENDIX

a. Verification of (9.5)

In the mean square on the left-hand side of (9.4), to $\tilde{\mathbf{u}} - \mathbf{u}$ add and subtract E[$\mathbf{u}|\mathbf{y}$], which, for convenience, will be denoted by $\mathbf{u}_0$; i.e.,

with

$$\mathbf{u}_0 \equiv \mathrm{E}[\mathbf{u}|\mathbf{y}],$$

$$\mathrm{E}[(\tilde{\mathbf{u}}-\mathbf{u})'\mathbf{A}(\tilde{\mathbf{u}}-\mathbf{u})] = \mathrm{E}[(\tilde{\mathbf{u}}-\mathbf{u}_0+\mathbf{u}_0-\mathbf{u})'\mathbf{A}(\tilde{\mathbf{u}}-\mathbf{u}_0+\mathbf{u}_0-\mathbf{u})]. \quad (9.40)$$

To choose a $\tilde{\mathbf{u}}$ that minimizes (9.40), note that in expanding it, the last term, $\mathrm{E}[(\mathbf{u}_0-\mathbf{u})'\mathbf{A}(\mathbf{u}_0-\mathbf{u})]$, does not involve $\tilde{\mathbf{u}}$. And in the cross-product term, using iterated expectation with E_y representing expectation with respect to the distribution of y,

$$\begin{aligned}\mathrm{E}[(\tilde{\mathbf{u}}-\mathbf{u}_0)'\mathbf{A}(\mathbf{u}_0-\mathbf{u})] &= \mathrm{E}_y\left[\mathrm{E}_{u|y}[(\tilde{\mathbf{u}}-\mathbf{u}_0)'\mathbf{A}(\mathbf{u}_0-\mathbf{u})|\mathbf{y}]\right] \\ &= \mathrm{E}_y[(\tilde{\mathbf{u}}-\mathbf{u}_0)'\mathbf{A}(\mathbf{u}_0-\mathbf{u}_0)] = 0\end{aligned}$$

since, for a given $\mathbf{y}$, only $\mathbf{u}$ is not fixed and has $\mathrm{E}_{u|y}[\mathbf{u}|\mathbf{y}] = \mathbf{u}_0$. Therefore

$$\mathrm{E}[(\tilde{\mathbf{u}}-\mathbf{u})'\mathbf{A}(\tilde{\mathbf{u}}-\mathbf{u})] = \mathrm{E}[(\tilde{\mathbf{u}}-\mathbf{u}_0)'\mathbf{A}(\tilde{\mathbf{u}}-\mathbf{u}_0)] + \text{terms without } \tilde{\mathbf{u}}.$$

Since $\mathrm{E}[(\tilde{\mathbf{u}}-\mathbf{u}_0)'\mathbf{A}(\tilde{\mathbf{u}}-\mathbf{u}_0)]$ must be non-negative, it is minimized by choosing $\tilde{\mathbf{u}} = \mathbf{u}_0$; i.e., the best predictor is $\tilde{\mathbf{u}} = \mathbf{u}_0 = \mathrm{E}[\mathbf{u}|\mathbf{y}]$. Thus the problem of predicting a random variable is simply that of predicting its conditional mean.

b. Verification of (9.7) and (9.8)

Deriving (9.7) comes from (1.14) with y_{ij} replaced by $\tilde{\mathbf{u}}-\mathbf{u}$ and a_i by $\mathbf{y}$. The two results in (9.8) are established by using (1.16) with $\mathbf{y}$, $\mathbf{w}$ and $\mathbf{u}$ replaced, respectively, by $\tilde{\mathbf{u}}, \mathbf{u}'$, and $\mathbf{y}$. This gives

$$\mathrm{cov}(\tilde{\mathbf{u}}, \mathbf{u}') = \mathrm{cov}\left(\mathrm{E}[\tilde{\mathbf{u}}|\mathbf{y}], \mathrm{E}[\mathbf{u}'|\mathbf{y}]\right) + \mathrm{E}_y\left[\mathrm{cov}(\tilde{\mathbf{u}}, \mathbf{u}'|\mathbf{y})\right].$$

The second term here involves the covariance of $\mathbf{u}$ (conditional on $\mathbf{y}$) with its mean $\mathrm{E}[\mathbf{u}|\mathbf{y}]$. It is therefore zero. Hence

$$\mathrm{cov}(\tilde{\mathbf{u}}, \mathbf{u}') = \mathrm{cov}(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}') = \mathrm{var}(\tilde{\mathbf{u}}),$$

which is the first result in (9.8). Likewise, for the second result we start with

$$\mathrm{cov}(\mathbf{u}, \mathbf{y}') = \mathrm{cov}\left(\mathrm{E}[\mathbf{u}|\mathbf{y}], \mathrm{E}[\mathbf{y}'|\mathbf{y}]\right) + \mathrm{E}_y\left[\mathrm{cov}(\mathbf{u}, \mathbf{y}'|\mathbf{y})\right].$$

In the second term, the covariance is of $\mathbf{u}$ with $\mathbf{y}'$, which is constant conditional on $\mathbf{y}$. Therefore it is zero and so

$$\mathrm{cov}(\mathbf{u}, \mathbf{y}') = \mathrm{cov}(\tilde{\mathbf{u}}, \mathbf{y}').$$

Thus (9.8) is established.

9.9 EXERCISES

E 9.1 Suppose $y_{ij}|\lambda_i \sim$ indep. Poisson(λ_i) and $\lambda_i \sim$ Gamma(r, β); $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$. Find the BP and the BLP of λ_i.

E 9.2 Establish (9.7) and (9.8).

E 9.3 For the random effects 1-way classification model with unbalanced data of Section 2.2c, use (9.16) to derive

$$\text{BLP}(a_i) = \frac{n_i\sigma_a^2}{\sigma^2 + n_i\sigma_a^2}(\bar{y}_{i\cdot} - \mu).$$

Under what condition is this also BP(a_i) of Section 2.4b–ii?

E 9.4 Use (9.24) and (9.25) to derive BLUE$(\mathbf{X}\boldsymbol{\beta})$.

E 9.5 Derive (9.28) using a partitioned inverse in (9.27).

E 9.6 Derive (9.30) and (9.31).

E 9.7 Derive (9.35) and use it to obtain the solutions in Section 9.7b.

E 9.8 Prove

(a) $\mathbf{V}^{-1} = (\mathbf{ZDZ}' + \mathbf{R})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$.

(b) $\mathbf{DZ}'\mathbf{V}^{-1} = (\mathbf{Z}'\mathbf{R}^{-1} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$.

Chapter 10

COMPUTING

10.1 INTRODUCTION

A common theme throughout this book has been the difficulty of calculation of likelihood-based inference. As noted in Chapter 8, computing the likelihood itself is often difficult for GLMMs, requiring the calculation of high-dimensional integrals. In the case of the leaf blight example of Section 8.8, the integral is more than 200 dimensions. Unfortunately, the current state of software does not include any well-tested and general-purpose routines for performing such calculations. For certain subclasses, e.g., linear mixed models, they do exist, but not for the full generality of GLMMs. In this chapter we identify some of the common methods used for likelihood calculation and maximization and briefly describe and give references to some current research topics in computing for GLMMs.

10.2 COMPUTING ML ESTIMATES FOR LMMs

We first consider computing ML estimates for linear mixed models since their structure simplifies the calculations somewhat.

a. The EM algorithm

The EM algorithm (McLachlan and Krishnan, 1996) is an iterative algorithm for calculating ML (or REML) estimates, its name standing for *expectation maximization*: It alternates between calculating conditional *expected* values and *maximizing* simplified likelihoods. It generates only estimates and requires extra computations (e.g., Louis, 1982)

for obtaining variance estimates.

The EM algorithm is designed for situations where the recognition or invention of "missing" data simplifies the maximum likelihood calculations. Starting with initial guesses for the model's parameters, the EM algorithm typically fills in the missing data by calculating conditional expected values (given the observed or *incomplete* data $\mathbf{y}$) of the sufficient statistics. The combination of the observed data and the missing data is usually referred to as *complete data*.

For estimating variance components in LMMs with $\mathrm{E}[\mathbf{y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \sum_i \mathbf{Z}_i\mathbf{u}_i$, the missing data are typically taken to be the realized values of the random effects. Knowledge of the random effects simplifies the calculations from two viewpoints. First, if they were known we could simply estimate $\sigma_i^2 = \mathrm{var}(u_i)$ as:

$$\hat{\sigma}_i^2 = \mathbf{u}_i'\mathbf{u}_i/q_i, \tag{10.1}$$

where q_i is the dimension of $\mathbf{u}_i$. This the ML estimator under the assumption that $\mathbf{u}_i \sim \mathcal{N}(0, \mathbf{I}\sigma_i^2)$. Second, if they were known, we could subtract them from $\mathbf{y}$ leaving the resulting data independent and following a linear model, to which we could apply ordinary least squares:

$$\mathbf{y} - \sum_i \mathbf{Z}_i\mathbf{u}_i \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2). \tag{10.2}$$

All this is very nice, but in reality we do not know the realized values of the $\mathbf{u}_i$. To counter this, the EM algorithm calculates values to use in place of the unknown realized values (the missing data) in order to effect estimation. The conditional expected values of the $\mathbf{u}_i'\mathbf{u}_i$ are used in place of the $\mathbf{u}_i'\mathbf{u}_i$ in (10.1) and the conditional expected values of the $\mathbf{u}_i$ are used in place of the $\mathbf{u}_i$ in (10.2) to form improved estimates of the parameters. This is the maximization step since those equations represent ML estimation from the complete data. The new estimates are then used to recalculate conditional expected values; and so on. This iterative scheme is used until convergence.

Details of applying the preceding ideas to the linear mixed model are given in Searle et al. (1992, Sec. 8.3). They result in three procedures, one for ML, a variation of that, and one for REML. We now describe the procedures, in each case by indicating iteratively computed values of (functions of) parameters using parenthesized superscripts. Thus $\sigma_i^{2(m)}$ is the computed value of σ_i^2 after the mth round of iteration; and $\mathbf{V}^{-1(m)}$ is $\mathbf{V}^{-1}$ with σ_i^2 in $\mathbf{V}$ replaced by $\sigma_i^{2(m)}$ for $i = 0, 1, \ldots, r$.

– i. ***EM for ML***

Step 0. Set $m = 0$, and choose starting values $\boldsymbol{\beta}^{(0)}$ and $\sigma_i^{2(0)}$.
Step 1. Calculate

$$\mathbf{X}\boldsymbol{\beta}^{(m+1)} = \mathbf{X}\boldsymbol{\beta}^{(m)} + \sigma_0^{2(m)}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m)}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(m)}). \quad (10.3)$$

and, for $\mathbf{r}^{(m)} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(m)}$,

$$\sigma_i^{2(m+1)} = \sigma_i^{2(m)} + (\sigma_i^{4(m)}/q_i)[\mathbf{r}^{(m)'}\mathbf{V}^{-1(m)}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1(m)}\mathbf{r}^{(m)} - \mathrm{tr}(\mathbf{Z}_i'\mathbf{V}^{-1(m)}\mathbf{Z}_i)]. \quad (10.4)$$

Step 2. If convergence is reached, set $\hat{\sigma}_i^2 = \sigma_i^{(m+1)}$ and $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta}^{(m+1)}$; otherwise increase m by 1 and return to step 1.

– ii. ***EM (a variant) for ML***

Step 0. Set $m = 0$, and choose starting values $\sigma_i^{2(0)}$.
Step 1. Calculate

$$\sigma_i^{2(m+1)} = \sigma_i^{2(m)} + (\sigma_i^{4(m)}/q_i)[\mathbf{y}'\mathbf{P}^{(m)}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{y} - \mathrm{tr}(\mathbf{Z}_i'\mathbf{V}^{-1(m)}\mathbf{Z}_i)]. \quad (10.5)$$

Step 2. If convergence is reached, set $\hat{\sigma}_i^2 = \sigma_i^{(m+1)}$ and then calculate $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1(m+1)}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m+1)}\mathbf{y}$; otherwise increase m by 1 and return to step 1.

– iii. ***EM for REML***

Step 0. Set $m = 0$, and choose starting values $\sigma_i^{2(0)}$.
Step 1. Calculate

$$\sigma_i^{2(m+1)} = \sigma_i^{2(m)} + (\sigma_i^{4(m)}/q_i)[\mathbf{y}'\mathbf{P}^{(m)}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{y} - \mathrm{tr}(\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{Z}_i)]. \quad (10.6)$$

Step 2. If convergence is reached, set $\hat{\sigma}_i^2 = \sigma_i^{(m+1)}$; otherwise increase m by 1 and return to step 1.

b. Using E[u|y]

An alternative to EM is to begin with the REML equations of (6.67):

$$\left\{ {}_c \operatorname{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') \right\}_{i=0}^{r} = \left\{ {}_c \, \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} \right\}_{i=0}^{r} \tag{10.7}$$

which have to be solved for the σ^2s inherent in $\mathbf{P}$. We now show that those equations can also be written as equating calculated and expected best predicted values:

$$\left\{ {}_c \, \mathrm{E}[(\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0] \right\}_{i=0}^{r} = \left\{ {}_c \, (\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0 \right\}_{i=0}^{r} \tag{10.8}$$

for

$$\tilde{\mathbf{u}}^0 = \left\{ {}_c \, \tilde{\mathbf{u}}_i^0 \right\}_{i=0}^{r} = \mathrm{E}[\mathbf{u}|\mathbf{y}]\Big|_{\beta=\hat{\beta}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{10.9}$$

of (6.43). First,

$$\begin{aligned} \tilde{\mathbf{u}}^0 &= \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{y} \\ &= \left\{ {}_d \, \sigma_i^2 \mathbf{I}_{q_i} \right\} \left\{ {}_c \, \mathbf{Z}_i' \right\} \mathbf{P}\mathbf{y}. \end{aligned}$$

Hence

$$\tilde{\mathbf{u}}_i^0 = \sigma_i^2 \mathbf{Z}_i'\mathbf{P}\mathbf{y} \tag{10.10}$$

and so

$$(\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0/\sigma_i^4 = \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}. \tag{10.11}$$

Moreover,

$$\begin{aligned} \mathrm{E}\left[(\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0/\sigma_i^4\right] &= \operatorname{tr}\left(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathrm{E}[\mathbf{y}\mathbf{y}']\right) \\ &= \operatorname{tr}\left[\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}(\mathbf{V} + \mathbf{X}\beta\beta'\mathbf{X}')\right] \\ &= \operatorname{tr}\left[\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{V}\right] = \operatorname{tr}\left[\mathbf{P}\mathbf{V}\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\right] \\ &= \operatorname{tr}\left[\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\right]. \end{aligned} \tag{10.12}$$

Therefore the REML equations of (10.7) are equivalent to

$$\mathrm{E}[(\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0] = (\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0. \tag{10.13}$$

Moreover, from (10.7), the REML equations can, for $i = 1, 2, \ldots, r$ be written as

$$\begin{aligned} \sigma_i^4 \operatorname{tr}\left[\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\right] &= \sigma_i^4 \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} \\ &= (\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0 \qquad \text{from (10.11)} \end{aligned} \tag{10.14}$$

or

$$\sigma_i^2 = \frac{(\tilde{\mathbf{u}}_i^0)'\tilde{\mathbf{u}}_i^0}{\sigma_i^2 \mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i')}. \tag{10.15}$$

This suggests a substitution algorithm: starting with initial guesses $\sigma_i^{2(0)}$ find successive iterates of $\sigma_i^2, \sigma_i^{2(m)}$, from

$$\sigma_i^{2(m+1)} = \frac{\tilde{\mathbf{u}}_i^{\prime(m)}\tilde{\mathbf{u}}_i^{(m)}}{\sigma_i^{2(m)} \mathrm{tr}(\mathbf{P}^{(m)}\mathbf{Z}_i\mathbf{Z}_i')}, \tag{10.16}$$

where

$$\begin{aligned} \tilde{\mathbf{u}}_i^{(m)} &= \sigma_i^{2(m)}\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{y} \quad \text{and} \\ \mathbf{P}^{(m)} &= \mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)}\mathbf{X}[\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{X}]^-\mathbf{X}'\mathbf{V}^{-1(m)}. \end{aligned}$$

Successive substitutions would be performed in (10.16) until convergence.

c. Newton–Raphson method

The Newton–Raphson method is an old and celebrated method that can be used for maximization of a nonlinear function. More precisely, it is a root-finding algorithm. Starting from a function $f(\boldsymbol{\theta})$ we wish to find a root of

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{10.17}$$

which we hope is a maximum.

We expand $\partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ about $\boldsymbol{\theta}_0$ as

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = f'(\boldsymbol{\theta}) \doteq f'(\boldsymbol{\theta}_0) + \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{10.18}$$

Equating (10.18) to $\mathbf{0}$, solve for the root as

$$f'(\boldsymbol{\theta}_0) + \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{0} \tag{10.19}$$

which gives

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 - \left[\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]^{-1} f'(\boldsymbol{\theta}_0). \tag{10.20}$$

This can be used iteratively to refine the estimate of the root:

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - \left[\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]^{-1}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}} f'(\boldsymbol{\theta}^{(m)}). \tag{10.21}$$

To use (10.21) we need the first and second derivatives of the function.

We illustrate this method on the profile log likelihood, (6.61), of the components-of-variance model:

$$\log l_P(\mathbf{V}) = -\tfrac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{y} - \tfrac{1}{2}\log|\mathbf{V}| - \tfrac{N}{2}\log(2\pi). \tag{10.22}$$

where $\mathbf{V} = \sum_{i=0}^{r}\mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2$ and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$. Since this is a profile log likelihood it depends only on the σ^2s and not on β. We therefore need partial derivatives of $\log l_P$ with respect to σ_i^2 and the mixed partial derivatives with respect to σ_i^2 and σ_j^2.

The ingredients for these are given in Section 6.12. From (6.78) we have

$$\frac{\partial \mathbf{P}}{\partial \sigma_i^2} = -\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}$$

and therefore

$$\frac{\partial^2 \mathbf{P}}{\partial \sigma_i^2\,\partial \sigma_j^2} = \mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P} + \mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}.$$

Also, from (M.20)

$$\frac{\partial \log|\mathbf{V}|}{\partial \sigma_i^2} = \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right)$$

and using (M.18)

$$\frac{\partial^2 \log|\mathbf{V}|}{\partial \sigma_i^2\,\partial \sigma_j^2} = -\operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right).$$

From these derivatives it is straightforward to calculate

$$l_{\sigma^2} = \frac{\partial \log l_P}{\partial \boldsymbol{\sigma}^2} = \left\{ {}_c\ \tfrac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} - \tfrac{1}{2}\operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right)\right\} \tag{10.23}$$

and

$$\begin{aligned} l_{\sigma^2\sigma^2} \;=\; \frac{\partial^2 \log l_P}{\partial \boldsymbol{\sigma}^2\,\partial(\boldsymbol{\sigma}^2)'} &= \left\{ {}_m\ \tfrac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{y} + \tfrac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}\right. \\ &\qquad \left. -\tfrac{1}{2}\operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right)\right\}. \end{aligned} \tag{10.24}$$

Based on (10.21) the Newton–Raphson algorithm would take the form

$$\boldsymbol{\sigma}^{2(m+1)} = \boldsymbol{\sigma}^{2(m)} - (l_{\sigma^2\sigma^2})^{-1} l_{\sigma^2}, \tag{10.25}$$

where $l_{\sigma^2\sigma^2}$ and l_{σ^2} are evaluated at $\sigma^2 = \sigma^{2(m)}$.

The Newton–Raphson method is not without its drawbacks. First, it does not guarantee convergence, even to a local maximum. It can fail when the linearized approximation in (10.18) is a poor one. Second, it does not necessarily keep iterations in the parameter space. For example, (10.25) could lead to negative $\sigma^2 s$.

Various improvements are possible to remedy these defects. For example, σ_i can be estimated in place of σ_i^2 (and then squared to get σ_i^2) to keep iterative estimates of σ_i^2 positive. Also, taking smaller steps by modifying (10.25) to be of the form

$$\sigma^{2(m+1)} = \sigma^{2(m)} - \alpha(l_{\sigma^2\sigma^2})^{-1} l_{\sigma^2}, \tag{10.26}$$

where $0 < \alpha \leq 1$ is often a good idea. For more details on implementation and for some calculational details for alternative models, see Jennrich and Schluchter (1986), Lindstrom and Bates (1988) and Press et al. (1996).

10.3 COMPUTING ML ESTIMATES FOR GLMMs

a. Numerical quadrature

Generalized linear mixed models pose special challenges beyond linear mixed models because of the high-dimensional integration required to evaluate (and hence maximize) the likelihood. The direct numerical evaluation of integrals has a long history in mathematics and is a natural first place for considering how to deal with the computational complexity of GLMMs. We start by considering a GLMM with a single, normally distributed random effect. Let y_{ij} be the jth observation corresponding to the ith level of the random effect so that

$$\begin{aligned} y_{ij}|\mathbf{u} &\sim \text{indep. } f_{Y_{ij}|\mathbf{U}}(y_{ij}|\mathbf{u}) \\ f_{Y_{ij}|\mathbf{u}}(y_{ij}|\mathbf{u}) &= \exp\{[y_{ij}\gamma_{ij} - b(\gamma_{ij})]/\tau^2 - c(y_{ij},\tau)\} \\ \mathrm{E}[y_{ij}|\mathbf{u}] &= \mu_{ij} \\ g(\mu_{ij}) &= \mathbf{x}_{ij}'\beta + u_i \\ u_i &\sim \text{i.i.d. } \mathcal{N}(0,\sigma_u^2). \end{aligned} \tag{10.27}$$

The likelihood for this model is

$$L = \int \prod_{i,j} f_{Y_{ij}|U_i}(y_{ij}|u_i) f_{U_i}(u_i)\, du_i$$

$$= \prod_i \int_{-\infty}^{\infty} e^{\sum_j [y_{ij}\gamma_{ij} - b(\gamma_{ij})]/\tau^2 - \sum_j c(y_{ij},\tau)} \frac{e^{-u_i^2/(2\sigma_u^2)}}{\sqrt{2\pi\sigma_u^2}} du_i,$$

$$= \prod_i \int_{-\infty}^{\infty} h_i(u_i) \frac{e^{-u_i^2/(2\sigma_u^2)}}{\sqrt{2\pi\sigma_u^2}} du_i, \tag{10.28}$$

where $h_i(u_i) = e^{\sum_j [y_{ij}\gamma_{ij} - b(\gamma_{ij})]/\tau^2 - \sum_j c(y_{ij},\tau)}$ and γ_{ij} is a function of u_i.

It can be seen that the likelihood is the product of one-dimensional integrals of the form

$$\int_{-\infty}^{\infty} h(u) \frac{e^{-u^2/(2\sigma_u^2)}}{\sqrt{2\pi\sigma_u^2}} du$$

which, upon a change of variables of $u = \sqrt{2}\sigma_u v$, can be written as

$$\int_{-\infty}^{\infty} h(\sqrt{2}\sigma_u v) \frac{e^{-v^2}}{\sqrt{\pi}} dv \equiv \int_{-\infty}^{\infty} h^*(v) e^{-v^2} dv, \tag{10.29}$$

where $h^*(\cdot) \equiv h(\sqrt{2}\,\sigma_u \cdot)/\sqrt{\pi}$.

– i. *Gauss–Hermite quadrature*

Numerical integration over an unbounded range can be difficult. However, for integrals of smooth functions $h^*(\cdot)$ multiplied by the function e^{-v^2}, the method of Gauss–Hermite quadrature is available. This approximates the integral in (10.29) as a weighted sum:

$$\int_{-\infty}^{\infty} h^*(v) e^{-v^2} dv \doteq \sum_{k=1}^{d} h^*(x_k) w_k, \tag{10.30}$$

where the weights, w_k, and the evaluation points, x_k, are designed to provide an accurate approximation in the case where $h^*(\cdot)$ is a polynomial. More specifically, when the sum is from 1 to d, Gauss–Hermite quadrature gives the exact answer for all polynomials up to degree $2d - 1$. Table 10.1 lists the x_k and w_k for $d = 3$, 4, and 5. More extensive tables are available in, for example, Abramowitz and Stegun (1964), or the x_k and w_k can be calculated via mathematical software since

$$\begin{aligned} x_k &= i\text{th zero of } H_n(x) \\ w_k &= \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(x_k)]^2}, \end{aligned} \tag{10.31}$$

Table 10.1: Constants for Gauss–Hermite Quadrature

	x_k	w_k
$d = 3$	-1.22474487	0.29540898
	0	1.18163590
	1.22474487	0.29540898
$d = 4$	-1.65068012	0.08131284
	-0.52464762	0.80491409
	0.52464762	0.80491409
	1.65068012	0.08131284
$d = 5$	-2.02018287	0.01995324
	-0.95857246	0.39361932
	0	0.94530872
	0.95857246	0.39361932
	2.02018287	0.01995324

where $H_n(x)$ is the Hermite polynomial of degree n.

As an illustration consider

$$\int_{-\infty}^{\infty}(1+x^2)e^{-x^2}\,dx = \frac{3}{2}\sqrt{\pi} \doteq 2.65868.$$

This would be approximated using 3-point quadrature as

$$\begin{aligned}\int_{-\infty}^{\infty}(1+x^2)e^{-x^2}\,dx &\doteq (1+[-1.22474]^2)(0.29541) + (1+0^2)(1.18164) \\ &\quad + (1+1.22474^2)(0.29541) \\ &= (2.5)(0.29541)(2) + 1.18164 \\ &= 2.65868,\end{aligned}$$

as expected. On the other hand,

$$\int_{-\infty}^{\infty}(1+x^6)e^{-x^2}\,dx = \frac{23}{8}\sqrt{\pi} \doteq 5.0958,$$

which is approximated by $(4.375)(0.29541)(2) + 1.18163 = 3.76645$ with 3-point quadrature, a poor approximation. But 4-point quadrature gets the answer exactly right.

By using quadrature of a high-enough degree, accurate approximations can be calculated to integrals of functions that are similar to those

of any high-degree polynomial. For the likelihood calculations we have in mind, practical experience shows that quadrature with less than 10 points often gives inaccurate answers, while 20 is usually enough for a good degree of approximation.

Two cautions are in order. First, if the function is not properly "centered," Gauss–Hermite quadrature can give a poor approximation and second, if the function whose integral is to be approximated is not a smooth one, the approximation can also be poor. To illustrate the idea of centering, consider

$$\int_{-\infty}^{\infty} e^{2xa-a^2} e^{-x^2} dx,$$

which is easily shown to be $\sqrt{\pi}$ for any value of a. Five-point quadrature gives $\sqrt{\pi}$ as the answer when $a = 0$ but has an error of 0.001 when $a = 1$, and an error of 0.240 when $a = 2$, eventually giving an answer of 0 as $a \to \pm\infty$. This shows that the approximation is more accurate when the values of x_k are near where the function is nonzero and can be inaccurate otherwise. Exercise E 10.4 illustrates that the approximation can be poor for non-smooth functions.

– ii. *Likelihood calculations*

Gauss–Hermite quadrature can be used to calculate integrals with respect to the normal density as

$$\int_{-\infty}^{\infty} h(x) \frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} dx \doteq \sum_{k=1}^{d} h(\sqrt{2}\,\sigma x_k) w_k / \sqrt{\pi}. \tag{10.32}$$

To derive an approximation to a likelihood such as (10.28), (10.32) would be used repeatedly. For example, suppose that our model was, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$,

$$\begin{aligned} y_{ij}|\mathbf{a} &\sim \text{indep. Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) &= \mu + a_i \\ a_i &\sim \text{i.i.d. } \mathcal{N}(0, \sigma_a^2). \end{aligned} \tag{10.33}$$

The log likelihood would be

$$l = \sum_i \log \int_{-\infty}^{\infty} e^{(\mu+a_i)y_{i\cdot} - n\log(1+e^{\mu+a_i})} \frac{e^{-a_i^2/(2\sigma_a^2)}}{\sqrt{2\pi\sigma_a^2}} da_i$$

$$\doteq \sum_i \log\left(\sum_k e^{(\mu+x_k)y_{i\cdot}-n\log(1+e^{\mu+x_k})}w_k/\sqrt{\pi}\right). \qquad (10.34)$$

This log likelihood needs to be maximized numerically to get estimates of μ and σ_a. Derivatives of the log likelihood, which are often required by numerical maximization algorithms, can be approximated similarly. Alternatively, quasi–Newton or derivative-free maximization methods can be used.

A likelihood ratio test, or best predicted values, would require similar numerical calculation. For example, the best predicted values for model (10.33) are

$$\begin{aligned}
\mathrm{E}[a_i|\mathbf{y}] &= \int a_i f_{a_i|\mathbf{y}}(a_i|\mathbf{y})\,da_i \\
&= \int a_i f_{\mathbf{y}|a_i}(\mathbf{y}|a_i) f_{a_i}(a_i)/f_{\mathbf{y}}(\mathbf{y})\,da_i \\
&= \frac{\int a_i e^{(\mu+x_k)y_{i\cdot}-n\log(1+e^{\mu+x_k})} f_{a_i}(a_i)\,da_i}{\int e^{(\mu+x_k)y_{i\cdot}-n\log(1+e^{\mu+x_k})} f_{a_i}(a_i)\,da_i}.
\end{aligned}$$

The denominator is exactly the likelihood, the approximation for which is displayed in (10.34), and the numerator would be approximated similarly. If the MLEs $\hat{\mu}$ and $\hat{\sigma}_a$ were used in the calculation, then the approximation would be for the estimated best predictor.

– iii. *Limits of numerical quadrature*

Numerical quadrature is limited in its application. The calculations above show that with clustered data the computations are feasible. It is also possible to use numerical quadrature to approximate the likelihood of a model with two nested random effects (see E 10.5). However, crossed random factors and higher levels of nesting lead to integrals that are not amenable to Gauss–Hermite quadrature.

The possible distributions available for the random effect distribution are also limited. The quadrature techniques described in the preceding sections are appropriate only when integrating products of functions with e^{-x^2}, that is, for normally distributed random effects. To employ other random effects distributions we need alternative quadrature methods. Although these could be derived conceptually they are not readily available, except for integrating products with e^{-x}, which would correspond to exponentially distributed random effects. This

methodology is called *Laguerre integration* (Abramowitz and Stegun, 1964, p. 923). Another way to extend the class of random effects distributions is to consider transformations of normally distributed random effects (Piepho and McCulloch, 1999), for example, by using e^{a_i} to introduce lognormally distributed random effects.

b. EM algorithm

As noted in Section 10.2a, for mixed models a typical missing data configuration is to assume the random effects to be the missing data. Since the random effects introduce correlation in the model, once they are filled in by the EM algorithm and can be treated as fixed known values, the problem often simplifies. The variance components version of the linear mixed model, for example, treated in Section 10.2a, simplifies to the traditional homoscedastic linear model, for which maximum likelihood is ordinary least squares. So, for that model, EM reduces maximum likelihood to a series of least squares problems.

We return to the GLMM of Chapter 8 as our most general model:

$$\begin{aligned} y_i|\mathbf{u} &\sim \text{indep. } f_{Y_i|\mathbf{U}}(y_i|\mathbf{u}) \\ f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) &= \exp\{[y_i\gamma_i - b(\gamma_i)]/\tau^2 - c(y_i,\tau)\} \\ \mathrm{E}[y_i|\mathbf{u}] &= \mu_i \qquad (10.35) \\ g(\mu_i) &= \mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u} \qquad (10.36) \\ \mathbf{u} &\sim f_{\mathbf{U}}(\mathbf{u}|\mathbf{D}), \end{aligned}$$

where $\mathbf{D}$ represents the parameters governing the distribution of $\mathbf{u}$ in keeping with the notation in Chapter 6.

To set up the EM algorithm we declare $\mathbf{u}$ to be the missing data so that the *complete data* are $\mathbf{w}' = (\mathbf{y}', \mathbf{u}')$. The EM algorithm proceeds by forming the log likelihood of the complete data, calculating its expectation with respect to the conditional distribution of $\mathbf{u}$ given $\mathbf{y}$ and then maximizing with respect to the parameters. The algorithm is iterative since we now recalculate the log likelihood of the complete data given the new parameter estimates, and so on.

The distribution of the complete data, $\mathbf{w}$, can be factored as $f_{\mathbf{Y},\mathbf{U}} = f_{\mathbf{Y}|\mathbf{U}} f_{\mathbf{U}}$, so that the complete data log likelihood, $\log L_w$, is

$$\log L_w = \log f_{\mathbf{Y}|\mathbf{U}} + \log f_{\mathbf{U}}$$

$$= \sum_{i=1}^{n} \log f_{Y_i|\mathbf{U}} + \log f_{\mathbf{U}}$$

$$= [\sum y_i\gamma_i - b(\gamma_i)]/\tau^2 - \sum c(y_i, \tau) + \log f_{\mathbf{U}}. \quad (10.37)$$

This choice of missing data has two advantages. First, conditional on $\mathbf{u}$, the y_i are independent. Second, β and τ enter only the first portion of the log likelihood (the GLM portion) whereas $\mathbf{D}$ enters only through $f_{\mathbf{U}}$, the portion coming from the random effects. The *maximization* or *M–step* of the algorithm with respect to β and τ will be similar to the calculations for GLMs in Section 5.4e. Maximizing with respect to $\mathbf{D}$ is akin to ML using the distribution of $\mathbf{u}$. In fact, if the distribution of $\mathbf{u}$ is a member of the exponential family, then the M–step for $\mathbf{D}$ simplifies to maximum likelihood after replacing the sufficient statistics with their conditional expected values.

The EM algorithm takes the following form:

1. Choose starting values $\beta^{(0)}, \tau^{(0)}$, and $\mathbf{D}^{(0)}$. Set $m = 0$.
2. Calculate (with expectations evaluated under current values)
   (a) $\beta^{(m+1)}$ and $\tau^{(m+1)}$ to maximize $\mathrm{E}[\log f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}, \beta, \tau)|\mathbf{y}]$.
   (b) $\mathbf{D}^{(m+1)}$ to maximize $\mathrm{E}[\log f_{\mathbf{U}}(\mathbf{u}|\mathbf{D})|\mathbf{y}]$.
   (c) Set $m = m + 1$.
3. If convergence is achieved, declare the current values to be the MLEs; otherwise return to step 2.

In general, the expectations in neither steps 2(a) nor 2(b) can be computed in closed form for the model. This is because the conditional distribution of $\mathbf{u}|\mathbf{y}$ involves $f_{\mathbf{Y}}$, i.e., the likelihood, which we are trying to avoid calculating directly. However, because it is possible to produce random draws from the conditional distribution of $\mathbf{u}|\mathbf{y}$ without specifying $f_{\mathbf{Y}}$, one can then use those draws to form Monte Carlo approximations to the required expectations. We describe this approach in the next section.

c. Markov chain Monte Carlo algorithms

There are a number of ways to generate draws from a difficult-to-calculate density, e.g., Gibbs sampling or Markov chain Monte Carlo methods (Robert and Casella, 1999). McCulloch (1994, 1997) uses the

Gibbs sampler for probit models, and the Metropolis–Hastings algorithm for general GLMM problems, while Booth and Hobert (1999) use the independence sampler.

– i. *Metropolis*

As an example, we consider a Metropolis algorithm, which generates a Markov chain sequence of values that eventually stabilizes to draws from the candidate distribution. To specify a Metropolis algorithm, a candidate distribution, $h_{\mathbf{U}}(\mathbf{u})$, must be selected, from which potential new values are drawn. The *acceptance function*, which gives the probability of accepting a new value (as opposed to keeping the previous value) is given by

$$A_k(\mathbf{u}^*, \mathbf{u}) = \min\left\{1, \frac{f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}^*|\mathbf{y}, \boldsymbol{\beta}, \tau, \mathbf{D}) h_{\mathbf{U}}(\mathbf{u})}{f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta}, \tau, \mathbf{D}) h_{\mathbf{U}}(\mathbf{u}^*)}\right\}, \tag{10.38}$$

where $\mathbf{u}^* = (u_1, u_2, \ldots, u_{k-1}, u_k^*, u_{k+1}, \ldots, u_q)'$, which is the candidate new value and has all entries equal to the previous value except the kth.

What can be used for the candidate distribution? Upon choosing $h_{\mathbf{U}} = f_{\mathbf{U}}$, the ratio term in (10.38) simplifies to

$$\begin{aligned}
&\frac{f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}^*|\mathbf{y}, \boldsymbol{\beta}, \tau, \mathbf{D}) h_{\mathbf{U}}(\mathbf{u})}{f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta}, \tau, \mathbf{D}) h_{\mathbf{U}}(\mathbf{u}^*)} \\
&\qquad = \frac{\prod_{i=1}^n f_{Y_i|\mathbf{U}}(y_i|\mathbf{u}^*, \boldsymbol{\beta}, \tau) f_{\mathbf{U}}(\mathbf{u}^*|\mathbf{D}) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D})}{\prod_{i=1}^n f_{Y_i|\mathbf{U}}(y_i|\mathbf{u}, \boldsymbol{\beta}, \tau) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D}) f_{\mathbf{U}}(\mathbf{u}^*|\mathbf{D})} \\
&\qquad = \frac{\prod_{i=1}^n f_{Y_i|\mathbf{U}}(y_i|\mathbf{u}^*, \boldsymbol{\beta}, \tau)}{\prod_{i=1}^n f_{Y_i|\mathbf{U}}(y_i|\mathbf{u}, \boldsymbol{\beta}, \tau)}.
\end{aligned} \tag{10.39}$$

This calculation involves specifying only the generalized linear model portion of the model, i.e. the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$.

Incorporating this Metropolis step into the EM algorithm gives a Monte Carlo EM (MCEM) algorithm as follows:

1. Choose starting values $\boldsymbol{\beta}^{(0)}, \tau^{(0)}$, and $\mathbf{D}^{(0)}$. Set $m = 0$.

2. Generate M values, $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(M)}$, from the conditional distribution of $\mathbf{u}$ given $\mathbf{y}$ using the Metropolis algorithm described above.

(a) Calculate $\beta^{(m+1)}$ and $\tau^{(m+1)}$ to maximize a Monte Carlo estimate of $\mathrm{E}[\log f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\beta,\tau)|\mathbf{y}]$, i.e., choose them to maximize $(1/M)\sum_{k=1}^{M}\log f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}^{(k)},\beta,\tau)$.

(b) Calculate $\mathbf{D}^{(m+1)}$ to maximize $(1/M)\sum_{k=1}^{M}\log f_{\mathbf{U}}(\mathbf{u}^{(k)}|\mathbf{D})$.

(c) Set $m = m+1$.

3. If convergence is achieved, declare the current values to be the MLEs; otherwise return to step 2.

While computationally intensive, this approach remains feasible for a variety of data configurations.

– ii. *Monte Carlo Newton–Raphson*

There is also a simulation analog of the working variates or Fisher scoring approach which was used to fit GLMs in Section 5.4e. Whenever the marginal density of $\mathbf{y}$ is formed as a mixture as in (10.35) with separate parameters for $f_{\mathbf{Y}|\mathbf{U}}$ and $f_{\mathbf{U}}$, then the ML equations for $\boldsymbol{\theta} = (\beta', \tau)'$ and $\mathbf{D}$ take the following form (see Exercise E 10.6):

$$\mathrm{E}\left[\left.\frac{\partial f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|\mathbf{y}\right] = \mathbf{0} \tag{10.40}$$

$$\mathrm{E}\left[\left.\frac{\partial f_{\mathbf{U}}(\mathbf{u}|\mathbf{D})}{\partial\mathbf{D}}\right|\mathbf{y}\right] = \mathbf{0}. \tag{10.41}$$

Equation (10.41) involves only the distribution of $\mathbf{u}$ and is often fairly easy to solve, e.g., when the distribution is normal. On the other hand, (10.40) is amenable to Newton–Raphson or a scoring approach just as in Chapter 5.

Expanding $\partial f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})/\partial\beta$ as a function of β around a value $\boldsymbol{\theta}_0$ gives

$$\frac{\partial f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})}{\partial\beta} \doteq \left.\frac{\partial f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})}{\partial\beta}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \left.\frac{\partial^2 f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})}{\partial\beta\,\partial\beta'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\beta-\beta_0). \tag{10.42}$$

Specializing this to our model, and dropping the term with a conditional expected value of zero, the formula for a scoring-type algorithm

becomes

$$\frac{\partial f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \doteq \frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) - \frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \quad (10.43)$$

where $\mathbf{W} = \left\{{}_d[v(\mu_i)g'_\mu(\mu_i)]^{-1}\right\}$ and $\boldsymbol{\Delta} = \left\{{}_c g_\mu(\mu_i)\right\}$ and it is understood that $\mathbf{W}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}|\mathbf{u}]$ are all functions of $\mathbf{u}$ and that all parameters are evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Using this approximation in (10.40) leads to an iteration equation of the form

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{X}'\mathrm{E}[\mathbf{W}|\mathbf{y}]\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}\left[\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu})|\mathbf{y}\right]. \quad (10.44)$$

This analog of scoring would proceed by iteratively solving (10.44), (10.41), and an equation for τ. An advantage of the scoring approach over MCEM is that is makes automatic the maximization step 2(a).

Again, typically the expectations cannot be evaluated in closed form, which leads to a Monte Carlo Newton–Raphson (MCNR) approach. As before, an algorithm like the Metropolis algorithm is used to approximate the expectations in (10.44) since these are expectations with respect to the conditional distribution of $\mathbf{u}$ given $\mathbf{y}$.

d. Stochastic approximation algorithms

A different approach to fitting these models has been suggested recently by Gu and Kong (1998) through the use of a stochastic approximation (SA) algorithm, although the basic idea of using SA to find MLEs is certainly older (e.g. Moyeed and Baddeley, 1991; Ruppert, 1991). The basic concept is to write $f_{\mathbf{Y},\mathbf{U}}$ as $f_{\mathbf{Y}}f_{\mathbf{U}|\mathbf{Y}}$. It is then straightforward to derive

$$\frac{\partial \log f_{\mathbf{Y},\mathbf{U}}(\mathbf{y},\mathbf{u}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\partial \log f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}|\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \quad (10.45)$$

We are interested in finding the root of the likelihood equation, that is, the value of $\boldsymbol{\theta}$ such that $\partial \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$. SA algorithms are methods of finding roots of regression equations, so we need to rewrite (10.45) as a regression equation.

Write $\mathbf{m}(\boldsymbol{\theta})$ for the score function, $\partial \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, to emphasize we are regarding it as a function of $\boldsymbol{\theta}$ and that it is not a function of $\mathbf{u}$. Next observe that

$$\mathrm{E}\left[\left.\frac{\partial \log f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}|\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|\mathbf{y}\right] = \mathbf{0} \quad (10.46)$$

for fixed $\mathbf{y}$ when the expectation is taken with respect to the conditional distribution of $\mathbf{u}$ given $\mathbf{y}$. This is the usual score identity, e.g. (5.9), applied to the conditional distribution. Hence $\partial \log f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}|\mathbf{y},\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, can be regarded as a mean-zero, "error" term in the following "regression" equation, which is (10.45) rewritten:

$$\frac{\partial \log f_{\mathbf{Y},\mathbf{U}}(\mathbf{y},\mathbf{u}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{m}(\boldsymbol{\theta}) + \text{error}. \tag{10.47}$$

Thus, inserting random values of $\mathbf{u} \sim f_{\mathbf{U}|\mathbf{Y}}$ into $\partial \log f_{\mathbf{Y},\mathbf{U}}(\mathbf{y},\mathbf{u}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ gives "data" for performing the regression.

To implement an SA algorithm, we use the Metropolis algorithm of Section 10.2c to generate a sequence of values $\mathbf{u}^{(k)} \sim f_{\mathbf{U}|\mathbf{Y}}$ and use them to form data $\partial \log f_{\mathbf{Y},\mathbf{U}}(\mathbf{y},\mathbf{u}^{(k)}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. One can then apply a multivariate version of an SA algorithm in order to find the root of the likelihood equation. Ruppert (1991) provides a nice review.

An SA algorithm applied to maximum likelihood for the GLMM would generally take the form

$$\theta^{(m+1)} = \theta^{(m)} - a_m \frac{\partial \log f_{\mathbf{Y},\mathbf{U}}(\mathbf{y},\mathbf{u}^{(m)}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{10.48}$$

where a_m is chosen to decrease slowly to zero. Ideally, a_m also incorporates information about the derivative of $\partial \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ (with respect to $\boldsymbol{\theta}$) at the root, but this is rarely known in practice.

A reasonable choice for a_m allowing it to decrease to zero and using some information about the curvature of the surface to be maximized is

$$a_m = \frac{a}{(m+k)^\alpha}\left(\hat{\mathrm{E}}\left[\frac{\partial^2 \log f_{\mathbf{Y},\mathbf{U}}(\mathbf{y},\mathbf{u}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right]\right)^{-1}, \tag{10.49}$$

where $\hat{\mathrm{E}}$ denotes a Monte Carlo estimate of the expectation (taken with respect to the conditional distribution of $\mathbf{u}$ given $\mathbf{y}$). This choice of a_m follows recommendations in the literature; see, for example, Frees and Ruppert (1990) and Ruppert (1991). There is latitude in the choice of the constants a, k and α, although we have successfully used $a = 3, k = 50$ and $\alpha = 0.75$. Estimates are formed by iterating until convergence.

MCNR and SA are similar, with the main difference being that SA uses a single simulated value at each iteration. The multiplier a_m decreases the step size as the iterations increase in SA. This eventually serves to eliminate the stochastic error involved in the Metropolis step. To achieve a corresponding reduction using MCNR, the simulation size

would have to be increased as the iterations increase in order to eliminate the simulation noise.

SA seems to have advantages in that it can use all of the simulated data to calculate estimates and it uses the simulated values one at a time. A theoretical advantage of SA is that convergence proofs are worked out for many cases. Practical details of the implementation of both SA and MCNR have not yet been settled in the literature.

e. Simulated maximum likelihood

While both MCEM and MCNR work on the log of the likelihood, Geyer and Thompson (1992), Gelfand and Carlin (1993), and Durbin and Koopman (1997) have suggested simulation to estimate the value of the likelihood directly. Starting from the likelihood we have

$$\begin{aligned} L &= \int f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\beta,\tau) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D})\, d\mathbf{u} \\ &= \int \frac{f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\beta,\tau) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D})}{h_{\mathbf{U}}(\mathbf{u})} h_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} \\ &= \mathrm{E}_u\left[\frac{f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u},\beta,\tau) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D})}{h_{\mathbf{U}}(\mathbf{u})}\right] \\ &\doteq \sum_{k=1}^{M} \frac{f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}^{(k)},\beta,\tau) f_{\mathbf{U}}(\mathbf{u}^{(k)}|\mathbf{D})}{h_{\mathbf{U}}(\mathbf{u}^{(k)})}, \end{aligned} \tag{10.50}$$

where the subscript u on the expectation is a reminder that the expectation is with respect to $\mathbf{u}$, $h_{\mathbf{U}}(\mathbf{u})$ is a density with respect to which the expectation is taken, $\mathbf{u}^{(k)}$ are selected from this density, and M is the number of simulated values. This is an unbiased estimate no matter the choice of $h_{\mathbf{U}}(\mathbf{u})$. The simulated likelihood is then numerically maximized, either after a single simulation, or using multiple simulations in an iterative process where the importance sampling distribution is allowed to depend on the current parameter values.

Although unbiased, the approximation is sensitive to the choice of $h_{\mathbf{U}}(\mathbf{u})$ in the sense that it can be highly variable for choices far from the optimal choice (for the optimal choice see E 10.10). So implementation of simulated maximum likelihood must be done with care.

10.4 PENALIZED QUASI-LIKELIHOOD AND LAPLACE

The attractive features of quasi-likelihood, namely model robustness and less restrictive assumptions, have led to a search for generalizations applicable to GLMMs. Central to these is the use of a Laplace approximation (Tierney and Kadane, 1986) for evaluating the high-dimensional integral in the likelihood. The basic form of Laplace's approximation is based on a second-order Taylor series expansion and takes the form

$$\log \int_{\Re^q} e^{h(\mathbf{u})} d\mathbf{u} \doteq h(\mathbf{u}_0) + \frac{q}{2}\log 2\pi - \frac{1}{2}\log\left|-\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}\partial \mathbf{u}'}\right|, \tag{10.51}$$

where $\mathbf{u}_0$ is the solution to

$$\left.\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}}\right|_{\mathbf{u}=\mathbf{u}_0} = \mathbf{0}. \tag{10.52}$$

We utilize this result to approximate the log likelihood of the GLMM via

$$\begin{aligned} l &= \log \int f_{\mathbf{Y}|\mathbf{U}} f_{\mathbf{U}}\, d\mathbf{u} \\ &= \log \int e^{\log f_{\mathbf{Y}|\mathbf{U}} + \log f_{\mathbf{U}}}\, d\mathbf{u} \\ &= \log \int e^{h(\mathbf{u})}\, d\mathbf{u}, \end{aligned} \tag{10.53}$$

with $h(\mathbf{u}) = \log f_{\mathbf{Y}|\mathbf{U}} + \log f_{\mathbf{U}}$. To construct the Laplace approximation (10.52) must be solved and an expression for $\partial^2 h(\mathbf{u})/\partial \mathbf{u}\,\partial \mathbf{u}'$ is needed.

If we assume that $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ then

$$\log f_{\mathbf{U}} = -\frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u} - \frac{q}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{D}|$$

and $h(\mathbf{u})$ becomes

$$\log f_{\mathbf{Y}|\mathbf{U}} + \log f_{\mathbf{U}} = \log f_{\mathbf{Y}|\mathbf{U}} - \frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u} - \frac{q}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{D}|.$$

Differentiating with respect to $\mathbf{u}$ gives

$$\begin{aligned} \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} &= \frac{\partial \log f_{\mathbf{Y}|\mathbf{U}}}{\partial \mathbf{u}} - \mathbf{D}^{-1}\mathbf{u} \\ &= \frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu}) - \mathbf{D}^{-1}\mathbf{u}, \end{aligned} \tag{10.54}$$

where $\mathbf{W}$ and $\boldsymbol{\Delta}$ are defined below (10.43). The second equality comes about from derivations identical to (5.18) with $\mathbf{u}$ replacing $\boldsymbol{\beta}$ and $\mathbf{Z}$ replacing $\mathbf{X}$. To find $\mathbf{u}_0$ it is necessary to solve for $\mathbf{u}$ in

$$\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{D}^{-1}\mathbf{u}, \tag{10.55}$$

which is not as simple as it appears since $\mathbf{W}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}|\mathbf{u}]$ on the left-hand side of the equation are all functions of $\mathbf{u}$.

We will also need the second derivative:

$$\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}\,\partial \mathbf{u}'} = -\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\boldsymbol{\Delta}\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{u}'} + \frac{1}{\tau^2}\mathbf{Z}'\frac{\partial \mathbf{W}\boldsymbol{\Delta}}{\partial \mathbf{u}'}(\mathbf{y}-\boldsymbol{\mu}) - \mathbf{D}^{-1}. \tag{10.56}$$

For some models (e.g., the binomial or Poisson) $\mathbf{W}\boldsymbol{\Delta} = \mathbf{I}$ so the second term is zero. In general, the second term has expectation zero with respect to the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$. So it may be reasonable to consider it as negligible with respect to the other terms. If this is the case, (10.56) becomes

$$\begin{aligned} -\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u}\,\partial \mathbf{u}'} &\doteq \frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1}\mathbf{Z} + \mathbf{0} + \mathbf{D}^{-1} \\ &= \frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{D}^{-1} \\ &= \left(\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D} + \mathbf{I}\right)\mathbf{D}^{-1}. \end{aligned} \tag{10.57}$$

Using (10.57) in (10.51) gives

$$\begin{aligned} l &\doteq \log f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}_0) - \tfrac{1}{2}\mathbf{u}_0'\mathbf{D}^{-1}\mathbf{u}_0 - \tfrac{q}{2}\log 2\pi - \tfrac{1}{2}\log|\mathbf{D}| \\ &\qquad + \tfrac{q}{2}\log 2\pi - \tfrac{1}{2}\log|(\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D}/\tau^2 + \mathbf{I})\mathbf{D}^{-1}| \qquad (10.58) \\ &= \log f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}_0) - \tfrac{1}{2}\mathbf{u}_0'\mathbf{D}^{-1}\mathbf{u}_0 + \tfrac{1}{2}\log|\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D}/\tau^2 + \mathbf{I}|. \end{aligned}$$

This still must be maximized with respect to $\boldsymbol{\beta}$ to find the ML estimate. Differentiating with respect to $\boldsymbol{\beta}$ gives an approximate score equation of

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &\doteq \frac{\partial \log f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}_0)}{\partial \boldsymbol{\beta}} + \frac{\partial}{\partial \boldsymbol{\beta}}\frac{1}{2}\log|\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D}/\tau^2 + \mathbf{I}| \\ &= \frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) + \frac{\partial}{\partial \boldsymbol{\beta}}\frac{1}{2}\log|\mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D}/\tau^2 + \mathbf{I}| \\ &\doteq \frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}), \end{aligned} \tag{10.59}$$

where the second equality follows from (5.18) and for the third we have assumed that $\mathbf{W}$ changes negligibly as a function of $\boldsymbol{\beta}$. Thus we jointly solve the equations

$$\frac{1}{\tau^2}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0} \tag{10.60}$$

and

$$\frac{1}{\tau^2}\mathbf{Z}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{D}^{-1}\mathbf{u} \tag{10.61}$$

for $\boldsymbol{\beta}$ and $\mathbf{u}$. Of course, this only gives an estimate of $\boldsymbol{\beta}$; a subsidiary method is needed to estimate $\mathbf{D}$.

Equations (10.60) and (10.61) can also arise from jointly maximizing (with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$)

$$\log f_{\mathbf{Y}|\mathbf{U}} - \tfrac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u} \tag{10.62}$$

which is similar to a quasi–likelihood (the $f_{\mathbf{Y}|\mathbf{U}}$ term) with a "penalty" function added on (the $\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}$ term). In (10.62) the $\frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}$ term serves to prevent arbitrary values of $\mathbf{u}$ from being selected and forces them to be closer to zero (a shrinkage effect). Methods to solve these equations are thus frequently called *penalized quasi–likelihood* (PQL) methods. Green (1990), Schall (1991), and Wolfinger (1993) all discuss methods of this type.

In the "derivation" of the PQL equations quite a few approximations of undetermined accuracy are bandied about and the development has an air of ad hocery. How well do these methods work in practice? Unfortunately, not very.

Breslow and Lin (1995) and Lin and Breslow (1996) show that PQL methods lead to estimators which are asymptotically biased and hence inconsistent. Of course, inconsistency in itself may not be a worry if the asymptotic bias is small and the the small- or moderate-sized sample performance is good. After all, even full ML is not unbiased in small- or moderate-sized samples. Unfortunately, for situations like paired binary data the PQL estimator can perform quite badly. Its performance improves as the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$ gets closer to normal (and the Laplace approximation becomes more accurate), for example with a Poisson distribution with mean 7 or greater. However, from a practical point of view, we may prefer to transform such data to make them approximately normal and use LMM methods. We thus cannot recommend the use of simple PQL methods in practice.

Recently, there have been improvements in PQL methods (using more accurate Taylor expansions) that may lead to better-performing estimators. However, these have not yet been fully tested.

10.5 EXERCISES

E 10.1 If $\mathbf{u}_i$ of order q_i is distributed $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_i^2)$ show that the ML estimator of σ_i^2 is $\hat{\sigma}_i^2 = \mathbf{u}_i'\mathbf{u}_i/q_i$.

E 10.2 For w_k of (10.30) show that $\sum_k w_k = \sqrt{\pi}$ for any order Gauss–Hermite quadrature.

E 10.3 Calculate $\int_{-\infty}^{\infty}(1+x^2)\dfrac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx$ both analytically and using 3-point Gauss–Hermite quadrature. What relationship is there between w_1, w_2, and w_3?

E 10.4 Calculate $\mathrm{P}\{Z > 1.7\}$ when $Z \sim \mathcal{N}(0,1)$ using 3-, 4- and 5-point quadrature and compare to the value from a table. Is the approximation likely to improve by using a slightly higher-order quadrature? Why or why not?

E 10.5 Consider a nested logit-normal model:

$$\begin{aligned} y_{ijk}|\mathbf{a},\mathbf{b} &\sim \text{indep. Bernoulli}(p_{ijk}) \\ \text{logit}(p_{ijk}) &= \mu + a_i + b_{ij} \\ a_i &\sim \text{i.i.d. } \mathcal{N}(0,\sigma_a^2) \text{ independently of} \\ b_{ij} &\sim \text{i.i.d. } \mathcal{N}(0,\sigma_b^2). \end{aligned}$$

Write the likelihood in as simple a form as possible with regard to the integrations involved.

E 10.6 Derive (10.40) and (10.41).

E 10.7 For $u_i \sim$ i.i.d. $\mathcal{N}(0,\sigma^2)$ and hence $\mathbf{D} = \mathbf{I}\sigma^2$, write out (10.41).

E 10.8 Derive (10.44).

E 10.9 Show that (10.50) is an unbiased estimator of the likelihood independent of the choice of $h_{\mathbf{U}}(\mathbf{u})$.

E 10.10 Show that $h_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{U}|\mathbf{Y}}(\mathbf{u}|\mathbf{y})$ is the optimal choice of $h_{\mathbf{U}}(\mathbf{u})$ in the sense that it gives a zero variance estimator for (10.50).

E 10.11 For the case of a scalar u, derive (10.51) by approximating $h(u)$ in a second-order Taylor series about the point u_0 with $h'(u_0) = 0$.

E 10.12 Show that the Laplace approximation is exact for the case of a linear mixed model.

Chapter 11

NONLINEAR MODELS

11.1 INTRODUCTION

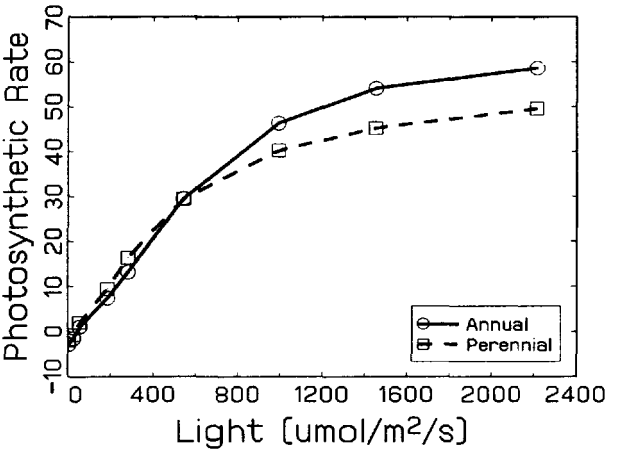
This chapter considers very briefly the topic of nonlinear mixed models (NLMMs). The main purpose is to emphasize that GLMMs are a proper subset of NLMMs, which comes with both advantages and disadvantages. We illustrate the ideas mostly in the context of an example. References are given to more complete coverage of these topics.

11.2 EXAMPLE: CORN PHOTOSYNTHESIS

Parker (1995) at Cornell University studied the photosynthetic ability of wild relatives of corn. The main question of interest was to compare two species (an annual and perennial) with respect to photosynthetic physiology. Seeds from two populations of each species were collected and grown in the greenhouse. The experimental design was a randomized complete block design with four blocks and three seeds from each population in each block (for a total of 12 seeds per block). After 24 days, photosynthesis was recorded at nine different light levels from full sunlight to darkness on one individual from each population in each block (N=16). Measurements on the same 16 plants were repeated after 48 days. From these data, photosynthesis versus irradiance response curves reflecting the change in photosynthetic rate with light level were derived.

The traits of interest are the maximum photosynthetic rate, dark respiration, the light compensation point, and the quantum yield. The maximum photosynthetic rate measures the maximum amount of carbon dioxide the plants are able to assimilate in full sunlight, the dark

Figure 11.1: Photosynthetic rate versus light for two plants.



respiration indicates how much carbon dioxide they respire in the dark, the light compensation point is the light level at which photosynthesis overcomes respiration and carbon assimilation becomes positive, and quantum yield is the efficiency of carbon assimilation at low light levels, or the slope of the light response curve as it crosses the light compensation point.

This is perhaps easier to describe mathematically and graphically. Figure 11.1 shows the graph of photosynthetic rate of two representative plants, one annual and one perennial species. The form of the curve typically used to describe this relationship as a function of light, l, is

$$\text{PHOTO}(l) = \beta_1 + \beta_2 e^{-\beta_3 l}. \tag{11.1}$$

Though a simple way to write the equation, not all of the parameters β_i are directly of interest. Equation (11.1) has value $\beta_1 + \beta_2$ at $l = 0$, asymptotes at β_2 at $l \to \infty$ (for $\beta_3 > 0$), and crosses the x-axis at $l = -\log(-\beta_1/\beta_2)/\beta_3$. Thus we define

$$\begin{aligned} \alpha &= \text{asymptote} = \text{maximum photosynthetic rate} = \beta_1 \\ \delta &= \text{dark respiration rate} = \beta_1 + \beta_2 \\ \lambda_0 &= \text{light compensation point} = -\log(-\beta_1/\beta_2)/\beta_3 \end{aligned} \tag{11.2}$$

and rewrite (11.1) (see E 11.1) as

$$\text{PHOTO}(l) = \alpha - (\alpha - \delta)\left(\frac{\alpha}{\alpha - \delta}\right)^{\frac{l}{\lambda_0}}. \tag{11.3}$$

This is the equation for a single plant. If we assume that (11.3) represents the mean response as a function of light, this cannot be a GLMM. This is because no function of rate will be linear in the parameters. Hence models like this one which are intrinsically nonlinear in the parameters are not GLMMs. Of course, GLMMs are special cases of nonlinear mixed models; but this example shows that they are not one and the same.

So far, only the effect of light is incorporated into (11.3). What about the effect of plants, blocks, populations and species? If the results in Figure 11.1 are typical then we might consider modeling α as a function of species and perhaps each of the other factors as well. Use i as a subscript for species, j for populations, k for blocks, m for plants, and t for time. A reasonable model for α itself would be

$$\alpha_{ijkmt} = \mu + \text{SPECIES}_i + \text{POPLN}_j + \text{BLOCK}_k + \text{PLANT}_m, \tag{11.4}$$

where BLOCK and PLANT might be considered random effects. This approach, of modeling the parameters as functions of the other factors, is sensible since they represent (three of) the traits of interest in this study. It is thus easy to assess, for example, the influence of species on the maximum photosynthetic rate, α. To complete the overall model we would need similar submodels for the dependence of δ and λ_0 on the foregoing factors. However, it is certainly also possible to entertain alternate ways of incorporating the factors.

Some of the advantages and disadvantages of GLMMs as compared to NLMMs are clear from this example. To specify the NLMM each of the sub-models for α, δ, and λ_0 must first be specified, then fit and perhaps simplified using the data. For example, the plant effects in each model would need to be considered and separate plant variance components would need to be estimated for each submodel. This would lead to a large number of parameters to be estimated. In contrast, for a GLMM we assume that all model terms enter into the mean of the distribution, simplifying the construction of the model.

Clearly this is a double-edged sword. Although the models are simpler, with fewer parameters to estimate, they may make unreasonably restrictive assumptions. In the photosynthesis example the model is fundamentally nonlinear and a GLMM will not suffice.

11.3 PHARMACOKINETIC MODELS

A common usage of nonlinear mixed models is in pharmacokinetic modeling, that is, models for describing the movement of drugs or other substances through the body. The mean structure for these models is typically derived from a system of differential equations. The differential equations are commonly set up by hypothesizing the existence of two or more *compartments* in the body with differential equations incorporating the rates of flow from one compartment to the next.

For example, suppose a dose D_0 of a drug is administered orally at time $t = 0$. Two compartments might be hypothesized, representing the stomach and the blood system. We model this as a system of differential equations that describe the flows between compartments. Let $S(t)$ be the amount in the stomach at time t and let $B(t)$ be the amount in the bloodstream. We assume that the drug moves from the stomach to the bloodstream at a rate r_{12}, leaves the system from the stomach at rate r_{13}, is reabsorbed from the bloodstream at rate r_{21}, and exits the system from the bloodstream at rate r_{23}. This would give the following set of equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= -(r_{12} + r_{13})S(t) + r_{21}B(t) \\ \frac{dB(t)}{dt} &= r_{12}S(t) - (r_{21} + r_{23})B(t) \\ S(0) &= D_0 \\ B(0) &= 0. \end{aligned} \tag{11.5}$$

Since this is a relatively simple system of differential equations, it can be solved explicitly for the amount of drug in the bloodstream at any time t. The solution is of the form

$$B(t) = \beta(e^{-\lambda_1 t} - e^{-\lambda_2 t}). \tag{11.6}$$

Let y_{ij} represent the amount of the drug found in the jth sample taken from the ith person's bloodstream, which occurred at time t_{ij}. Our model might then be

$$\begin{aligned} y_{ij}|\beta_i, \lambda_{1i}, \lambda_{2i} &\sim \mathcal{N}(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \beta_i(e^{-\lambda_{1i} t_{ij}} - e^{-\lambda_{2i} t_{ij}}) \end{aligned} \tag{11.7}$$

$$\begin{pmatrix} \beta_i \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \sim \text{i.i.d. } \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

which allows the parameters to vary from person to person.

Clearly, this is a nonlinear mixed model: The means are nonlinear in the parameters and cannot be linearized with a fixed transformation. And we are allowing the flow rates between the compartments to vary from person to person.

Models that arise from the solution of a differential equation or a system of such equations are typically nonlinear in the parameters. When we incorporate random effects to model variation from subject to subject or other forms of correlation we end up with nonlinear mixed models. A much more thorough treatment of these topics can be found in Giltinan and Davidian (1995) and Sheiner et al. (1997). Vonesh and Chinchilli (1997) cover the more general topic of nonlinear mixed models in more detail than here.

11.4 COMPUTATIONS FOR NONLINEAR MIXED MODELS

When the data are normally distributed and homoscedastic (or can be transformed to be) then the Laplace approximation methods described in Chapter 10 are more successful for NLMMs than they are for GLMMs in general. This is the basis of computing algorithms implemented in S-Plus and SAS for NLMMs. The conceptual basis for them is described in Giltinan and Davidian (1995), Pinheiro and Bates (1995), and Lindstrom and Bates (1990).

11.5 EXERCISES

E 11.1 Prove that (11.1) can be rewritten as (11.3) using the reparameterization given in (11.2).

E 11.2 Why might we expect the Laplace approximation method of Section 10.3 to work for models such as (11.7) when it fails for some of the models discussed previously?

Appendix M: Some Matrix Results

Readers of this book are assumed to have a working knowledge of matrix algebra. Nevertheless, we provide a few reminders in this appendix.

M.1 VECTORS AND MATRICES OF ONES

Vectors having every element equal to unity are denoted by **1**: Thus $\mathbf{1}_3' = [\ 1 \quad 1 \quad 1\ ]$. With $\mathbf{x}' = [\ x_1 \quad x_2 \quad x_3\ ]$, $\mathbf{1}'\mathbf{x} = \sum_{i=1}^{3} x_i$. The inner product of $\mathbf{1}_n$ with itself is n : $\mathbf{1}_n'\mathbf{1}_n = n$ and outer products of these vectors with each other are matrices having every element unity. They are denoted by **J**. For example,

$$\mathbf{1}_2\mathbf{1}_3' = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \mathbf{J}_{2\times 3}.$$

Square J-matrices are the most common form: $\mathbf{1}_n\mathbf{1}_n' = \mathbf{J}_n$. Products of Js with each other and with 1s are, respectively, Js and 1s multiplied by scalars. For square **J**

$$\mathbf{J}_n^2 = n\mathbf{J}_n \text{ and } \mathbf{J}_n\mathbf{1}_n = n\mathbf{1}_n; \text{ also } \operatorname{tr}(\mathbf{J}_n) = n.$$

Illustration The mean and variance of data $x_1, x_2, \ldots, x_n$ are easily expressed in terms of the preceding matrices. Thus

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{\mathbf{1}'\mathbf{x}}{n} = \frac{\mathbf{x}'\mathbf{1}}{n}, \quad \text{and}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\mathbf{x}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{x}.$$

Linear combinations of $\mathbf{I}$ (an identity matrix) and $\mathbf{J}$ often arise in a variety of circumstances, for which the following results are found useful.

1. $(a\mathbf{I}_n + b\mathbf{J}_n)(\alpha\mathbf{I}_n + \beta\mathbf{J}_n) = a\alpha\mathbf{I}_n + (a\beta + b\alpha + b\beta n)\mathbf{J}_n$.
2. $(a\mathbf{I}_n + b\mathbf{J}_n)^{-1} = \frac{1}{a}\left(\mathbf{I}_n - \frac{b}{a+nb}\mathbf{J}_n\right)$, for $a \neq 0$ and $a \neq -nb$.
3. $|a\mathbf{I}_n + b\mathbf{J}_n| = a^{n-1}(a + nb)$.
4. Eigenroots of $a\mathbf{I}_n + b\mathbf{J}_n$ are a, with multiplicity $n-1$, and $a+nb$.

M.2 KRONECKER (OR DIRECT) PRODUCTS

The Kronecker product of two matrices $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$ is

$$\mathbf{A} \otimes \mathbf{B} = \{a_{ij}\mathbf{B}\}.$$

Examples arising in linear models are

$$\mathbf{I}_2 \otimes \mathbf{1}_3 = \begin{bmatrix} 1(\mathbf{1}_3) & 0(\mathbf{1}_3) \\ 0(\mathbf{1}_3) & 1(\mathbf{1}_3) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{1}_2 \otimes \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Kronecker products have many properties. Assuming conformability

$$\begin{aligned} (\mathbf{A} \otimes \mathbf{B})' &= \mathbf{A}' \otimes \mathbf{B}' \\ (\mathbf{A} \otimes \mathbf{B})^{-1} &= \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \\ (\mathbf{A} \otimes \mathbf{B})(\mathbf{X} \otimes \mathbf{Y}) &= \mathbf{AX} \otimes \mathbf{BY} \\ \text{rank}(\mathbf{A} \otimes \mathbf{B}) &= \text{rank}(\mathbf{A})\text{rank}(\mathbf{B}) \\ \text{tr}(\mathbf{A} \otimes \mathbf{B}) &= \text{tr}(\mathbf{A})\text{tr}(\mathbf{B}) \\ |\mathbf{A}_{a\times a} \otimes \mathbf{B}_{b\times b}| &= |\mathbf{A}|^b|\mathbf{B}|^a. \end{aligned}$$

M.3 A MATRIX NOTATION IN TERMS OF ELEMENTS

Familiar notation for $\mathbf{A}$ of order $p \times q$ is

$$\mathbf{A} = \{a_{ij}\} \text{ for } i = 1, \ldots, p \text{ and } j = 1, \ldots, q,$$

where a_{ij} is the element in row i and column j of $\mathbf{A}$. We abbreviate this to

$$\mathbf{A} = \left\{ {}_m\, a_{ij} \right\}_{i=1,j=1}^{p \quad q} = \left\{ {}_m\, a_{ij} \right\}_{ij} = \left\{ {}_m\, a_{ij} \right\}$$

using only as much detail concerning i and j as is needed for the context. Similarly we use

$$\mathbf{u} = \left\{ {}_c\, u_i \right\}_{i=1}^{q} \quad \text{and} \quad \mathbf{u}' = \left\{ {}_r\, u_i \right\}_{i=1}^{q}$$

for a column and a row, respectively, of elements u_i. Also we use $\left\{ {}_d\, x_i \right\}_{i=1}^{t}$ for a diagonal matrix with t diagonal elements x_i.

The advantage of this notation is, for example, that instead of writing $\mathbf{A} = \{a_{ij}\}$ for $i = 1, \ldots, p$ and $j = 1, \ldots, q$, and $\mathbf{u}$ as a column vector of elements u_i for $i = 1, \ldots, q$ with $\mathbf{Au} = \sum_{j=1}^{q} a_{ij} u_j$ for $i = 1, \ldots, p$, one has no need of the symbols $\mathbf{A}$ and $\mathbf{u}$ but simply writes

$$\left\{ {}_m\, a_{ij} \right\}_{i=1,j=1}^{p \quad q} \left\{ {}_c\, u_j \right\}_{j=1}^{q} = \left\{ {}_c \sum_{j=1}^{q} a_{ij} u_j \right\}_{i=1}^{p}.$$

M.4 GENERALIZED INVERSES

a. Definition

Readers will be familiar with a nonsingular matrix $\mathbf{T}$ being a square matrix that has an inverse $\mathbf{T}^{-1}$ such that $\mathbf{TT}^{-1} = \mathbf{T}^{-1}\mathbf{T} = \mathbf{I}$. More generally, for any non-null matrix $\mathbf{A}$, be it rectangular, or square and singular, there are always matrices $\mathbf{A}^-$ satisfying

$$\mathbf{AA}^-\mathbf{A} = \mathbf{A}. \tag{M.1}$$

When $\mathbf{A}$ is non-singular, (M.1) leads to $\mathbf{A}^- = \mathbf{A}^{-1}$, but otherwise there is an infinite number of matrices $\mathbf{A}^-$ that, for each $\mathbf{A}$, satisfy (M.1). Each such $\mathbf{A}^-$ is called a *generalized inverse* of $\mathbf{A}$.

Example For

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 2 \\ 3 & 7 & 11 & 4 \\ 4 & 9 & 14 & 6 \end{bmatrix}, \quad \mathbf{A}^- = \begin{bmatrix} 7-t & -2-t & t \\ -3+2t & 1+2t & -2t \\ -t & -t & t \\ 0 & 0 & 0 \end{bmatrix}.$$

Calculation of $\mathbf{AA}^-\mathbf{A}$ yields $\mathbf{A}$ no matter what value of t is used, thus illustrating an infinity of matrices $\mathbf{A}^-$ satisfying (M.1).

A great deal has been written about generalized inverse matrices, with much of what is useful for linear models being available in books such as Rao (1962) and Searle (1997) and many others. We direct attention here solely to generalized inverses of $\mathbf{X}'\mathbf{X}$ and their properties, which are extremely useful in solving the normal equations $\mathbf{X}'\mathbf{X}\beta^0 = \mathbf{X}'\mathbf{y}$ of (4.18) or their more general form $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta^0 = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ of (6.19).

b. Generalized inverses of X'X

Clearly $\mathbf{X}'\mathbf{X}$ is square and symmetric; its generalized inverses are denoted by $(\mathbf{X}'\mathbf{X})^-$ and $\mathbf{G}$ interchangeably. Thus $\mathbf{G}$ is defined as

$$\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}. \tag{M.2}$$

Note that although $\mathbf{X}'\mathbf{X}$ is symmetric, $\mathbf{G}$ need not be symmetric. For example,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 7 & 3 & 2 & 2 \\ 3 & 3 & \cdot & \cdot \\ 2 & \cdot & 2 & \cdot \\ 2 & \cdot & \cdot & 2 \end{bmatrix} \text{ has } \mathbf{G} = \begin{bmatrix} 9 & 0 & 0 & 3 \\ 5 & -13\frac{2}{3} & -14 & 17 \\ 1 & -10 & -9\frac{1}{2} & -13 \\ 0 & -9 & -9 & -11\frac{1}{2} \end{bmatrix} \tag{M.3}$$

as a non-symmetric generalized inverse. Despite this, transposing (M.2) shows that when $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$, then so also is $\mathbf{G}'$. As a consequence, as may easily be verified,

$$(\mathbf{X}'\mathbf{X})^- = \mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}' \tag{M.4}$$

is a symmetric generalized inverse of $\mathbf{X}'\mathbf{X}$.

The following theorem is vital for linear model theory.

Theorem M.1. When $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$:

$\mathbf{G}'$ is also a generalized inverse of $\mathbf{X}'\mathbf{X}$. (M.5)

$\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}$. (M.6)

$\mathbf{X}\mathbf{G}\mathbf{X}'$ is the same for every $\mathbf{G}$. (M.7)

$\mathbf{X}\mathbf{G}\mathbf{X}'$ is symmetric, even if $\mathbf{G}$ is not. (M.8)

$\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{1} = \mathbf{1}$ when $\mathbf{1}$ is a column of $\mathbf{X}$. (M.9)

Proof. Condition (M.5) comes from transposing (M.2). Result (M.6) is true because, for real matrices, there is a theorem (e.g., Searle, 1982, p. 63) indicating that if $\mathbf{PX'X} = \mathbf{QX'X}$ then $\mathbf{PX'} = \mathbf{QX'}$; applying this to the transpose of (M.2) and then transposing it yields (M.6); and applying it to $\mathbf{XGX'X} = \mathbf{X} = \mathbf{XFX'X}$ for $\mathbf{F}$ being any other generalized inverse of $\mathbf{X'X}$ yields (M.7). Using $(\mathbf{X'X})^-$ of (M.4) in place of $\mathbf{G}$ in $\mathbf{X'GX}$ demonstrates the symmetry of (M.8) which, by (M.7), therefore holds for any $\mathbf{G}$. Finally, (M.9) follows from considering an individual column of $\mathbf{X}$ in (M.6). *Q.E.D.*

Notice that (M.5) and (M.6) spawn three other results similar to (M.6): $\mathbf{XG'X'X} = \mathbf{X}$, $\mathbf{X'XGX'} = \mathbf{X'}$, and $\mathbf{X'XG'X'} = \mathbf{X'}$.

A particularly useful matrix is $\mathbf{M} = \mathbf{I} - \mathbf{XGX'}$. Theorem M.1 provides the means for verifying that $\mathbf{M}$ has the following properties: $\mathbf{M}$ is symmetric, idempotent, invariant to $\mathbf{G}$, of rank $N - r_{\mathbf{X}}$ when $\mathbf{X}$ has N rows, and its products with $\mathbf{X}$ and $\mathbf{X'}$ are null. Thus

$$\mathbf{M} = \mathbf{M'} = \mathbf{M}^2, r_{\mathbf{M}} = N - r_{\mathbf{X}}, \mathbf{MX} = \mathbf{0} \text{ and } \mathbf{X'M} = \mathbf{0}. \tag{M.10}$$

c. Two results involving $\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^-\mathbf{X'V}^{-1}$

For $\mathbf{V}$ being symmetric and positive definite [as it usually is when it is var(**y**)]

$$\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^-\mathbf{X'V}^{-1} \text{ is invariant to } (\mathbf{X'V}^{-1}\mathbf{X})^- \tag{M.11}$$

and

$$\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^-\mathbf{X'V}^{-1}\mathbf{X} = \mathbf{X}.$$

Proof of these results stems from the nature of $\mathbf{V}$ ($= \mathbf{V'}$ and p.s.d.) enabling us to write $\mathbf{V}^{-1} = \mathbf{L'L}$ for some $\mathbf{L}$. Then

$$\begin{aligned}
\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^-\mathbf{X'V}^{-1} &= \mathbf{VV}^{-1}\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^-\mathbf{X'V}^{-1} \\
&= \mathbf{VL'LX}(\mathbf{X'L'LX})^-\mathbf{X'L'L} \\
&= \mathbf{VL'T}(\mathbf{T'T})^-\mathbf{T'L} \text{ for } \mathbf{T} = \mathbf{LX}
\end{aligned}$$

which by (M.6) is invariant to the generalized inverse.

Also

$$\begin{aligned}
\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^-\mathbf{X'V}^{-1}\mathbf{X} &= \mathbf{VL'T}(\mathbf{T'T})^-\mathbf{T'T} \\
&= \mathbf{VL'T} \qquad \text{by (M.6)} \\
&= \mathbf{X}. \qquad\qquad\qquad\qquad (\text{M.12})
\end{aligned}$$

d. Solving linear equations

Rao (1962) shows that equations $\mathbf{Ax} = \mathbf{y}$ have solutions $\mathbf{A}^-\mathbf{y} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{z}$ for any $\mathbf{z}$ of the appropriate order. The simplest application of this to

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}'\mathbf{y} \text{ is } \boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$$

for any $(\mathbf{X}'\mathbf{X})^-$. Equation (M.7) ensures that $\mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$ is invariant to $(\mathbf{X}'\mathbf{X})^-$. Extension to $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is obvious.

e. Rank results

The standard result for the rank of a product matrix is $r_{\mathbf{AB}} \leq r_{\mathbf{B}}$. Thus using $r(\mathbf{X})$ and $r_{\mathbf{X}}$ interchangeably to represent the rank of $\mathbf{X}$, we have $r(\mathbf{AA}^-) \leq r_{\mathbf{A}}$; and from $\mathbf{A} = \mathbf{AA}^-\mathbf{A}$ we have $r_{\mathbf{A}} \leq r(\mathbf{AA}^-)$. Therefore $r(\mathbf{AA}^-) = r_{\mathbf{A}}$. And so, because (M.5) shows that $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ is a generalized inverse of $\mathbf{X}$, $r[\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'] = r_{\mathbf{X}}$.

f. Vectors orthogonal to columns of X

Suppose $\mathbf{k}'$ is such that $\mathbf{k}'\mathbf{X} = \mathbf{0}$. Then $\mathbf{X}'\mathbf{k} = 0$ and, from the theory of solving linear equations (e.g., Searle, 1982, Sec. 9.4b), $\mathbf{k} = [\mathbf{I} - (\mathbf{X}')^-\mathbf{X}']\mathbf{c}$ for any vector $\mathbf{c}$, of the appropriate order. Therefore, since $(\mathbf{X}')^-$ is a generalized inverse of $\mathbf{X}'$ we can write $\mathbf{k}' = \mathbf{c}'(\mathbf{I} - \mathbf{XX}^-)$. Moreover, because $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ is a generalized inverse of $\mathbf{X}$, another form for $\mathbf{k}'$ is $\mathbf{k}' = \mathbf{c}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}']$. Thus two forms of $\mathbf{k}$ are

$$\mathbf{k}' = \mathbf{c}'[\mathbf{I} - \mathbf{XX}^-] \quad \text{or} \quad \mathbf{k}' = \mathbf{c}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'] = \mathbf{c}'\mathbf{M},$$

for $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ of Section M.4b.

With $\mathbf{X}$ of order $N \times p$ of rank r, there are only $N - r$ linearly independent vectors $\mathbf{k}'$ satisfying $\mathbf{k}'\mathbf{X} = \mathbf{0}$ (e.g., Searle, 1982, Sec. 9.7a). Using a set of such $N - r$ linearly independent vectors $\mathbf{k}'$ as rows of $\mathbf{K}'$, we then have the following theorem, for $\mathbf{K}'\mathbf{X} = \mathbf{0}$ with $\mathbf{K}'$ having maximum row rank $N - r$, and with $\mathbf{K}' = \mathbf{C}'\mathbf{M}$ for some $\mathbf{C}$.

g. A theorem for K′ with K′X being null

Theorem M.2. If $\mathbf{K}'\mathbf{X} = \mathbf{0}$, where $\mathbf{K}'$ has maximum row rank and $\mathbf{V}$ is positive definite then

$$\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}' = \mathbf{P}$$

for

$$\mathbf{P} \equiv \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}.$$

The proof of this is lengthy and technical and we do not show it here; it can be found in VC p. 452.

M.5 DIFFERENTIAL CALCULUS

a. Definition

Differentiation with respect to elements of a vector $\mathbf{x}=\left\{ {}_{c}\ x_i \right\}_{i=1}^{k}$ is defined by the notation

$$\frac{\partial}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_k} \end{bmatrix}.$$

b. Scalars

Thus

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}'\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{a}) = \mathbf{a}. \tag{M.13}$$

c. Vectors

For $\mathbf{y}' = [\ y_1 \quad y_2 \quad \cdots \quad y_p\ ]$

$$\frac{\partial \mathbf{y}'}{\partial \mathbf{x}} = \left\{ {}_{m}\ \frac{\partial y_j}{\partial x_i} \right\}_{i=1,j=1}^{k \quad p}, \text{ a matrix of order } k \times p.$$

Then

$$\frac{\partial \mathbf{x}'}{\partial \mathbf{x}} = \mathbf{I} \tag{M.14}$$

and for $\mathbf{A}$ not involving $\mathbf{x}$

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}) = \frac{\partial \mathbf{x}'}{\partial \mathbf{x}}\mathbf{A} = \mathbf{A}. \tag{M.15}$$

d. Inner products

Consider $\mathbf{u}$ and $\mathbf{v}$, of the same order, each having elements that are functions of the elements of $\mathbf{x}$. Then $\mathbf{u}'\mathbf{v}$ is a scalar, and so by (M.13)

$\partial(\mathbf{u}'\mathbf{v})/\partial\mathbf{x}$ is a column. Therefore, because differentiating the $\mathbf{u}'$ part of $\mathbf{u}'\mathbf{v}$ gives $(\partial\mathbf{u}'/\partial\mathbf{x})\mathbf{v}$ and because $\mathbf{u}'\mathbf{v} = \mathbf{v}'\mathbf{u}$, we have

$$\frac{\partial \mathbf{u}'\mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}'}{\partial \mathbf{x}}\mathbf{v} + \frac{\partial \mathbf{v}'}{\partial \mathbf{x}}\mathbf{u}. \tag{M.16}$$

e. Quadratic forms

To differentiate $\mathbf{x}'\mathbf{A}\mathbf{x}$ with respect to $\mathbf{x}$, use (M.16) with $\mathbf{u}'$ and $\mathbf{v}$ being $\mathbf{x}'$ and $\mathbf{A}\mathbf{x}$ respectively. This gives

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}}\mathbf{x}'\mathbf{A}\mathbf{x} &= \frac{\partial \mathbf{x}'}{\partial \mathbf{x}}\mathbf{A}\mathbf{x} + \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}}\mathbf{x} \\ &= \mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{x} \\ &= 2\mathbf{A}\mathbf{x} \text{ when } \mathbf{A} \text{ is symmetric,} \end{aligned} \tag{M.17}$$

which it usually is.

f. Inverse matrices

If $\mathbf{V}$ is non-singular of order n and has elements which are functions of a scalar w, differentiating $\mathbf{V}^{-1}$ with respect to w comes from differentiating the identity $\mathbf{V}^{-1}\mathbf{V} = \mathbf{I}$. Thus

$$\frac{\partial \mathbf{V}^{-1}}{\partial w}\mathbf{V} + \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial w} = \mathbf{0}$$

and so

$$\frac{\partial \mathbf{V}^{-1}}{\partial w} = -\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial w}\mathbf{V}^{-1} \tag{M.18}$$

where

$$\frac{\partial \mathbf{V}}{\partial w} = \left\{ m \; \frac{\partial v_{ij}}{\partial w} \right\}_{i,j=1}^{n}.$$

Note that (M.18) is a special case of (6.75) for generalized inverses.

Finally, using $\mathbf{P} = \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'$ note that

$$\begin{aligned} \frac{\partial \mathbf{P}}{\partial w} &= -\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\frac{\partial \mathbf{V}}{\partial w}\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}' \\ &= -\mathbf{P}\frac{\partial \mathbf{V}}{\partial w}\mathbf{P}. \end{aligned} \tag{M.19}$$

g. Determinants

$$\frac{\partial \log |\mathbf{V}|}{\partial w} = \frac{1}{|\mathbf{V}|}\frac{\partial |\mathbf{V}|}{\partial w} = \sum_i \sum_j \frac{|\mathbf{V}_{ij}|}{|\mathbf{V}|}\frac{\partial v_{ij}}{\partial w}$$

$$= \sum_i \sum_j v^{ij}\frac{\partial v_{ij}}{\partial w} = \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial w}\right), \qquad \text{(M.20)}$$

with v_{ij} being an element of $\mathbf{V}$, $|\mathbf{V}_{ij}|$ its cofactor, and $v^{ij} = |\mathbf{V}_{ij}|/|\mathbf{V}|$ representing an element of $\mathbf{V}^{-1}$. The last step arises from $\mathbf{V}$ being symmetric. Some intermediate details are given in VC pp. 456–457.

Appendix S: Some Statistical Results

As with Appendix M, we assume that a reader's background knowledge includes familiarity with basic mathematical statistics. Nevertheless, here are a few reminders.

S.1 MOMENTS

a. Conditional moments

For random variables $\mathbf{y}$ and $\mathbf{u}$, let $f(\mathbf{y},\mathbf{u})$ and $f(\mathbf{y}|\mathbf{u})$ denote, respectively, the joint density of $\mathbf{y}$ and $\mathbf{u}$, and the density of $\mathbf{y}$ conditional on $\mathbf{u}$. Also, let E_y and var_y denote expectation and variance with respect to the distribution of $\mathbf{y}$. There are three well-established results (Searle et al., 1992):

$$\mathrm{E}[\mathbf{y}] = \mathrm{E}_{y,u}[\mathbf{y}] = \mathrm{E}_u\left[\mathrm{E}[\mathbf{y}|\mathbf{u}]\right] \tag{S.1}$$

$$\mathrm{cov}(\mathbf{y},\mathbf{w}) = \mathrm{E}_u[\mathrm{cov}(\mathbf{y},\mathbf{w}|\mathbf{u})] + \mathrm{cov}_u\left(\mathrm{E}[\mathbf{y}|\mathbf{u}],\mathrm{E}[\mathbf{w}|\mathbf{u}]\right), \tag{S.2}$$

and using the latter with $\mathbf{w} = \mathbf{y}$,

$$\mathrm{var}(\mathbf{y}) = \mathrm{E}_u[\mathrm{var}(\mathbf{y}|\mathbf{u})] + \mathrm{var}_u\left(\mathrm{E}[\mathbf{y}|\mathbf{u}]\right). \tag{S.3}$$

(S.1) is established as follows

$$\begin{aligned}\mathrm{E}[\mathbf{y}] &= \int\int \mathbf{y} f(\mathbf{y}|\mathbf{u}) f(\mathbf{u})\, d\mathbf{y}\, d\mathbf{u} \\ &= \int \mathrm{E}_y[\mathbf{y}|\mathbf{u}] d\mathbf{u} = \mathrm{E}_u\left[\mathrm{E}_y[\mathbf{y}|\mathbf{u}]\right].\end{aligned} \tag{S.4}$$

When $\mathbf{y}$ of $\mathrm{E}[\mathbf{y}]$ is replaced by $(\mathbf{y}-\mathrm{E}[\mathbf{y}])(\mathbf{w}-\mathrm{E}[\mathbf{w}])$, the left-hand side of (S.1) becomes cov(y,w). That same replacement on the right-hand

side of (S.1) followed by some tedious algebra (see VC, p. 462), yields (S.2). And then replacing $\mathbf{w}$ by $\mathbf{y}$ in (S.2) gives it as

$$\begin{aligned} \text{var}(\mathbf{y}) &= \text{E}_u[\text{cov}(\mathbf{y}, \mathbf{y}|\mathbf{u})] + \text{cov}_u\left(\text{E}[\mathbf{y}|\mathbf{u}], \text{E}[\mathbf{y}|\mathbf{u}]\right) \\ &= \text{E}_u[\text{var}(\mathbf{y}|\mathbf{u})] + \text{var}_u\left(\text{E}[\mathbf{y}|\mathbf{u}]\right). \end{aligned} \tag{S.5}$$

which is (S.3).

b. Mean of a quadratic form

Suppose $\text{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\text{var}(\mathbf{y}) = \mathbf{V}$, i.e., $\mathbf{y} \sim (\boldsymbol{\mu}, \mathbf{V})$, not necessarily normally distributed. Then, for a quadratic form in $\mathbf{y}$, we have

$$\begin{aligned} \text{E}[\mathbf{y}'\mathbf{A}\mathbf{y}] &= \text{E}[\text{tr}(\mathbf{y}'\mathbf{A}\mathbf{y})] = \text{E}[\text{tr}(\mathbf{A}\mathbf{y}\mathbf{y}')] \\ &= \text{tr}(\mathbf{A}\text{E}[\mathbf{y}\mathbf{y}']) = \text{tr}(\mathbf{A}[\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}']) \\ &= \text{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \end{aligned} \tag{S.6}$$

c. Moment generating function

The moment generating function (m.g.f.) of a random variable $\mathbf{y}$ is a function [carefully defined; see Casella and Berger (1990, p. 61)] of a mathematical variable t. For y having a density $f(y)$ the m.g.f. is

$$\text{M}_Y(t) = \text{E}[e^{ty}] = \int_{-\infty}^{\infty} e^{ty} f(y)\, dy. \tag{S.7}$$

This yields the rth moment (about zero) of y as

$$\mu_Y^{(r)} = \left.\frac{\partial^r \text{M}_Y(t)}{\partial t^r}\right|_{t=0}. \tag{S.8}$$

Similarly, for a function $h(y)$ of y

$$\text{M}_{h(Y)}(t) = \text{E}[e^{th(y)}] \qquad \text{and} \qquad \mu_{h(Y)}^{(r)} = \left.\frac{\partial^r \text{M}_{h(Y)}(t)}{\partial t^r}\right|_{t=0}. \tag{S.9}$$

For a vector random variable $\mathbf{y}$, the m.g.f. is

$$\text{M}_Y(\mathbf{t}) = \text{E}[e^{t'y}]. \tag{S.10}$$

S.2 NORMAL DISTRIBUTIONS

a. Univariate

The scalar random variable y is said to be normally distributed with mean μ and variance σ^2 when it has probability density function

$$f(y) = \frac{e^{-\frac{1}{2}(y-\mu)^2/\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

We represent this as $y \sim \mathcal{N}(\mu, \sigma^2)$.

b. Multivariate

The vector of n random variables $\mathbf{y}' = [\ y_1 \quad y_2 \quad \dots \quad y_n\ ]$ is said to have a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and non-singular variance-covariance matrix $\mathbf{V}$ when it has probability density function

$$f(\mathbf{y}) = \frac{e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})}}{(2\pi)^{n/2}|\mathbf{V}|^{\frac{1}{2}}}. \tag{S.11}$$

This is represented as $\mathbf{y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$, often with the subscript n omitted when it is evident from the context. Many texts (e.g., Searle, 1997) have numerous details about these distributions, so we summarize just some of the properties of the multivariate normal, mostly those which are useful to the purposes of this book.

For $\mathbf{y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{V})$:

1. $\mathrm{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{y}) = \mathbf{V}$.
2. $\mathbf{Ky} \sim \mathcal{N}(\mathbf{K}\boldsymbol{\mu}, \mathbf{KVK}')$.

On writing

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \right),$$

3. the marginal distribution of $\mathbf{y}_1$ is

$$\mathbf{y}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{V}_{11}),$$

4. and the conditional distribution of $\mathbf{y}_1|\mathbf{y}_2$ is

$$\mathbf{y}_1|\mathbf{y}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right).$$

5. The moment generating function is

$$M_Y(\mathbf{t}) = E[e^{\mathbf{t}'\mathbf{y}}] = e^{\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}}.$$

6. $\text{var}(\mathbf{y}'\mathbf{A}\mathbf{y}) = 2\text{tr}[(\mathbf{A}\mathbf{V})^2] + 4\boldsymbol{\mu}'\mathbf{A}\mathbf{V}\mathbf{A}\boldsymbol{\mu}$.

Extensive details of deriving these, especially properties 4, 5 and 6, are available in Searle (1971, Chap. 2).

c. Quadratic forms in normal variables

Section S.1b shows the derivation of $E[\mathbf{y}'\mathbf{A}\mathbf{y}]$ no matter what the distribution of $\mathbf{y}$ is. When that distribution is normal, $\mathbf{y}'\mathbf{A}\mathbf{y}$ has three very useful properties, the first of which requires the prelude of describing the non-central chi-square (χ^2) distribution.

– i. *The non-central* χ^2

For $\mathbf{y}_{n\times 1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have the well-known result that $\sum_i y_i^2 = \mathbf{y}'\mathbf{y}$ is distributed as chi-square on n degrees of freedom, i.e., $\mathbf{y}'\mathbf{y} \sim \chi^2_n$. A well-known variant of this is that when $\mathbf{y}_{n\times 1} \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2\mathbf{I})$ then $\sum_{i=1}^n (y_i - \bar{y})^2 \sim \chi^2_{n-1}$. An extension of these two cases is when $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$. Then the resulting distribution of $\mathbf{y}'\mathbf{y}$ is known as the non-central chi-square. It is akin to the customary χ^2 (now called the central χ^2), with degrees of freedom n, but with a second parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu}$, known as the non-centrality parameter. And when $\boldsymbol{\mu} = \mathbf{0}$ the non-central chi-square [denoted by $\chi^{2\prime}(n, \lambda)$] reduces to being the central χ^2.

– ii. *Properties of* $\mathbf{y}'\mathbf{A}\mathbf{y}$ *when* $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$

When $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$

$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^{2\prime}(r_{\mathbf{A}}, \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$ if and only if $\mathbf{A}\mathbf{V}$ is idempotent;

$\mathbf{y}'\mathbf{A}\mathbf{y}$ and $\mathbf{y}'\mathbf{B}\mathbf{y}$ are independent if and only if $\mathbf{A}\mathbf{V}\mathbf{B} = \mathbf{0}$.

Details and proofs of these widely-known results can be found in Searle (1997, Chap. 2). The sufficient condition in each is easily proven, whereas the necessity conditions are not. Driscoll and Gundberg (1986) and Driscoll and Krasnicka (1995) have an interesting history of these necessity conditions.

Two further results for $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ are

$$\begin{aligned} \operatorname{var}(\mathbf{y}'\mathbf{A}\mathbf{y}) &= 2\operatorname{tr}[(\mathbf{A}\mathbf{V})^2] + 4\boldsymbol{\mu}'\mathbf{A}\mathbf{V}\mathbf{A}\boldsymbol{\mu} \text{ and} \\ \operatorname{cov}(\mathbf{y}'\mathbf{A}\mathbf{y}, \mathbf{y}'\mathbf{B}\mathbf{y}) &= 2\operatorname{tr}(\mathbf{A}\mathbf{V}\mathbf{B}\mathbf{V}). \end{aligned}$$

The first of these two is a special case of the more general result that the kth cumulant of $\mathbf{y}'\mathbf{A}\mathbf{y}$ is $2^{k-1}(k-1)![\operatorname{tr}(\mathbf{A}\mathbf{V})^k + k\boldsymbol{\mu}'\mathbf{A}(\mathbf{V}\mathbf{A})^{k-1}\boldsymbol{\mu}]$. And the second of the two comes from applying the first to $\operatorname{var}[\mathbf{y}'(\mathbf{A}+\mathbf{B})\mathbf{y}]$.

S.3 EXPONENTIAL FAMILIES

Probability densities which can be written in the form

$$\begin{aligned} f(\mathbf{y};\boldsymbol{\theta}) &= h(\mathbf{y})d(\boldsymbol{\theta})\exp\{\boldsymbol{\nu}(\boldsymbol{\theta})'\mathbf{T}(\mathbf{y})\} \\ &= h(\mathbf{y})d(\boldsymbol{\theta})\exp\{\sum_{i=1}^{k}\nu_i(\boldsymbol{\theta})'T_i(\mathbf{y})\}, \end{aligned} \tag{S.12}$$

are said to constitute an *exponential family*. Many of the commonly-used distributions are of this form: for example, normal, gamma, beta, binomial, and Poisson. An important consequence of the form (S.12) is that the sufficient statistics are $[T_1(\mathbf{y}), T_2(\mathbf{y}), \ldots, T_k(\mathbf{y})]'$.

S.4 MAXIMUM LIKELIHOOD

a. The likelihood function

Suppose a vector of random variables, $\mathbf{y}$, has density function $f(\mathbf{y})$. Let $\boldsymbol{\theta}$ be the vector of parameters involved in $f(\mathbf{y})$. Then $f(\mathbf{y})$ is a function of both $\mathbf{y}$ and $\boldsymbol{\theta}$. As a result, it can be viewed in two different ways. The first is (as above) as a density function, in which case $\boldsymbol{\theta}$ is usually assumed to be known. With this in mind we use the symbol $f(\mathbf{y}|\boldsymbol{\theta})$ in place of $f(\mathbf{y})$ to emphasize that $\boldsymbol{\theta}$ is being taken as known.

A second viewpoint is where $\mathbf{y}$ represents a known vector of data and where $\boldsymbol{\theta}$ is unknown. Then $f(\mathbf{y})$ will be a function of just $\boldsymbol{\theta}$. It is called the *likelihood function* for the data $\mathbf{y}$; and because in this context $\boldsymbol{\theta}$ is unknown and $\mathbf{y}$ is known we use the notation $L(\boldsymbol{\theta}|\mathbf{y})$ or just $L(\boldsymbol{\theta})$ or even just L. Thus, although $f(\mathbf{y}|\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}|\mathbf{y})$ represent the same thing mathematically, i.e.,

$$f(\mathbf{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y}),$$

it is convenient to use each in its appropriate context.

b. Maximum likelihood estimation

The likelihood function $L(\boldsymbol{\theta}|\mathbf{y})$ is the foundation of the widely-used method of estimation known as *maximum likelihood estimation*. It yields estimators that have many good properties. ML is used as an abbreviation for maximum likelihood and MLE for maximum likelihood estimate—with whatever suffix is appropriate to the context: estimate, estimator (and their plurals) or estimation.

The essence of the ML method is to view $L(\boldsymbol{\theta}|\mathbf{y})$ as a function of the mathematical variable $\boldsymbol{\theta}$ and to derive $\hat{\boldsymbol{\theta}}$ as the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\mathbf{y})$. The only proviso is that this maximization must be carried out within the range of permissible values for $\boldsymbol{\theta}$. For example, if one element of $\boldsymbol{\theta}$ is a variance then permissible values for that variance are non-negative values. This aspect of ML estimation is very important in estimating variances of random effects.

Under widely existing regularity conditions on $f(\mathbf{y}|\boldsymbol{\theta})$, a general method of establishing equations that yield MLEs is to differentiate $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and equate the derivative to $\mathbf{0}$. But finding the values of $\boldsymbol{\theta}$ that maximize L is equivalent to maximizing $\log L$, which we denote by l, and it is often easier to use l rather than L. Thus for $l = \log L(\boldsymbol{\theta}|\mathbf{y})$ the equations

$$\left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\dot{\boldsymbol{\theta}}} = \mathbf{0} \tag{S.13}$$

are known as the ML equations, with $\dot{\boldsymbol{\theta}}$, their solution, being called the ML solution. When this solution is the global maximum and is in the parameter space it is also the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}$. When it is not within the permissible range, then adjustments must be made to the solution to find the MLE; these adjustments depend on the context and form of $f(\mathbf{y}|\boldsymbol{\theta})$.

c. Asymptotic variance-covariance matrix

A useful property of the ML estimator, $\hat{\boldsymbol{\theta}}$, is that its large-sample, or asymptotic, variance-covariance matrix is easy to calculate. From $\mathbf{I}(\boldsymbol{\theta})$, known as the information (or Fisher information) matrix, and defined as

$$\mathbf{I}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\frac{\partial l}{\partial \boldsymbol{\theta}'}\right] = \mathrm{E}\left[\left\{{}_m\,\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right\}_{i,j}\right], \tag{S.14}$$

the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\text{var}(\hat{\boldsymbol{\theta}}) \approx [\mathbf{I}(\boldsymbol{\theta})]^{-1}. \tag{S.15}$$

Note that this is available without even needing a formula or the sampling distribution of $\hat{\boldsymbol{\theta}}$. An alternative form of the information matrix that is valid in many situations is

$$\mathbf{I}(\boldsymbol{\theta}) = -\text{E}\left[\frac{\partial^2 l}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right] = -\text{E}\left[\left\{{}_m\,\frac{\partial^2 l}{\partial\theta_i\,\partial\theta_j}\right\}_{i,j}\right]. \tag{S.16}$$

Proof of this is widely available (e.g., Searle et al., 1992, p. 473).

d. Asymptotic distribution of MLEs

No matter what the distribution of one's data vector $\mathbf{y}$, it is ordinarily the case for an MLE that, as the sample size increases, the MLE of $\boldsymbol{\theta}$ is consistent and asymptotically normally distributed with mean $\boldsymbol{\theta}$ and variance $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$; we summarize this by writing

$$\hat{\boldsymbol{\theta}} \sim \mathcal{AN}\left(\boldsymbol{\theta}, [\mathbf{I}(\boldsymbol{\theta})]^{-1}\right), \tag{S.17}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is given in (S.14) or (S.16).

S.5 LIKELIHOOD RATIO TESTS

The likelihood ratio test is a standard test for composite hypotheses. It has the advantage of an easily derived large–sample distribution. Suppose the parameter vector $\boldsymbol{\theta}$ is partitioned into two components $\boldsymbol{\theta}' = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2']$ and suppose interest focuses on $\boldsymbol{\theta}_1$ while $\boldsymbol{\theta}_2$ is left unspecified. $\boldsymbol{\theta}_2$ is often called a *nuisance parameter*. Either or both of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ could be vector-valued and, if the entire parameter vector is of interest, $\boldsymbol{\theta}_2$ would be null.

Suppose our hypothesis is of the form $H_0: \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$, where $\boldsymbol{\theta}_{1,0}$ is a specified value of $\boldsymbol{\theta}_1$, and let $\hat{\boldsymbol{\theta}}_{2,0}$ be the MLE of $\boldsymbol{\theta}_2$ under the restriction that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$.

With $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ being the likelihood, the likelihood ratio statistic is

$$\Lambda = \frac{L(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0})}{L(\hat{\boldsymbol{\theta}})}. \tag{S.18}$$

The test is to reject H_0 when $\Lambda \leq k$ and with k determined such that

$$\sup_{\theta_2} \mathrm{P}\{\Lambda \leq k | \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}\} \leq \alpha.$$

An equivalent rejection region is when

$$-2\log\Lambda \geq -2\log k \equiv k^*.$$

Under regularity conditions, a notable one being that $\boldsymbol{\theta}_{1,0}$ is not on the boundary of the parameter space, the large-sample distribution of $-2\log\Lambda$ is χ^2 with degrees of freedom equal to ν, the dimension of $\boldsymbol{\theta}_1$. The large-sample value of k^* is then given by $\chi^2_{\nu,1-\alpha}$. This can be written in terms of the log likelihood, l, as follows:

$$-2\log\Lambda = -2\left[l(\boldsymbol{\theta}_{1,0}, \hat{\boldsymbol{\theta}}_{2,0}) - l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right] \tag{S.19}$$

with the large-sample critical region of the test given by

$$-2\log\Lambda > \chi^2_{\nu,1-\alpha}. \tag{S.20}$$

If $\boldsymbol{\theta}_{1,0}$ *is* on the boundary of the parameter space then special care must be taken. This can arise in the analysis of random effects since we may be interested in testing the null hypothesis that the variance of the random effect is zero. See Self and Liang (1987) for details.

S.6 MLE UNDER NORMALITY

In this section we consider maximum likelihood estimation under the linear model: $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.

a. Estimation of $\boldsymbol{\beta}$

If we assume $\mathbf{V}$ to be known then, from (S.11), with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$

$$l = \log L = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{S.21}$$

In Section 6.3 we use the general results of Section 6.12 to differentiate l with respect to $\boldsymbol{\beta}$. Now we confirm that result by differentiating l of (S.21) using the rules in Section M.5:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{S.22}$$

Equating this to $\mathbf{0}$ and using $\boldsymbol{\beta}^0$ for the solution gives $\boldsymbol{\beta}^0$ of (6.19):

$$\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{S.23}$$

And then, as in (6.20),

$$\mathrm{ML}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{S.24}$$

which, by (M.11), is invariant to the choice of $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-$.

It is to be noted in passing when $\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, whether normally distributed or not, that (S.24) is the generalized least squares estimator (GLSE) of $\mathbf{X}\boldsymbol{\beta}$. Moreover, if $\mathbf{V} = \sigma^2\mathbf{I}$, then (S.24) simplifies to

$$\mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} \tag{S.25}$$

which is known as the ordinary least squares estimator (OLSE) of $\mathbf{X}\boldsymbol{\beta}$.

b. Estimation of variance components

Equation (6.60) for obtaining $\mathrm{ML}(\sigma^2)$ is

$$\left\{ {}_c\,\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i') \right\}_{i=0}^{r} = \left\{ {}_c\,\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} \right\}_{i=0}^{r}. \tag{S.26}$$

c. Asymptotic variance-covariance matrix

The asymptotic variance of the ML estimators $\mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ is shown in Section 6.8c. Using (S.16) the terms for $\mathbf{I}\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \sigma^2 \end{pmatrix}$ are

$$-\mathrm{E}\left[\frac{\partial^2 l}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right] = \mathrm{E}\left[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right] = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \tag{S.27}$$

and based on $\partial l/\partial\sigma_i^2$ following (6.57)

$$\begin{aligned} -\mathrm{E}\left[\frac{\partial^2 l}{\partial\boldsymbol{\beta}\,\partial\sigma_i^2}\right] &= \mathrm{E}\left[\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{y}\right] \\ &= \mathbf{0}, \text{ since } \mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{S.28}$$

And

$$\begin{aligned} -\mathrm{E}\left[\frac{\partial^2 l}{\partial\sigma_j^2\,\partial\sigma_i^2}\right] = \mathrm{E}\big[&\tfrac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i') \\ &- \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &- \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big]. \end{aligned}$$

The last term is a scalar and so equals its transpose, which is the penultimate term. Moreover, a scalar equals its own trace. Thus the expected value of the sum of those two (equal) terms is

$$\begin{aligned} -\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathrm{E}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\right]\right) \\ = \; -\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{V}\right) \\ = \; -\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right). \end{aligned} \tag{S.29}$$

Thus

$$-\mathrm{E}\left[\frac{\partial^2 l}{\partial\sigma_j^2\,\partial\sigma_j^2}\right] = -\tfrac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right). \tag{S.30}$$

From these results we get (6.62), (6.63), and (6.64).

d. Restricted maximum likelihood (REML)

The underlying concept of restricted maximum likelihood (REML) is described at the beginning of Section 6.9. In that section we give a wholly technical derivation of the REML methodology whereas here we describe the derivation from basic principles. That involves estimating variance components from linear combinations of the data that do not involve $\boldsymbol{\beta}$. This is achieved by using maximum likelihood on $\mathbf{K}'\mathbf{y}$ where $\mathbf{K}'$ is chosen to have as many linearly independent rows as possible satisfying $\mathbf{K}'\mathbf{X} = 0$. This results in $\mathbf{K}'$ having row rank $r_{\mathbf{K}} = N - r_{\mathbf{X}}$. Then with

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \;\text{ we have }\; \mathbf{K}'\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}'\mathbf{V}\mathbf{K}). \tag{S.31}$$

– i. *Estimation*

Using (S.31) the log likelihood of $\mathbf{K}'\mathbf{y}$ is

$$l_* = \tfrac{1}{2}r_{\mathbf{K}}\log 2\pi - \tfrac{1}{2}\log|\mathbf{K}'\mathbf{V}\mathbf{K}| - \tfrac{1}{2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}. \tag{S.32}$$

To apply maximum likelihood we need

$$\begin{aligned} \frac{\partial l}{\partial\sigma_i^2} = & -\tfrac{1}{2}\mathrm{tr}\left[(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{K}\right] \\ & + \tfrac{1}{2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}. \end{aligned} \tag{S.33}$$

Equating this to zero, and using $\mathbf{P}= \mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'$ of Section M.4g in doing so, gives

$$\mathrm{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') = \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}.$$

This equation written for each $i = 0, 1, \ldots, r$ is equation (6.66) which is the REML procedure.

– ii. *Asymptotic variance*

Using (S.33) together with (M.19) of Section M.5f gives

$$\frac{\partial^2 l}{\partial\sigma_j^2\,\partial\sigma_i^2} = -\tfrac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') \qquad \text{(S.34)}$$

$$- \tfrac{1}{2}\left[\mathbf{y}'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} + \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{y}\right].$$

Each quadratic in $\mathbf{y}$ is a scalar and so equals its transpose; hence the last two terms are equal.

To derive the asymptotic variance we first note that, for any $\mathbf{A}$,

$$\begin{aligned} \mathrm{E}\left[\mathrm{tr}(\mathbf{y}'\mathbf{P}\mathbf{A}\mathbf{P}\mathbf{y})\right] &= \mathrm{tr}(\mathbf{A}\mathbf{P}\mathrm{E}\left[\mathbf{y}\mathbf{y}'\right]\mathbf{P}) \\ &= \mathrm{tr}(\mathbf{A}\mathbf{P}[\mathbf{V} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}']\mathbf{P}) \qquad \text{(S.35)} \\ &= \mathrm{tr}(\mathbf{A}\mathbf{P}) \text{ because } \mathbf{P}\mathbf{V}\mathbf{P} = \mathbf{P} \text{ and } \mathbf{P}\mathbf{X} = \mathbf{0}. \end{aligned}$$

Thus on using (S.34) in

$$\mathbf{I}(\boldsymbol{\sigma}^2) = -\left\{ \underset{m}{\mathrm{E}}\left[\frac{\partial^2 l}{\partial\sigma_j^2\,\partial\sigma_i^2}\right]\right\}_{i,j=0}^{r},$$

and then applying (S.35) with $\mathbf{A} = \mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'$, we get

$$\begin{aligned} \mathbf{I}(\boldsymbol{\sigma}^2) &= \tfrac{1}{2}\left\{ \underset{m}{\mathrm{tr}}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P})\right\}_{i,j=0}^{r} \\ &= \tfrac{1}{2}\left\{ \underset{m}{\mathrm{tr}}(\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i[\mathbf{Z}_j'\mathbf{P}\mathbf{Z}_i]')\right\}_{i,j=0}^{r} \qquad \text{(S.36)} \end{aligned}$$

leading to

$$\mathrm{var}_\infty(\hat{\boldsymbol{\sigma}}^2_{\mathrm{REML}}) = [I(\boldsymbol{\sigma}^2)]^{-1}. \qquad \text{(S.37)}$$

References

Abramowitz, M. and I. Stegun (eds.) (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington, D.C.

Abu-Libdeh, H., B. W. Turnbull, and L. C. Clark (1990). Analysis of multi-type recurrent events in longitudinal studies: Application to a skin cancer prevention trial. *Biometrics*, **46**:1017–1034.

Arnold, S. (1981). *The Theory of Linear Models and Multivariate Analysis*. Wiley, New York.

Atwill, E., H. O. Mohammed, J. W. Lopez, C. E. McCulloch, and E. J. Dubovi (1996). Cross-sectional evaluation of environmental, host, and management factors associated with the risk of seropositivity to *Ehrlichia risticii* in horses of New York State. *American Journal of Veterinary Research*, **57**:278–285.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.

Bliss, C. (1934). The method of probits. *Science*, **79**:38–39.

Bliss, C. (1935). The calculation of the dose-mortality curve. *Annals of Applied Biology*, **22**:134–167.

Blyth, C. R. and H. A. Still (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, **78**:108–116.

Booth, J. G. and J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**:265–285.

Box, G. E. P. and D. R. Cox (1962). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**:211–252.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**:9–25.

Breslow, N. E. and X. Lin (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**:81–91.

Casella, G. and R. L. Berger (1990). *Statistical Inference*. Wadsworth and Brooks/Cole, Pacific Grove, Calif.

Chakravorti, S. R. and J. E. Grizzle (1975). Analysis of data from multiclinic experiments. *Biometrics*, **31**:325–338.

Churchill, G. (1995). Personal communication. Cornell University.

Cochran, W. G. (1951). Improvement by means of selection. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*. University of California Press, Berkeley and Los Angeles.

Commenges, D. and H. Jacqmin-Gadda (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society, Series B*, **59**:157–171.

Commenges, D., L. Letenneur, H. Jacqmin, T. Moreau, and J.-F. Dartigues (1994). Test of homogeneity of binary data with explanatory variables. *Biometrics*, **50**:613–620.

Cortesi, P. and M. Milgroom (1998). Genetics of vegetative incompatibility in *Cryphonectria parasitica*. *Applied Environmental Microbiology*, **64**:2988–2994.

Cortesi, P., M. Milgroom, and M. Bisiach (1996). Distribution and diversity of vegetative incompatibility types in subpopulations of *Cryphonectria parasitica* in Italy. *Mycological Research*, **100**:1087–1093.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman & Hall, London.

David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *The American Statistician*, **29**:21–31.

Devore, J. and R. Peck (1993). *Statistics: The Exploration and Analysis of Data, 2nd ed.* Brooks/Cole, Pacific Grove, Calif.

Diggle, P., K.-Y. Liang, and S. L. Zeger (1994). *Longitudinal Data Analysis*. Oxford University Press, Oxford.

Driscoll, M. F. and W. R. Gundberg (1986). The history of the development of Craig's theorem. *The American Statistician*, **40**:65–71.

Driscoll, M. F. and B. Krasnicka (1995). An accessible proof of Craig's theorem in the general case. *The American Statistician*, **49**:59–62.

Durbin, J. and S. J. Koopman (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, **84**:669–684.

Finney, D. J. (1952). *Probit Analysis*. Cambridge University Press, Cambridge.

Finney, D. J. (1971). *Probit Analysis, 3rd Ed.* Cambridge University Press, Cambridge.

Firth, D. (1987). On the efficiency of quasi–likelihood estimation. *Biometrika*, **74**:233–245.

Fisher, R. A. (1935). Appendix to the calculation of the dose-mortality curve (by Bliss). *Annals of Applied Biology*, **22**:164–165.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, **51**:309–317.

Frees, E. W. and D. Ruppert (1990). Estimation following a sequentially designed experiment. *Journal of the American Statistical Association*, **85**:1123–1129.

Gelfand, A. and B. Carlin (1993). Maximum-likelihood estimation for constrained- or missing-data models. *Canadian Journal of Statistics*, **21**:303–311.

Geyer, C. J. and E. A. Thompson (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, **54**:657–699.

Giltinan, D. and M. Davidian (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized regression model. *Journal of the American Statistical Association*, **57**:369–375.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**:43–56.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Duxbury, Mass.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B*, **52**:443–452.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, **29**:637–648.

Gu, M. G. and F. H. Kong (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences*, **95**:7270–7274.

Hartley, H. O. and J. N. K. Rao (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**:93–108.

Hayes, K. and J. Haslett (1999). Simplifying general least squares. *The American Statistician*, **53**:376–381.

Hedeker, D. and R. D. Gibbons (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**:933–944.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**:226–252.

Henderson, C. R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding* (edited by W. D. Hansen and H. F. Robinson). Publication 982, National Academy of Sciences and National Research Council. Washington, D.C.

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. N. von Krosigk (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**:192–218.

Herbach, J. H. (1959). Properties of Model II type analysis of variance tests, A: optimum nature of the F-test for Model II in the balanced case. *Annals of Mathematical Statistics*, **30**:939–959.

Heyde, C. C. (1997). *Quasi-likelihood and Its Application: A General Approach to Optimal Parameter Application*. Springer-Verlag, New York.

Hocking, R. R. (1985). *The Analysis of Linear Models*. Brooks/Cole, Pacific Grove, Calif.

Jacqmin-Gadda, H. and D. Commenges (1995). Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association*, **90**:1237–1246.

Jennrich, R. I. and M. D. Schluchter (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**:805–820.

Johnson, N. L. and S. Kotz (1970). *Continuous Univariate Distributions - Vol. 2*. Wiley, New York.

Kackar, R. N. and D. A. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**:853–862.

Kenward, M. G. and J. H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**:983–997.

Khuri, A. I., T. Mathew, and B. K. Sinha (1998). *Statistical Tests for Mixed Linear Models*. Wiley, New York.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, **38**:963–974.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models. With discussion. *Journal of the Royal Statistical Society, Series B*, **58**:619–678.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Wiley, New York.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**:13–22.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**:309–326.

Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**:1007–1016.

Lindstrom, M. J. and D. M. Bates (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**:1014–1022.

Lindstrom, M. J. and D. M. Bates (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**:673–687.

Liu, Q. and D. A. Pierce (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81**:624–629.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**:226–233.

McCullagh, P. (1983). Quasi–likelihood functions. *Annals of Statistics*, **11**:59–67.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models, 2nd Ed.* Chapman & Hall, London.

McCulloch, C. E. (1994). Maximum likelihood estimation of variance components for binary data. *Journal of the American Statistical Association*, **89**:330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**:162–170.

McLachlan, G. J. and T. Krishnan (1996). *The EM Algorithm and Extensions*. Wiley, New York.

Mehta, C. and N. Patel (1992). *StatXact-Turbo User's Manual*. Cytel Software, Cambridge, MA.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics, 3rd Ed.* McGraw-Hill, New York.

Moyeed, R. A. and A. J. Baddeley (1991). Stochastic approximation of the MLE for a spatial point pattern. *Scandinavian Journal of Statistics*, **18**:39–50.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**:370–384.

Neyman, J. and E. S. Pearson (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, **20A**:175–240.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**:1–32.

Parker, H. (1995). Personal communication. Cornell University.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**:545–554.

Piepho, H. P. and C. E. McCulloch (1999). Transformations in mixed models: Application to risk analysis for a multi-environment trial. Department of Biometrics Technical Report BU-1460-M, Cornell University, Ithaca, NY.

Pinheiro, J. C. and D. M. Bates (1995). Approximations to the log-likelihood function in nonlinear mixed-effects models. *Journal of Computational and Graphical Statistics*, **4**:12–35.

Portnoy, S. (1982). Maximizing the probability of correctly ordering random variables using linear predictors. *Journal of Multivariate Analysis*, **12**:256–269.

Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**:163–171.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1996). *Numerical Recipes in Fortran 90, 2nd ed.*. Cambridge University Press, Cambridge.

Puntanen, S. and G. P. H. Styan (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, **43**:153–164.

Rao, C. R. (1962). A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society, Series B*, **24**:152–158.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Ruppert, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis*. Marcel Dekker, New York, 503–529.

Ruppert, D., N. Cressie, and R. J. Carroll (1989). A transformation/weighting model for estimating Michaelis-Menten parameters. *Biometrics*, **45**:637–656.

Santner, T. J. and D. E. Duffy (1990). *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.

Santner, T. J. and M. K. Snell (1980). Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. *Journal of the American Statistical Association*, **75**:386–394.

SAS Institute (1998). *SAS/STAT User's Guide, Version 7-1*. SAS Institute Inc., Cary, N.C.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**:110–114.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**:719–727.

Searle, S. R. (1971). *Linear Models*. Wiley, New York.

Searle, S. R. (1974). Prediction, mixed models and variance components. In *Reliability and Biometry* (edited by F. Proschan and R. Serfling). Society for Industrial and Applied Mathematics, Philadelphia.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.

Searle, S. R. (1987). *Linear Models for Unbalanced Data*. Wiley, New York.

Searle, S. R. (1997). *Linear Models*. Classic Edition, Wiley, New York. (Reprinted from 1971).

Searle, S. R. (1999). On Linear Models with Restrictions on Parameters. Department of Biometrics Technical Report BU-1450-M, Cornell University, Ithaca, NY.

Searle, S. R., G. Casella, and C. E. McCulloch (1992). *Variance Components*. Wiley, New York.

Seber, G. A. F. (1977). *Linear Regression Analysis*. Wiley, New York.

Self, S. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**:605–610.

Sheiner, L. B., S. L. Beal, and A. Dunne (1997). Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials (c/r: pp. 1245–1255). *Journal of the American Statistical Association*, **92**:1235–1244.

Snedecor, G. W. and W. G. Cochran (1989). *Statistical Methods, 8th ed.* Iowa State University Press, Ames, Iowa.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**:82–86.

Verbyla, A. P. (1993). Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society, Series B*, **55**:493–508.

Vonesh, E. F. and V. M. Chinchilli (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Dekker, New York.

Wald, A. (1941). Asymptotically most powerful tests of statistical hypotheses. *Annals of Mathematical Statistics*, **12**:1–19.

Wedderburn, R. W. M. (1974). Quasi–likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**:439–447.

Weisberg, S. (1980). *Applied Linear Regression*. Wiley, New York.

Williams, D. A. (1975). The analysis of binary response from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31**:949–952.

Wolfinger, R. W. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**:791–795.

Yu, H., S. R. Searle, and C. E. McCulloch (1994). Properties of maximum likelihood estimators of variance components in the one-way classification, balanced data. *Communications in Statistics, Series B*, **23**:897–914.

Zeger, S. L. and K.-Y. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**:121–130.

Index

This page intentionally left blank

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Probability and Statistics Section

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Probability and Statistics (Continued)

HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
IMAN and CONOVER · A Modern Approach to Statistics
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KELLY · Probability, Statistics, and Optimization
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
LINDVALL · Lectures on the Coupling Method
MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
PRESS · Bayesian Statistics: Principles, Models, and Applications
PUKELSHEIM · Optimal Experimental Design
RAO · Asymptotic Theory of Statistical Inference
RAO · Linear Statistical Inference and Its Applications, *Second Edition*
RAO and SHANBHAG · Choquet-Deny Type Functional Equations with Applications to Stochastic Models
ROBERTSON, WRIGHT, and DYKSTRA · Order Restricted Statistical Inference
ROGERS and WILLIAMS · Diffusions, Markov Processes, and Martingales, Volume I: Foundations, *Second Edition;* Volume II: Îto Calculus
RUBINSTEIN and SHAPIRO · Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method
RUZSA and SZEKELY · Algebraic Probability Theory
SCHEFFE · The Analysis of Variance
SEBER · Linear Regression Analysis
SEBER · Multivariate Observations
SEBER and WILD · Nonlinear Regression
SERFLING · Approximation Theorems of Mathematical Statistics
SHORACK and WELLNER · Empirical Processes with Applications to Statistics
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYANOV · Counterexamples in Probability
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON and SEBER · Adaptive Sampling
WELSH · Aspects of Statistical Inference
WHITTAKER · Graphical Models in Applied Multivariate Statistics
YANG · The Construction Theory of Denumerable Markov Processes

Applied Probability and Statistics Section

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · Categorical Data Analysis

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies

*ARTHANARI and DODGE · Mathematical Programming in Statistics

ASMUSSEN · Applied Probability and Queues

*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences

BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*

BARTHOLOMEW, FORBES, and McLEAN · Statistical Techniques for Manpower Planning, *Second Edition*

BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems

BATES and WATTS · Nonlinear Regression Analysis and Its Applications

BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons

BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression

BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

BHAT · Elements of Applied Stochastic Processes, *Second Edition*

BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications

BIRKES and DODGE · Alternative Methods of Regression

BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization

BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*

BOLLEN · Structural Equations with Latent Variables

BOULEAU · Numerical Methods for Stochastic Processes

BOX · Bayesian Inference in Statistical Analysis

BOX and DRAPER · Empirical Model-Building and Response Surfaces

*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement

BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation

BUNKE and BUNKE · Nonlinear Regression, Functional Relations and Robust Methods: Statistical Methods of Model Building

CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression

CHERNICK · Bootstrap Methods: A Practitioner's Guide

CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty

CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*

*COCHRAN and COX · Experimental Designs, *Second Edition*

CONOVER · Practical Nonparametric Statistics, *Second Edition*

CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Second Edition*

*COX · Planning of Experiments

CRESSIE · Statistics for Spatial Data, *Revised Edition*

DANIEL · Applications of Statistics to Industrial Experimentation

DAVID · Order Statistics, *Second Edition*

*DEGROOT, FIENBERG, and KADANE · Statistics and the Law

DODGE · Alternative Methods of Regression

DOWDY and WEARDEN · Statistics for Research, *Second Edition*

GALLANT · Nonlinear Statistical Models

GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems

GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*

GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

*HAHN · Statistical Models in Engineering

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HAND · Construction and Assessment of Classification Rules
HAND · Discrimination and Classification
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HINKELMAN and KEMPTHORNE: · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance
HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variables
HOGG and KLUGMAN · Loss Distributions
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HØYLAND and RAUSAND · System Reliability Theory: Models and Statistical Methods
HUBERTY · Applied Discriminant Analysis
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ, and KEMP · Univariate Discrete Distributions, *Second Edition*
JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations
KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KELLY · Reversability and Stochastic Networks
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data
LLOYD · The Statistical Analysis of Categorical Data
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions
McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
MEEKER and ESCOBAR · Statistical Methods for Reliability Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

MONTGOMERY and PECK · Introduction to Linear Regression Analysis, *Second Edition*
MYERS and MONTGOMERY · Response Surface Methodology: Process and Product in Optimization Using Designed Experiments
NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
NELSON · Applied Life Data Analysis
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, and SUGIHARA · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
PORT · Theoretical Probability for Applications
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
RACHEV · Probability Metrics and the Stability of Stochastic Models
RÉNYI · A Diary on Information Theory
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Statistical Methods for Quality Improvement, *Second Edition*
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE, CASELLA, and McCULLOCH · Variance Components
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
THOMPSON · Empirical Model Building
THOMPSON · Sampling
THOMPSON · Simulation: A Modeler's Approach
TIJMS · Stochastic Modeling and Analysis: A Computational Approach
TIJMS · Stochastic Models: An Algorithmic Approach
TITTERINGTON, SMITH, and MAKOV · Statistical Analysis of Finite Mixture Distributions
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume 1: Point Pattern and Quantitative Data
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN RIJCKEVORSEL and DE LEEUW · Component and Correspondence Analysis
VIDAKOVIC · Statistical Modeling by Wavelets
WEISBERG · Applied Linear Regression, *Second Edition*
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment
WHITTLE · Systems in Stochastic Equilibrium
*ZELLNER · An Introduction to Bayesian Inference in Econometrics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Biostatistics Section

ARMITAGE and DAVID (editors) · Advances in Biometry
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies
DUNN · Basic Statistics: A Primer for the Biomedical Sciences, *Second Edition*
DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, *Second Edition*
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Second Edition*
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
LAWLESS · Statistical Models and Methods for Lifetime Data
LEE · Statistical Methods for Survival Data Analysis, *Second Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
McNEIL · Epidemiological Research Methods
McFADDEN · Management of Data in Clinical Trials
*MILLER · Survival Analysis, *Second Edition*
PIANTADOSI · Clinical Trials: A Methodologic Perspective
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON · Statistical Methods for the Analysis of Biomedical Data

Financial Engineering Section

HUNT and KENNEDY · Financial Derivatives in Theory and Practice
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance

Texts, References, and Pocketbooks Section

AGRESTI · An Introduction to Categorical Data Analysis
ANDEL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Second Edition*
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and COLTON · Encyclopedia of Biostatistics: Volumes 1 to 6 with Index
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*
BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BHATTACHARYA and JOHNSON · Statistical Concepts and Methods
BILLINGSLEY · Probability and Measure, *Second Edition*
BOX · R. A. Fisher, the Life of a Scientist
BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Texts, References, and Pocketbooks (Continued)

CHATTERJEE and PRICE · Regression Analysis by Example, *Third Edition*
COOK and WEISBERG · Applied Regression Including Computing and Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
COX · A Handbook of Introductory Statistical Methods
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Sixth Edition*
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
*DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DUDEWICZ and MISHRA · Modern Mathematical Statistics
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FREEMAN and SMITH · Aspects of Uncertainty: A Tribute to D. V. Lindley
GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HELLER · MACSYMA for Statisticians
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
KHURI · Advanced Calculus with Applications in Statistics
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, REED, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, REED, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
MARDIA · The Art of Statistical Science: A Tribute to G. S. Watson
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science
McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
RENCHER · Linear Models in Statistics
RENCHER · Methods of Multivariate Analysis
RENCHER · Multivariate Statistical Inference with Applications
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROHATGI · An Introduction to Probability Theory and Mathematical Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Texts, References, and Pocketbooks (Continued)

ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
RYAN · Modern Regression Methods
SCHOTT · Matrix Analysis for Statistics
SEARLE · Matrix Algebra Useful for Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization

JWS/SAS Co-Publications Section

KHATTREE and NAIK · Applied Multivariate Statistics with SAS Software, *Second Edition*
KHATTREE and NAIK · Applied Descriptive Multivariate Statistics Using SAS Software

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors
Robert M. Groves, Graham Kalton, J. N. K. Rao, Norbert Schwarz, Christopher Skinner

Survey Methodology Section

BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
COCHRAN · Sampling Techniques, *Third Edition*
COUPER, BAKER, BETHLEHEM, CLARK, MARTIN, NICHOLLS, and O'REILLY (editors) · Computer Assisted Survey Information Collection
COX, BINDER, CHINNAPPA, CHRISTIANSON, COLLEDGE, and KOTT (editors) · Business Survey Methods
*DEMING · Sample Design in Business Research
DILLMAN · Mail and Telephone Surveys: The Total Design Method, *Second Edition*
DILLMAN · Mail and Internet Surveys: The Tailored Design Method
GROVES and COUPER · Nonresponse in Household Interview Surveys
GROVES · Survey Errors and Survey Costs
GROVES, BIEMER, LYBERG, MASSEY, NICHOLLS, and WAKSBERG · Telephone Survey Methodology
*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume 1: Methods and Applications
*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume II: Theory
KISH · Statistical Design for Research

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Survey Methodology (Continued)

*KISH · Survey Sampling

KORN and GRAUBARD · Analysis of Health Surveys

LESSLER and KALSBEEK · Nonsampling Error in Surveys

LEVY and LEMESHOW · Sampling of Populations: Methods and Applications, *Third Edition*

LYBERG, BIEMER, COLLINS, de LEEUW, DIPPO, SCHWARZ, TREWIN (editors) · Survey Measurement and Process Quality

SIRKEN, HERRMANN, SCHECHTER, SCHWARZ, TANUR, and TOURANGEAU (editors) · Cognition and Survey Research

VALLIANT, DORFMAN, and ROYALL · A Finite Population Sampling and Inference