Topics in Modeling Multilevel and Longitudinal Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting Wiskunde
te verdedigen door

Didier RENARD

Promotor : Prof. dr. Geert Molenberghs

Acknowledgements

This work could not have been accomplished without assistance of a number of persons and I would like to take these few lines to thank them warmly for their help.

First of all, I owe my deep gratitude to my supervisor, Prof. dr. Geert Molenberghs, without whom accomplishment of this work would not have been made possible. Geert is a very kind, generous and open person and it was a real pleasure working with him. I can only hope that our collaboration will keep on going in the future.

My stay at the Limburgs Universitair Centrum was also a great pleasure. Not only did I enjoy the past four years that I have spent there, but I also learned a lot through the many consultancy projects I was involved in and the fruitful collaboration with Geert and other colleagues at L.U.C. I would like, in particular, to thank all members of what we happen to call the "surrogate endpoints working group", that is: Marc Buyse, Geert Molenberghs, Tomasz Burzykowski, Helena Geys, Ariel Aronso, Jose Cortiñas Abrahantes, and Fabian Tibaldi. I would also like to thank my room-mates: Tomasz Burzykowski, Ziv Shkedy and Veerle Vandersmissen. My special thanks go to Tomasz for the many interesting discussions we have had.

I am very grateful to my family, whose presence and support have always been important to me. I dedicate this work to the woman who shares my life, Mélissa, and to our newly born daughter, Chloé.

Finally, I gratefully acknowledge the financial support from an LUC Bijzonder Onderzoeksfonds grant that allowed me to stay at L.U.C. and carry out this work.

Didier Renard

Diepenbeek
September 2002

Contents

List of Abbreviations

ARE	Asymptotic Relative Efficiency
CGI	Clinical Global Impression
EB	Empirical Bayes
EM	Expectation Maximization
GEE	Generalized Estimating Equations
GLMM	Generalized Linear Mixed Model
HIS	Health Interview Survey
IID	Independent and Identically Distributed
IGLS	Iterative Generalized Least Squares
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MPL	Maximum Pairwise Likelihood
MPLE	Maximum Pairwise Likelihood Estimator
PANSS	Positive And Negative Syndrome Scale
PL	Pairwise Likelihood
PQL	Penalized Quasi Likelihood
PQL2	Penalized Quasi Likelihood (2^{nd} order approximation)
PSA	Prostate Specific Antigen
REML	Restricted Maximum Likelihood
RIGLS	Restricted Iterative Generalized Least Squares
SD	Standard Deviation
SE	Standard Error
WPL	Weighted Pairwise Likelihood

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Clustered Data

Often in applied statistics, after some empirical data have been collected, the purpose of the analysis is to construct a statistical model. Otherwise said, we are interested in situations where the aim is to explain how an *outcome*, or *response*, variable of particular interest is related to a set of *explanatory variables*, or *covariates*.

Classically, a single observation on the response variable is obtained for each observational unit and one of the fundamental hypotheses of standard statistical modeling in this case is independence between observations. Many types of studies, however, have designs which imply gathering data in dependent groups or clusters. Familiar examples of clusters are animal litters, families or schools. In each of these examples, a cluster is a collection of subunits on which observations are made. Another usual form of clustering arises when data are measured repeatedly on the same unit.

In all cases the elements of a cluster, whether they are meaningful subunits or repeated measures on the same unit, share some common characteristics. Therefore, the distinguishing feature of *clustered* data is that observations within a cluster tend to be more alike than observations from different clusters or, stated otherwise, they are *correlated*. Thus, unlike in the classical setting where there exists a single source of variation between observational units, the heterogeneity between clusters introduces an additional source of variation and complicates the analysis.

When this variation cannot be explained by measured covariates, we require statistical analysis methods which acknowledge explicitly for the clustering in the data.

Failure to account for the effect of clustering can result in erroneous estimation of the variability of parameter estimates, and hence in misleading inference. Although this fact has long been recognized, it was not until quite recently that the wide availability and advances in computer power have permitted the development of appropriate statistical techniques for the analysis of correlated measurements.

Correlated data arise naturally in many different ways in scientific disciplines such as the biological, health or social sciences, and this generic term actually covers a multitude of data structures. We now briefly describe those which are mostly encountered in the statistical literature, that is, multivariate data, clustered data, longitudinal data, spatial data and multilevel data.

1.2 Types of Correlated Data

Researchers frequently collect measurements on several variables in practice. Multivariate statistical analysis (Johnson and Wichern, 1992) is concerned with statistical methods for describing and analyzing a group of variables simultaneously. As an example, suppose that a clinical trial was designed to compare the effect of a new drug to placebo for the treatment of chronic schizophrenia and that several measuring instruments (or scales) are considered to assess a patient's condition. A multivariate model could be assumed for studying the (joint) relationship of the scores on each instrument with treatment and possibly other predictor variables of interest. One should of course acknowledge that the response variables are correlated, each being measured on the same sample of patients.

Suppose now that one of the scales comprises a subjective clinical evaluation by the treating physician. Patients rated by the same physician will tend to have scores on that instrument that are more alike, maybe because the physician rates, on average, higher or lower than other physicians. Thus, we have yet another illustration of clustered data.

If the outcome variable were to be measured under different experimental conditions, we would face a so-called repeated-measures study design. When the outcome is measured repeatedly over time, we are dealing with longitudinal data (Diggle, Liang and Zeger, 1994). For example, the test could be administered a number of times (at baseline and then monthly for a year, say) in order to investigate the individual patterns of change over time.

We can usually distinguish between three sources of variation with longitudinally

measured data: inter-individual variability, serial correlation and measurement error. The second component arises owing to the fact that pairs of measurements taken closer in time often show a stronger similarity than pairs of measurements taken further apart. Thus, time — a one-dimensional scale — plays a critical role in defining a component of serial correlation. If instead measures were obtained at different spatial locations (that is, using two or even three dimensions), we would talk about spatial data (Cressie, 1991) and an autocorrelation structure could be assumed provided some distance measure is introduced.

So far, we have focused on data structures involving a single level of clustering. It is frequent, however, to face problems that have a hierarchical or *multilevel* structure (Goldstein, 1995), that is, when data have been collected in different layers, or levels, of a hierarchy. To expand on our initial example, suppose that physicians were affiliated to some hospital and that a number of hospitals were contacted for recruiting patients. This defines a three-level structure wherein hospitals are level 3 units, physicians are level 2 units and, assuming we focus on a single scale of measurement, patients would correspond to level 1 units. Should we consider all available measurements (from different measuring instruments) on each patient, an additional level could then be introduced, resulting in a four-level data structure.

1.3 Statistical Models for Correlated Data

An important consideration in the statistical modeling of correlated data concerns the type of outcome. Methods for continuous (read 'normally distributed') data are undeniably the best developed and the linear mixed model (Laird and Ware, 1982; Verbeke and Molenberghs, 1997, 2000) has played a prominent role in extending the general linear model to handle correlated continuous data. Owing to the elegant properties of the multivariate normal distribution, its theory and implementation are greatly simplified. Software programs, such as the SAS procedure MIXED (Littell *et al.*, 1996), are therefore widely available to fit this kind of models and have facilitated dissemination of the methodology among the statistical community.

When the outcome variable is discrete (e.g. counts) or categorical (nominal or ordinal data), a first issue arises which is the lack of a discrete analogue to the multivariate normal distribution. Complete specification of the joint distribution of the response vector becomes more problematic and fully likelihood-based methods are generally awkward.

Another issue raised by this type of outcomes is that the researcher must distinguish between three broad model families. For simplicity, let us concentrate on the special case of correlated binary outcomes. A *marginal* model is one in which marginal probalities of response are directly modeled. There is an extensive statistical literature on marginal modeling of correlated binary responses. For example, Bahadur (1961) and Zhao and Prentice (1990), describe maximum likelihood estimation where marginal correlations are used to account for the association among responses. Alternatively, the within-cluster association can be parameterized in terms of marginal odds ratios, as shown by Dale (1986), Liang, Zeger and Qaqish (1992), Lang and Agresti (1994), Molenberghs and Lesaffre (1994), and Glonek and McCullagh (1995) for instance.

Since few joint probability models for multivariate categorical data permit tractable modeling of marginal probabilities, alternative methods have been in demand. Thus, Liang and Zeger (1986) and Zeger and Liang (1986) proposed so-called generalized estimating equations (GEE) which do not require assumptions about the complete joint distribution of the response vector. Their approach relies on estimating functions and provides a natural extension of quasilikelihood (Wedderburn, 1974) to the multivariate response setting. Standard GEE require only correct specification of the univariate marginal probabilities while adopting some working assumptions about the association structure. Extensions of GEE that allow modeling of pairwise associations were given, for example, by Prentice (1988), Lipsitz, Laird and Harrington (1991), Liang, Zeger and Qaqish (1992) and Carey, Zeger and Diggle (1993), using the correlation or the odds ratio as a measure of association.

Drawing on direct analogies with linear models for continuous responses, another way to model the joint distribution of the response vector is to postulate the existence of unobserved latent variables, often called random effects. These can be thought of as reprensenting various features shared by the subunits of a cluster and hence introduce correlation among observations. Such cluster-specific effects are usually assumed to be independent and identically distributed according to a certain mixing distribution. An additional assumption that is frequently used is that the observations within a cluster are conditionally independent given the random effects. When the mixing distribution is assumed Gaussian, the families of linear mixed models and generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) can be combined to form the class of generalized linear mixed models (GLMM). These models have been studied, among others, by Stiratelli, Laird and Ware (1984), Anderson and Aitkin

(1985), Gilmour, Anderson and Rae (1985), Zeger, Liang and Albert (1988), Breslow and Clayton (1993), Wolfinger and O'Connell (1993) and Goldstein and Rasbach (1996). Some authors have advocated different distributional assumptions for the mixing distribution (e.g. Lee and Nelder, 1996). Also, cluster-specific approaches are not limited to mixed models, as demonstrated by the popular beta-binomial model (Williams, 1975).

There are some important distinctions between the two model families described so far. In marginal models parameters may be interpreted with respect to the marginal or population-averaged distribution; therefore, such models are often referred to as *population-averaged* models. In random-effects models, on the other hand, parameters have cluster-specific effects and these models are consequently also called *cluster-specific*. This distinction is, in effect, irrelevant for normal outcomes since parameters have both population-averaged and cluster-specific interpretations in this case, but it becomes critical with categorical data. Zeger, Liang and Albert (1988), for example, discuss these two approaches to modeling of longitudinal data using GEE to estimate model parameters.

The third class of models that is commonly used to model correlated data is that of (response) *conditional* models. In a conditional model the parameters describe a feature (e.g. response probability) of a set of outcomes conditionally on the other outcomes. Due to the popularity of marginal and random-effects models, these have received relatively little attention, especially in the case of clustered data. An example in the specific context of clustered binary outcomes was given by Molenberghs and Ryan (1999). Conditional models have been much more popular in the context of longitudinal data, where they have been termed *transition* models (Diggle, Liang and Zeger, 1994). The (response) conditional approach, however, is usually criticized because of the conditional interpretation of the parameters on other outcomes and on cluster sizes.

The debate continues about the relative merits of the different approaches. For several years, it seemed that marginal models, particularly GEE, were the most popular, perhaps due to their relative computational ease and the availibility of good software (e.g. SAS procedure GENMOD). More recently, there has been a renewed interest in random-effects models partly provoked by the availability of the SAS procedure NLMIXED. There are merits and disadvantages to all three model families, however. Arguably, model choice will depend not only on the application of interest but also on the specific analysis goals.

Finally, in an attempt to overcome the limitations of a single class of models, some researchers have also proposed to combine two approaches, thus getting the best of two worlds. A first possibility is to modify the natural parameterization of the response conditional models to allow likelihood-based estimation of marginal mean parameters. Fitzmaurice and Laird (1993), for instance, discuss marginalized log-linear models, while Azzalini (1994) presents a marginalized transition model. More recently, Heagerty (1999) and Heagerty and Zeger (2000) have proposed a class of marginalized multilevel models, wherein the mean structure is modeled marginally and the association among responses is accounted for by random effects. This approach has the advantage of allowing a marginal interpretation of the regression parameters, while enabling parsimonious parameterization of the covariance structure using random effects.

1.4 Organization of Subsequent Chapters

Throughout the present work, we will mainly be concerned with applications of the random-effects modeling approach. Emphasis will be on multilevel modeling (Chapters 2-4) on the one hand, and modeling of longitudinal data (Chapters 5-7) on the other hand.

In Chapter 2, we introduce multilevel models for continuous (Gaussian) and binary outcomes and briefly review estimation methods to fit this type of models. We then employ multilevel modeling techniques to examine clustering in the 1997 Belgian Health Interview Survey (HIS1997). The sampling scheme in this survey was a combination of several sampling techniques and, in particular, of multistage sampling. It is therefore natural to use multilevel models as an attempt to account for the sampling design when analyzing such data. The basic hierarchy defined by multistage sampling in the HIS1997 is constituted of three levels: individuals, households and municipalities. Using this hierarchical structure, we present an illustration of both a multilevel linear model and a multilevel logistic model.

In Chapter 3, we will focus on multilevel models for binary responses. More specifically, the aim of this chapter is to describe and investigate in more details an alternative estimation method to standard maximum likelihood (ML) and penelized quasi likelihood (PQL) procedures. This estimation method, termed maximum pairwise likelihood (MPL), is based on a particular form of pseudo-likelihood (Besag, 1975). After examining asymptotic properties of pseudo-likelihood estimators, we look more

specifically into the pairwise likelihood method within the multilevel probit model. We present some asymptotic relative efficiency calculations and a series of simulations to compare MPL to ML and PQL estimators.

The goal of the next chapter is to provide an illustration of the MPL estimation method in the context of surrogate endpoint validation. The framework is that developed by Buyse *et al.* (2000), in which the surrogate endpoint validation issue is approached from a meta-analytic standpoint and examined at each of two levels: the individual and the trial levels. We propose an extension of their methodology, applicable to normally distributed endpoints, to the case of two binary outcomes. This is done using a latent variable approach yielding a three-level probit model with a four-dimensional random-effects structure, which makes of pairwise likelihood a perfectly suitable estimation method in this situation.

In the second part of the thesis, we will mostly deal with longitudinal data modeling. In Chapter 5 we use linear mixed model methodology to analyze data from a hepatitis B vaccination program. This enables us to discuss some problems commonly faced in the modeling of continuous (normally distributed) longitudinal measurements. In particular, we examine the use of fractional polynomials as a flexible tool for parametric modeling in this context.

In Chapter 6 we focus on the issue of estimating reliability in clinical trial data with longitudinal binary outcomes. We first discuss how this can be accomplished in the framework of generalized linear mixed models. One limitation with such models is that they address the issue of clustering solely by introducing random effects in the model but do not allow for residual sources of autocorrelation. To overcome this problem, we propose a latent variable probit model in which the (latent) residual error terms are assumed to be realizations of a Gaussian process. This affords us an alternative way to parameterize association among the longitudinal outcomes. Although likelihood estimation of this model is awkward, pairwise likelihood estimation is relatively easy to implement, the procedure being a straightforward extension of that described in Chapter 3.

In Chapter 7 we return to the surrogate endpoint validation issue, when it is of interest to use a longitudinally measured biomarker as a surrogate for a time-to-event endpoint. To extend the approach of Buyse *et al.* (2000), we need to formulate a joint model for longitudinal measurements and event time data. To this end, the model of Henderson, Diggle and Dobson (2000) is adopted. We examine how trial- and individual-level surrogacy measures can be adapted in this context and use a set of

two randomized clinical trials in advanced prostate cancer to evaluate the usefulness of prostate-specific antigen (PSA) level as a surrogate for survival.

Finally, in Chapter 8 we formulate concluding remarks and also indicate some topics for further research.

Part I :

Multilevel Modeling

Chapter 2

Introduction to Multilevel Models

2.1 Introduction

Many sets of data collected in human and biological sciences have a *multilevel* or *hierarchical* structure. By hierarchy we mean that units at a certain level (also called micro units) are grouped into, or nested within, higher level (or macro) units. Schooling systems, for instance, present an obvious hierarchical structure, with pupils nested within classrooms, which are themselves nested within schools, and so forth.

Multilevel structures are common in practice and in fact, it could be argued that they are the norm rather than the exception. Over the last twenty years, there has been increasing interest in developing suitable techniques for the statistical modeling and analysis of hierarchically structured data, and this has resulted in a broad class of models known under the generic name of *multilevel models* (Goldstein, 1995).

An area where multilevel structures frequently occur is in survey sampling, where for cost-related reasons or administrative considerations, *multistage sampling* schemes are often adopted. In multistage sampling, the sample is selected in different stages, with the sampling units at each stage being sub-sampled from the larger units drawn at the previous stage.

Clustering in sampling surveys has been traditionally handled using design-based procedures, which are *ad hoc* corrections to account for the sampling design (Skinner

et al., 1989). In this approach the population structure, insofar as it is mirrored in the sampling design, is seen as a nuisance factor. On the contrary, in the multilevel modeling approach the population structure is considered of potential interest in its own right and is an integral part of the model. In addition, a model-based approach enables one to incorporate design-related information directly into the model, thus obviating the need to carry out special procedures to adjust for the effects of the sampling design.

As an illustrative example throughout the chapter, we will consider data from a Heath Interview Survey which took place for the first time in 1997 in Belgium. The sampling scheme of the HIS1997 was a combination of several sampling techniques and, in particular, of multistage sampling. The different levels of the hierarchy defined by this multistage sampling scheme are individuals, households and municipalities.

After a general description of the survey and its objectives in Section 2.2, we review linear and nonlinear multilevel models in Sections 2.3 and 2.4 respectively. Some attention will be given to the issue of unequal selection probability weighting in Section 2.5 as this is not, in principle, a simple extension of conventional weighing methods. In Section 2.6, multilevel models are used to examine the extent of clustering in the HIS1997 data based on specific (continuous and binary) outcomes. Note that the content of this chapter is mainly based on the papers of Renard *et al.* (1998) and Renard and Molenberghs (2002).

2.2 The 1997 Belgian Health Interview Survey

Health interview surveys aim to provide a global description of the health status of the population. Individuals are questioned about a wide variety of health related domains such as general health perception, morbidity, use of health services, lifestyle or socio-economic characteristics. This information, in turn, is useful to provide rational bases for the health policy makers and aids in the identification of health priorities, the description of health needs of the population, the estimation of the prevalence and distribution of health determinants, the analysis of social (in)equities in health and health services, and the study of health care consumption and its determinants.

The first Belgian Health Interview Survey was undertaken by the Scientific Institute for Public Health in 1997. A detailed account of the HIS1997 (simply referred to as HIS in the sequel) sampling design is given by Quataert *et al.* (1997). See Van Oyen *et al.* (1997) for a more concise description of the survey. Here, we briefly out-

line the main aspects of the final sampling scheme for the selection of the households and respondents in the HIS.

The target number of interviews in the HIS was 10,000. The sampling of the households and respondents was a combination of different sampling techniques such as stratification and multistage sampling. Stratification was performed at the regional level (Flemish, Walloon and Brussels regions) and at the provincial level. At the regional level, unequal sampling rates were taken to guarantee sufficient precision of the results. Within a region, sampling was taken proportional to population size in each province. An extra refinement was needed for the German community which was considered a proper entity on its own and was oversampled. Regional and provincial stratification aim at achieving a geographical spread of the interviews. The quota of interviews were also evenly distributed over quarters of the study year to obtain reasonable spread over time.

Within each stratum, a sample of individuals was obtained in three stages. At the first stage, municipalities (primary sampling units) were drawn by a systematic sampling procedure with probability proportional to their size. Each time a municipality was selected, a group of 50 individuals had to be successfully contacted. The next stage of random selection operated at household level (secondary sampling units) according to a clustered systematic sampling procedure upon ordering of the households by statistical sector, size and age of the reference person. Additional replacement households with matching characteristics were also identified in case some household would refuse to participate or could not be contacted. Finally, individuals (tertiary sampling units) were selected within households in such a way that at most four persons were interviewed in each household and the reference person and his or her partner were automatically selected.

2.3 Linear Multilevel Models

In this section, multilevel models for normally distributed response variables are briefly discussed. Keeping an eye towards the HIS, we consider a three-level population for notational convenience. Thus, we assimilate level 3 units to municipalities, level 2 units to households, and level 1 units to individuals.

For comprehensive accounts on multilevel modeling, we refer to the books by Bryk and Raudenbush (1992), Longford (1993), Goldstein (1995) and Snijders and Bosker (1999). Kreft and de Leeuw (1998) provide a more informal and introductory

approach to the subject.

Suppose that we have a sample consisting of K municipalities, with J_k households within the kth municipality ($k = 1, \ldots, K$) and N_{jk} individuals within the jth household from the kth municipality ($j = 1, \ldots, J_k$; $k = 1, \ldots, K$). In conformity with the standard index notation in multilevel models, we shall let y_{ijk} denote the value of the response variable recorded on the ith individual within the jth household from the kth municipality. The standard linear three-level model has the following structure:

$$y_{ijk} = \boldsymbol{x}_{ijk}^T \boldsymbol{\beta} + \boldsymbol{z}_{3,ijk}^T \boldsymbol{u}_k^{(3)} + \boldsymbol{z}_{2,ijk}^T \boldsymbol{u}_{jk}^{(2)} + \epsilon_{ijk}, \tag{2.1}$$

where $\boldsymbol{x}_{ijk}$ is a vector of covariates having fixed effects $\boldsymbol{\beta}$, $\boldsymbol{z}_{3,ijk}$ is a vector of covariates having random effects $\boldsymbol{u}_k^{(3)}$ at the municipality level, $\boldsymbol{z}_{2,ijk}$ is a vector of covariates having random effects $\boldsymbol{u}_{jk}^{(2)}$ at the household level and ϵ_{ijk} is an error term. All random terms in the model are assumed to be mutually independent and normally distributed:

$$\begin{aligned} \boldsymbol{u}_k^{(3)} &\sim N(\boldsymbol{0}, \Omega_u^{(3)}), \\ \boldsymbol{u}_{jk}^{(2)} &\sim N(\boldsymbol{0}, \Omega_u^{(2)}), \\ \epsilon_{ijk} &\sim N(0, \sigma_e^2). \end{aligned}$$

With a slight abuse of language, we will sometimes call $\boldsymbol{\beta}$ the fixed parameters, while random parameters will denote variance and covariance parameters characterizing the random terms in the model.

The vectors of covariates $\boldsymbol{z}_{3,ijk}$ and $\boldsymbol{z}_{2,ijk}$ will usually be subsets of the fixed-effects covariates $\boldsymbol{x}_{ijk}$, although they do not need to. The associated random effects are used to account for variation in the data that is attributable to clustering at the corresponding levels of the hierarchy. Specifically, $\boldsymbol{u}_{jk}^{(2)}$ represents the effect of the jth household in the kth municipality on the covariates $\boldsymbol{z}_{2,ijk}$ and is characteristic of between-household variability. A similar interpretation holds for $\boldsymbol{u}_k^{(3)}$ at the municipality level. Thus, we can see that multilevel models provide a natural way to decompose complex patterns of variability associated with hierarchical structures.

In (2.1) the standard assumption of homoscedasticity was made, i.e. the residual variance is assumed constant. This assumption can nevertheless be relaxed and dependence on specific covariates can be introduced by replacing ϵ_{ijk} by a term of the form $\boldsymbol{z}_{1,ijk}\epsilon_{ijk}$. This permits to represent more complex variation at level one, including subgroup variability and heteroscedasticity.

Model (2.1) can be rewritten as a special case of the general linear mixed model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{y}$ is the vector of responses and X the matrix for the fixed effects obtained by stacking the responses y_{ijk} and the covariates $\boldsymbol{x}_{ijk}$ respectively; $\boldsymbol{u}$ is the vector of random effects obtained by stacking the household effects $\boldsymbol{u}_{jk}^{(2)}$ on top of the municipality effects $\boldsymbol{u}_{k}^{(3)}$; Z is the matrix for the random effects obtained by padding with 0s and stacking the covariates $\boldsymbol{z}_{2,ijk}$ and $\boldsymbol{z}_{3,ijk}$ to conform to the structure of $\boldsymbol{u}$; and $\boldsymbol{\epsilon}$ is the vector of error terms obtained by stacking the ϵ_{ijk}'s. In other words, the theory of linear multilevel models is embodied in that of linear mixed models. The latter, however, does not explicitly recognize, nor does it take full advantage of, specific features of hierarchically structured data, a reason why multilevel modeling has been an area of research on its own.

Parameter estimation in the linear multilevel model can be carried out by maximizing the likelihood function. To this end, direct maximization using the Newton-Raphson or Expectation-Maximization (EM) algorithm can be performed. An equivalent procedure, called Iterative Generalized Least Squares (IGLS), was proposed by Goldstein (1986). This algorithm iterates between the estimation of the fixed and the random parameters using standard generalized least squares principles, hence the name. IGLS is an attractive procedure as it is computationally efficient compared to direct maximization of the likelihood, especially with large sets of data such as those typically found in multilevel modeling applications. The IGLS algorithm can also be modified to obtain residual or restricted maximum likelihood (REML) estimates, which are unbiased for random parameters. It is then refereed to as RIGLS.

Nowadays, most general purpose software packages, such as SAS, S-Plus, Stata or SPSS, have built-in capabilities to fit linear mixed (hence multilevel) models. Specialized package for multilevel modeling are also available. MLwiN (Goldstein *et al.*, 1998), a program that was developed by researchers working on the Multilevel Models Project at the Institute of Education in London, is certainly the most extensive multilevel package. It allows fitting of linear multilevel models (using the IGLS or RIGLS algorithm) and can handle discrete (binomial and count) response data as well. Furthermore, it offers parametric and nonparameteric bootstrap estimation and Markov Chain Monte Carlo (MCMC) methods to fit Bayesian models, as well as various macros to deal with more complicated types of outcomes (survival data, multi-categorical responses, time series, etc.).

2.4 Nonlinear Multilevel Models

We restrict attention to the case of a binary response, but the discussion below applies more generally, to models for binomial or count data for example. In fact, the theory can be developed for any nonlinear multilevel model (Goldstein, 1995). Besides general references to multilevel modeling, we refer to Agresti *et al.* (2000) and McCulloch and Searle (2000) for an overview of the closely related generalized linear mixed model.

Keeping the same notation as in the previous section, we consider the multilevel model

$$g(\pi_{ijk}) = \boldsymbol{x}_{ijk}^{T}\boldsymbol{\beta} + \boldsymbol{z}_{3,ijk}^{T}\boldsymbol{u}_{k}^{(3)} + \boldsymbol{z}_{2,ijk}^{T}\boldsymbol{u}_{jk}^{(2)}, \tag{2.3}$$

where $\pi_{ijk} = P[y_{ijk} = 1|\boldsymbol{u}_{k}^{(3)}, \boldsymbol{u}_{jk}^{(2)}]$ and $g(.)$ is a link function such as the logit, probit or log-log functions. Analogously to (2.2), the model can be rewritten as

$$g(\boldsymbol{\pi}) = X\boldsymbol{\beta} + Z\boldsymbol{u}, \tag{2.4}$$

where $\boldsymbol{\pi}$ is the vector of response probabilities π_{ijk}. As in the linear multilevel model, all components of the vector $\boldsymbol{u}$ are assumed to be mutually independent and normally distributed. It is further assumed that conditionally on $\boldsymbol{u}$, the binary responses y_{ijk} are independent. This assumption, known as the local independence assumption, greatly simplifies likelihood inference, as shown below.

A particularity of the above model, in contrast to the linear multilevel model, is that level one variability is not directly comparable to variability at higher levels. Indeed, since g is in general different from the identity link function, random disturbances at level 2 and above appear on a transformed scale (e.g. logit), whereas the level one variance characterizes binomial variation.

We now discuss in more details how inference can proceed in nonlinear multilevel models. We first describe maximum likelihood (ML) inference and then turn to approximate methods.

2.4.1 Maximum Marginal Likelihood

For notational simplicity, we consider a two-level model in the remainder of this section:

$$g(\pi_{ij}) = \boldsymbol{x}_{ij}^{T}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{T}\boldsymbol{u}_{j} \tag{2.5}$$

with $\pi_{ij} = P[y_{ij} = 1|\boldsymbol{u}_j]$ $(j = 1, \ldots, n;\ i = 1, \ldots, n_j)$ and $\boldsymbol{u}_j \sim N(0, \Omega_u)$.

Because of the local independence assumption, the conditional likelihood of (level 2) unit j takes the binomial form; that is, its contribution to the log *marginal* likelihood, obtained by integrating over the random effects, can be written as

$$\ell_j(\boldsymbol{\beta}, \Omega_u) = \log \int \prod_{i=1}^{n_j} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \phi(\boldsymbol{u}_j; \Omega_u) d\boldsymbol{u}_j, \tag{2.6}$$

where $\phi(\boldsymbol{u}; \Omega_u)$ is the normal density function $N(0, \Omega_u)$. The log marginal (or integrated) likelihood

$$\ell(\boldsymbol{\beta}, \Omega_u) = \sum_{j=1}^{n} \ell_j(\boldsymbol{\beta}, \Omega_u) \tag{2.7}$$

can then be maximized, using any standard optimization routine, to obtain estimates of the parameters $\boldsymbol{\beta}$ and Ω_u.

Unfortunately, expression (2.6) is intractable and necessitates the use of numerical integration techniques. When the dimension of integration is small, Gauss-Hermite quadrature can be used (Anderson and Aitkin, 1985; Crouch and Spiegelman, 1990). This method, however, suffers from the curse of dimensionality. In addition, no simple and effective error bounds are available for multidimensional integrals approximated by quadrature rules (Monaham, 2001). For these reasons ML estimation in GLMMs has been somewhat hampered in the past, being limited to simplistic models with low-dimensional random effects, and there has been a need for alternative, performant integration methods to overcome computational limitations. Among those, Monte Carlo-based methods have received a great deal of attention due to parallel developments in Bayesian statistics. For example, Monte Carlo Expectation Maximization (EM) and Monte Carlo Newton Raphson algorithms (McCulloch, 1994, 1997) were proposed to obtain ML estimates. Alternatively, the generalized linear random effects model can be cast in a fully Bayesian framework (see e.g. Zeger and Karim, 1991). A drawback of these methods is that they are computationally intensive in general.

As indicated in Chapter 1, there has been renewed interest in fitting GLMMs with quadrature-based methods over recent years owing to the wider availability of good software. The computer program MIXOR (Hedeker and Gibbons, 1996), for example, can be used for random-effects modeling of binary and ordinal responses (Hedeker and Gibbons, 1994). The software package STATA (StataCorp., 2001) offers several functions to fit GLMMs using Gaussian quadrature as well. Recently, the

procedure NLMIXED has been implemented in SAS (SAS Institute Inc., Carey, USA) to allow fitting of nonlinear random-effects models. A noticeable feature of PROC NLMIXED is that an adaptive version of Gaussian quadrature is available. With adaptive Gaussian quadrature, the quadrature nodes are centered around the mode of the function to be integrated and rescaled using the curvature of the function at its mode. The nodes therefore lie in the region of bigger "mass" and the resulting integral approximation is generally more accurate than that obtained with standard Gaussian quadrature based on the same number of nodes (Pinheiro and Bates, 1995; Lesaffre and Spiessens, 2001). Finally, Bayesian estimation using Markov chain Monte Carlo methods is possible with the BUGS software package (Spiegelhalter *et al.*, 1995).

2.4.2 Approximate Methods

Much of the early literature on random-effects models for discrete data has been concerned with ways of circumventing the computational burden caused by the need for numerical integration, and several authors have suggested to use approximations of the likelihood. Breslow and Clayton (1993), for instance, exploit the penalized quasi-likelihood (PQL) method by applying Laplace's integral approximation. They also consider marginal quasi-likelihood (MQL), a name they give to a procedure previously proposed by Goldstein (1991). PQL and MQL can be viewed as iterative procedures that entail fitting of linear multilevel models based on a first-order Taylor expansion of the mean function about the current estimated fixed part predictor (MQL) or the current predicted value (PQL). There are also a number of closely related approaches. This includes the work of Shall (1991), Wolfinger and O'Connell (1993), Longford (1994) and McGilchrist (1994).

Since these approximate procedures do not involve integrated likelihoods, they are simpler to program and computationally more efficient than ML methods. They are not without defect, however. On the basis of a large number of simulations, Rodríguez and Goldman (1995) demonstrate that these approximate procedures may be seriously biased. Their simulations reveal that both fixed effects and variance components may suffer from substantial, if not severe, attenuation bias under certain circumstances. Of note is that the PQL approximation seems to deteriorate as the data depart from normal (especially binary data) and the random effects become large.

To reduce the extent of these biases, some authors have advised the introduction of bias-correction terms (Breslow and Lin, 1995; Lin and Breslow, 1996) or the use of

iterative boostrap (Kuk, 1995). Goldstein and Rasbash (1996) show that the inclusion of a second-order term in the PQL expansion considerably reduces the biases described by Rodriguez and Goldman (1995). This method will be referred to as PQL2 in the sequel.

As to software availability, the PQL algorithm has been implemented in the %GLIMMIX macro in SAS and can be used, as well as PQL2, in MLwiN. A bootstrap method can also be used in MLwiN to correct for bias associated with PQL and Bayesian inference can be made through MCMC methods.

2.5 Weighting in Multilevel Models

Typically, analyses of data arising from complex sample surveys are adjusted by methods that incorporate sampling weights defined as the reciprocals of the sample inclusion probabilities. These sampling weights effectively represent the number of individuals in the population that each sampled individual represents (Graubard and Korn, 1996).

The issue of weighting in multilevel models has not been extensively investigated until quite recently (Pfeffermann *et al.*, 1998). A reason might be that sampling schemes are commonly ignored in multilevel analyses of survey data since multilevel models enable the data analyst to incorporate certain characteristics of the sampling design as covariates (e.g. stratification variables), even though this argument breaks down when the relevant information is not made available to the analyst or when sampling-related variables are not scientifically meaningful to include in the model. When the sample selection probabilities are related to the response variable even after conditioning on covariates of interest, the conventional estimators of the model parameters may be biased, hence the need to study weighting procedures that attempt to correct for this problem.

It should be emphasized that weighting in multilevel models is not a trivial extension of conventional methods of weighting. A key feature is that sample inclusion probabilities can be defined at any stage of the hierarchy, conditionally on cluster membership at above levels. Thus, municipality k is selected with inclusion probability π_k, household j is selected with probability $\pi_{j|k}$ within municipality k, and individual i is sampled with probability $\pi_{i|jk}$ within household j from municipality k. Unconditional selection probabilities can be derived from suitable products of conditional probabilities (e.g., $\pi_{jk} = \pi_k \pi_{j|k}$ denotes the probability that municipality k is

sampled and that, within this municipality, household j is selected).

The approach proposed by Pfeffermann *et al.* (1998) is to substitute, in the IGLS sample estimators, each sum over units at a given level by a correspondingly weighted sum, using (inverse) conditional selection probabilities as defined above. When the sample inclusion probabilities (and hence the weights) are independent of the random effects, they show that a simple transformation of the variables specified in the random part of the model is sufficient. The appropriate transformation is to:

- replace $z_{ijk}^{(1)}$ by $w_k^{-1/2} w_{j|k}^{-1/2} w_{i|jk}^{-1/2} z_{ijk}^{(1)} = w_{ijk}^{-1/2} z_{ijk}^{(1)}$,
- replace $z_{ijk}^{(2)}$ by $w_k^{-1/2} w_{j|k}^{-1/2} z_{ijk}^{(2)} = w_{ij}^{-1/2} z_{ijk}^{(2)}$,
- replace $z_{ijk}^{(3)}$ by $w_k^{-1/2} z_{ijk}^{(3)}$,

where the weights are defined by

$$w_k = \pi_k^{-1}, \;\; w_{j|k} = \pi_{j|k}^{-1}, \;\; w_{i|jk} = \pi_{i|jk}^{-1}.$$

Note that the weights should be rescaled in such a way to have unit mean. The main advantage of this procedure is that it can easily be implemented within any standard software package which allows fitting of multilevel models.

When the weights are not independent of the random effects at a certain level (the sampling mechanism is then said to be informative), this results in a more complicated procedure. The authors point out, however, that the weighted scaling method should produce acceptable results in many cases, although it can clearly give biased estimates in certain circumstances.

It should be noted that standard errors of parameter estimates cannot be obtained straightforwardly with standard software but as suggested in MLwiN, an alternative solution is to use the sandwich estimator (Liang and Zeger, 1986), which allows more robust inference.

2.6 Application to the HIS

For the sake of illustration, we consider two response variables: body mass index (BMI), which will be log-transformed and analyzed as a normally-distributed outcome, and a binary indicator for subjective or perceived health, which was originally rated by the interviewees on a 5-point scale and was dichotomized as good/very good

versus other. In an attempt to find a satisfactory model for these data, the following covariates were examined: sex, age (eight categories), education (five categories), household income (5 categories) and smoking behavior. Note that the question about smoking behavior was addressed only to persons aged 15 or more, thus reducing the effective sample size from 10,221 to 8560.
In addition to the aforementioned covariates, information about the sample design can be taken into consideration:

- stratification variables: quarter and provinces;
- size variables: province, municipality, household;
- other variables: number of groups to be interviewed within a municipality, interviewee status (indicating whether he/she is the reference person or his/her partner).

2.6.1 Linear Multilevel Model

Due to unit and item non-response, 7422 out of 8560 (87%) observations were available with complete information[†] on the selected covariates and BMI. The model we will fit is

$$y_{ijk} = \boldsymbol{x}_{ijk}^T \boldsymbol{\beta} + v_k + u_{jk} + e_{ijk}, \tag{2.8}$$

with $v_k \sim N(0, \sigma_v^2)$, $u_{jk} \sim N(0, \sigma_u^2)$ and $e_{ijk} \sim N(0, \sigma_e^2)$, and where $\boldsymbol{x}_{ijk}$ contains covariates listed above. This is an example of random-intercept (or variance components) model. The total variation in log BMI can be decomposed into that between individuals within households (σ_e^2), that between households within municipalities (σ_u^2) and that between municipalities (σ_v^2). Among covariates in $\boldsymbol{x}_{ijk}$, only sex, age, education and smoking behavior were found to have a significant effect and were included in the model. Second-order interaction terms of sex with age and education were also included. Among sampling-related variables, only province, household size and interviewee status were retained.

The variance partitioning in (2.8) allows interpretions in terms of 'intra-unit correlation'. Thus, the intra-municipality correlation coefficient can be defined as

$$\rho_{MUN} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_u^2 + \sigma_e^2}, \tag{2.9}$$

[†]See Burzykowski *et al.* (1999) for a study of missing data in the HIS.

while the intra-household correlation coefficient is equal to

$$\rho_{HH} = \frac{\sigma_v^2 + \sigma_u^2}{\sigma_v^2 + \sigma_u^2 + \sigma_e^2}. \tag{2.10}$$

These coefficients reflect the proportion of the total variability in the outcome variable that is attributable to the clustering effect at a certain level and, therefore, are measures of within-group homogeneity.

Table 2.1 shows the result of fitting model (2.8) to the data, using weighted and unweighted estimators. Robust standard errors are reported for parameter estimates. We see that there is generally good agreement (within standard error) between weighted and unweighted estimators, but that standard errors of the weighted estimators are subject to a sometimes severe loss of efficiency. Whether this is due to the use of the sandwich estimator or to weighting itself is not entirely clear. As Korn and Graubard (1995) show, weighted estimates tend to be more variable than unweighted ones, and variability in the weighted estimates increases as sampling weights become more heterogenous. In the HIS, the unscaled weights w_k and $w_{i|jk}$ were characterized by a mean of 4.17 and 1.04 and standard deviation of 4.43 and 0.21 respectively, thus revealing substantial variability at the municipality level. Note that it is assumed there is no differential sampling effect at the household level.

The estimated variance components show that there is little clustering effect at the municipality level and a moderate effect at the household level, with an estimated value of 0.19 for ρ_{HH}. The standard error for this parameter was obtained using the delta method for a ratio of two parameters (Herson, 1975).

To check the model, some diagnostic plots were examined. Thus, a plot of the level 1 residuals versus fixed part predictor did not reveal any special pattern, while normal probability plots of standardized residuals (at levels 1, 2 and 3) did not show any severe departures from the normality assumption, only pointing to a few extreme values.

2.6.2 Multilevel Logistic Model

Due to unit and item non-response, 7254 out of 8560 (85%) observations were available for analysis. We fit model (2.3) with a logit link function and include random intercepts at the household and municipality levels. The same covariates as before were considered for inclusion in the model.

Table 2.1. *Linear multilevel regression model on log(BMI). Weighted and unweighted estimators are reported with robust standard errors given in parentheses.*

	Unweighted	**Weighted**	**Unweighted**	**Weighted**
$\boldsymbol{\beta}$ Intercept	3.140 (0.021)	3.120 (0.034)		
Smoking status: (1=smoker)	-0.038 (0.004)	-0.035 (0.007)		
Age (categorical):	Males		Females	
15-24	-	-	-0.073 (0.022)	0.076 (0.033)
25-34	0.053 (0.011)	0.041 (0.018)	-0.038 (0.022)	-0.014 (0.034)
35-44	0.088 (0.009)	0.087 (0.023)	-0.035 (0.023)	-0.017 (0.038)
45-54	0.125 (0.011)	0.115 (0.019)	-0.041 (0.022)	-0.027 (0.033)
55-64	0.158 (0.012)	0.152 (0.018)	-0.067 (0.027)	-0.040 (0.037)
65-74	0.149 (0.012)	0.124 (0.016)	-0.068 (0.029)	-0.033 (0.037)
75+	0.079 (0.014)	0.068 (0.020)	-0.047 (0.025)	-0.035 (0.036)
Education:	Males		Females	
No diploma	-	-	-	-
Primary	-0.065 (0.021)	-0.044 (0.033)	0.073 (0.027)	0.055 (0.035)
Lower secondary	-0.089 (0.019)	-0.082 (0.033)	0.095 (0.021)	0.098 (0.034)
Higher secondary	-0.106 (0.019)	-0.091 (0.033)	0.097 (0.022)	0.088 (0.035)
Higher	-0.151 (0.020)	-0.135 (0.034)	0.122 (0.023)	0.108 (0.034)
Province:				
Brussels	-	-		
Antwerpen	-0.001 (0.007)	0.005 (0.008)		
Vlaamse Brabant	-0.005 (0.008)	-0.003 (0.009)		
Limburg	0.013 (0.012)	0.012 (0.021)		
Oost Vlaanderen	0.007 (0.008)	0.009 (0.008)		
West Vlaanderen	-0.017 (0.007)	0.028 (0.013)		
Brabant Wallon	0.022 (0.014)	0.028 (0.014)		
Hainaut	0.029 (0.007)	0.036 (0.010)		
Liege	0.022 (0.009)	0.010 (0.017)		
Luxembourg	0.025 (0.010)	0.023 (0.009)		
Namur	0.022 (0.009)	0.025 (0.013)		
German community	0.009 (0.008)	0.008 (0.010)		
Household size:	0.005 (0.002)	0.006 (0.002)		
Interviewee status:	0.034 (0.007)	0.035 (0.010)		
σ_v^2	0.0001 (0.0001)	0.0002 (0.0001)		
σ_u^2	0.004 (0.001)	0.004 (0.001)		
σ_e^2	0.019 (0.001)	0.019 (0.003)		
$\rho_{HH}^{\dagger}$	0.190 (0.016)	0.199 (0.025)		

† Standard errors were calculated using the delta method.

Interpretation of variance components in terms of intra-unit correlation coefficients is no longer simple in generalized linear multilevel models. We already touched upon the issue that variance components at level one and higher levels are not directly comparable, which precludes using formulas such as (2.9) and (2.10). In addition, unlike in the Gaussian case where the mean and variance are independent, here the level one variance, $\text{var}(y_{ijk}) = \pi_{ijk}(1 - \pi_{ijk})$, depends on the expected value and hence on the fixed predictors included in the model. Goldstein, Browne and Rasbash (2002) discuss four extensions of the intra-unit correlation coefficient to GLMMs.

Another issue is related to weighting. The paper of Pfeffermann *et al.* (1998) was framed in the context of linear multilevel models. It can be argued that a similar procedure applies to generalized linear multilevel models, though. Weighted explanatory variables at level 2 and above are defined as before. Level 1 being characterized by the binomial variation, a method of incorporating the weights is now to use $w_{ijk}n_{ijk}$ instead of n_{ijk} as the denominator of the binomial response variable. We found this procedure to be numerically unstable in our application, however. We ran the procedure on three different binary outcomes. With two of them (including the perceived health indicator), it did not converge, and in the last instance the final results were quite dramatically different from the unweighted analysis. It is not clear if the problem stems from the weighting scheme of this particular survey or from the procedure itself, but we cannot recommend its use until it is more thoroughly explored.

Table 2.2 shows the results of the unweighted analysis for the perceived health indicator. Among the original covariates, all were found important and included in the model. An interaction term between smoking behavior and age was also included. Among design-related covariates, only province indicators and size (household, municipality) variables were retained.

We allowed for an extra-dispersion parameter in the model, that is, we assume that

$$y_{ijk} = \pi_{ijk} + z^{(1)}_{ijk} e_{ijk},$$

where e_{ijk} has mean zero and variance σ^2_e, and $z^{(1)}_{ijk} = \sqrt{\pi_{ijk}(1 - \pi_{ijk})}$. As can be seen in Table 2.2, the estimated value of σ^2_e is strongly indicative of under-dispersion. Estimated variance components again reveal that there is considerably less variability at the municipality level compared to the household level. Note that a normal probability plot of standardized residuals at the household level showed a marked departure from normality, therefore making the appropriateness of the model to these data questionable.

Table 2.2. *Multilevel logistic regression model on perceived health indicator. Unweighted estimators are reported with robust standard errors given in parentheses.*

Parameters	**Estimates (S.E.)**	**Parameters**	**Estimates (S.E.)**
$\boldsymbol{\beta}$ Intercept	1.427 (0.290)		
Sex (1=male):	0.508 (0.080)		
Age (categorical):	Non smokers		Smokers
15-24	-		-0.540 (0.303)
25-34	-0.531 (0.221)		-0.081 (0.245)
35-44	-1.298 (0.204)		-0.368 (0.166)
45-54	-1.701 (0.204)		-0.226 (0.180)
55-64	-2.250 (0.203)		0.037 (0.251)
65-74	-2.445 (0.209)		-0.324 (0.233)
75+	-2.814 (0.217)		-0.112 (0.424)
Education:		Household income:	
No diploma	-	$<$ 20,000	-
Primary	0.240 (0.206)	20,000-30,000	-0.180 (0.177)
Lower secondary	0.560 (0.191)	30,000-40,000	0.179 (0.164)
Higher secondary	0.998 (0.197)	40,000-60,000	0.510 (0.170)
Higher	1.301 (0.208)	$>$ 60,000	0.982 (0.213)
Province:		Household size:	0.110 (0.035)
Brussels	-	Municipality size:	-0.163 (0.058)
Antwerpen	0.727 (0.185)		
Vlaamse Brabant	-0.080 (0.186)	σ_v^2	0.054 (0.036)
Limburg	0.118 (0.224)	σ_u^2	3.050 (0.157)
Oost Vlaanderen	0.454 (0.216)	σ_e^2	0.476 (0.023)
West Vlaanderen	0.664 (0.228)		
Brabant Wallon	-0.118 (0.141)		
Hainaut	-0.431 (0.155)		
Liege	-0.417 (0.204)		
Luxembourg	0.274 (0.225)		
Namur	-0.155 (0.151)		
German community	0.350 (0.234)		

Chapter 3

Pairwise Likelihood Estimation in Multilevel Probit Models

3.1 Introduction

In the previous chapter, we discussed two approaches to fitting multilevel models with binary responses: maximum (marginal) likelihood (ML) and penalized quasi-likelihood (PQL) estimation. Advantages and disadvantages of both methods can be roughly summarized as follows:

- ML: can be easily employed in low-dimensional problems but becomes computationally demanding as the dimension of integration grows. Whenever feasible, this ought to be the preferred method.

- PQL/PQL2: is computationally efficient but parameter estimates are biased, albeit to a lesser extent with PQL2.

Consequently, a method enjoying attractive asymptotic properties (such as consistency), while being computationally efficient in complex problems, is not without interest. This will be our principal motivation for investigating maximum pairwise likelihood (MPL) as an estimation tool in multilevel models with binary responses.

Pairwise likelihood is a special example of what is called *pseudo-likelihood*, first proposed by Besag (1975) and also termed *composite likelihood* by Lindsay (1988). The motivation behind pseudo-likelihood estimation is to replace the likelihood by a function that is easier to evaluate, and hence to maximize. The function in question is a product of conditional or marginal densities. Thus, the main feature of a pseudo-likelihood function is that it is composed of (pieces of) likelihoods and this can be exploited to prove general results about the consistency and asymptotic normality of pseudo-likelihood estimators, as shown in Section 3.2.

The aim of this chapter is to study a particular type of pseudo-likelihood function, namely, pairwise likelihood. Using a multilevel probit model, we present the method in Section 3.3 and discuss some of its computational properties. In Section 3.4 some calculations are reported to compare the asymptotic efficiency of the maximum pairwise likelihood estimator (MPLE) relative to the MLE. The issue of weighting each contribution in the log PL by the inverse cluster size (minus one) is examined in Section 3.5. Different authors have argued that in marginal models weighting should be applied for the estimation of marginal regression parameters but not for the estimation of association parameters. In particular, we shall look closely at the argument given by Kuk and Nott (2000) based on the theory of optimal estimating functions and also report some asymptotic efficiency calculations. Next, an illustration of the estimation method is given in Section 3.6 using data from a set of five clinical trials comparing antipsychotic agents for the treatment of chronic schizophrenia. In Section 3.7, we investigate finite-sample performances of the MPLE and report results of a simulation study that was conducted to compare ML, PQL2 and MPL estimators in relatively simple models. We conclude this chapter by a more extensive discussion of the relative merits of the three estimation methods. Note that a shortened version of this chapter is given in Renard, Molenberghs and Geys (2002).

3.2 Pseudo-Likelihood Estimation

In this section, the concept of pseudo-likelihood is introduced and general results on the consistency and asymptotic normality of pseudo-likelihood estimators are proved.

3.2.1 Pseudo-Likelihood Definition

For notational convenience, we define pseudo-likelihoods for a two-level data structure but this can be extended to a higher number of levels. For simplicity, we also assume that all clusters are of equal size, L say. The pseudo-likelihood definition presented hereafter is in accordance with that given by Arnold and Strauss (1991) and Geys (1999).

Let $\boldsymbol{y}_j = (y_{ij} : i = 1, \ldots, L)^T$ denote the vector of measurements on unit j ($j = 1, \ldots N$). Later on, we shall assume that the y_{ij}'s are binary variables but for the time being, we put no restriction on the type of outcomes. We make the following standard assumptions about the vectors of observations $\boldsymbol{y}_j$:

1. They are independently distributed with probability density function $f(\boldsymbol{y}_j; \boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta}$ is an element of the p-dimensional domain Ω.

2. The distributions of the $\boldsymbol{y}_j$'s are distinct and have common support.

To construct a pseudo-likelihood, one starts with a set of conditional or marginal likelihoods involving the y_{ij}'s. Since the vectors $\boldsymbol{y}_j$ are assumed to be independent, we restrict attention to joint densities, or ratios thereof, which involve observations y_{ij} sharing the same index j.

More formally, let us define the set S of all $2^L - 1$ vectors of length L consisting solely of zeros and ones, with at least one non zero entry. Denote by $\boldsymbol{y}_j^{(\boldsymbol{s})}$ ($\boldsymbol{s} \in S$) the subvector of $\boldsymbol{y}_j$ corresponding to the components of $\boldsymbol{s}$ that are non zero, and by $f_{\boldsymbol{s}}(\boldsymbol{y}_j^{(\boldsymbol{s})}; \boldsymbol{\theta})$ the associated joint density function. To complete the notation, let $\delta = \{\delta_{\boldsymbol{s}} \mid \boldsymbol{s} \in S\}$ be a set of real numbers (not all zero).

The log of the pseudo-likelihood function is defined as

$$\begin{aligned} \log PL(\delta; \boldsymbol{\theta}) &= \sum_{j=1}^{N} \sum_{\boldsymbol{s} \in S} \delta_{\boldsymbol{s}} \log f_{\boldsymbol{s}}(\boldsymbol{y}_j^{(\boldsymbol{s})}; \boldsymbol{\theta}) \\ &= \sum_{j=1}^{N} p\ell_j(\boldsymbol{\theta}). \end{aligned} \tag{3.1}$$

In this expression, the numbers $\delta_{\boldsymbol{s}}$ are not allowed to take on any value. In particular, some of them may be negative but they must have arisen from a product of likelihoods and conditional likelihoods.

The classical log-likelihood function is found by setting $\delta_{\boldsymbol{s}} = 1$ if $\boldsymbol{s}$ is the vector consisting solely of ones, and $\delta_{\boldsymbol{s}} = 0$ otherwise. Another example of pseudo-likelihood

function consists of the product of full conditional densities $f(y_{ij}|\{y_{i'j} : i' \neq i\};\theta)$ obtained by conditioning a component on all the others. In this case, the log pseudo-likelihood is defined by taking $\delta_{\mathbf{1}_L} = L$ and $\delta_{s_i} = -1$ $(i = 1, \ldots, L)$, where $\mathbf{1}_L$ is the vector of ones (of dimension L) and s_i consists of ones everywhere except for the ith entry. The full conditional pseudo-likelihood is particularly useful in exponential family models where it has the effect of replacing a joint density with a possibly complicated normalizing constant by a product of simple univariate functions (Geys, 1999).

By analogy with ML estimation, a *pseudo-likelihood estimator* can be constructed as the value of the parameter $\boldsymbol{\theta}$ which maximizes the log pseudo-likelihood (3.1). This value can be found by solving the *pseudo-score equations* obtained after differentiating (3.2) and setting the derivatives to zero, i.e.,

$$\frac{\partial}{\partial\theta_k}\log PL(\boldsymbol{\theta};\delta) = \sum_{j=1}^{N}\sum_{\boldsymbol{s}\in S_j}\delta_{\boldsymbol{s}}\frac{\partial f_{\boldsymbol{s}}(\boldsymbol{y}_j^{(\boldsymbol{s})};\boldsymbol{\theta})/\partial\theta_k}{f_{\boldsymbol{s}}(\boldsymbol{y}_j^{(\boldsymbol{s})};\boldsymbol{\theta})} = 0, \quad k = 1, \ldots, p. \tag{3.2}$$

As shown in the next section, the pseudo-likelihood estimator enjoys attractive asymptotic properties, making it valuable for estimation purposes.

3.2.2 Asymptotic Properties of Pseudo-Likelihood Estimators

We first state some assumptions:

(A1) There exists an open set ω of Ω containing the true parameter value $\boldsymbol{\theta}_0$ such that for all $\boldsymbol{s} \in S$ and for almost all $\boldsymbol{y}^{(\boldsymbol{s})}$, the density $f(\boldsymbol{y}^{(\boldsymbol{s})};\boldsymbol{\theta})$ admits all third derivatives

$$\frac{\partial^3}{\partial\theta_k\partial\theta_l\partial\theta_m}f(\boldsymbol{y}^{(\boldsymbol{s})};\boldsymbol{\theta}),$$

for all $\boldsymbol{\theta} \in \omega$.

(A2) For all $\boldsymbol{s} \in S$, the first and second logarithmic derivatives of $f_{\boldsymbol{s}}$ satisfy the equations

$$E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\theta_k}\log f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})};\boldsymbol{\theta})\right] = 0, \qquad k = 1, \ldots, p, \tag{3.3}$$

and

$$E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\theta_k}\log f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})};\boldsymbol{\theta})\cdot\frac{\partial}{\partial\theta_l}\log f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})};\boldsymbol{\theta})\right] \tag{3.4}$$
$$= E_{\boldsymbol{\theta}}\left[\frac{-\partial^2}{\partial\theta_k\partial\theta_l}\log f_{\boldsymbol{s}}(\boldsymbol{y}^{(\boldsymbol{s})};\boldsymbol{\theta})\right], \qquad k, l = 1, \ldots, p,$$

where $E_{\boldsymbol{\theta}}[.]$ denotes expectation taken with respect to the density $f_s(\boldsymbol{y}^{(s)};\boldsymbol{\theta})$.

(A3) The matrix $J(\boldsymbol{\theta})$ defined by

$$J_{kl}(\boldsymbol{\theta}) = -\sum_{s\in S} \delta_s E_{\boldsymbol{\theta}} \left[\frac{\partial^2}{\partial\theta_k \partial\theta_\ell} \log f_s(\boldsymbol{y}^{(s)};\boldsymbol{\theta}) \right] \tag{3.5}$$

has finite elements and is positive definite for all $\boldsymbol{\theta}$ in ω.

(A4) There exist functions M_{klm} such that

$$\left| \sum_{s\in S} \delta_s E_{\boldsymbol{\theta}} \frac{\partial^3}{\partial\theta_k \partial\theta_l \partial\theta_m} \log f_s(\boldsymbol{y}^{(s)};\boldsymbol{\theta}) \right| \leq M_{klm}(\boldsymbol{y}) \qquad \text{for all } \boldsymbol{\theta} \in \omega,$$

with $\mu_{klm} = E_{\boldsymbol{\theta}_0}[M_{klm}(\boldsymbol{y})] < \infty$.

Theorem (Consistency and Asymptotic Normality). *Under assumptions (A1) – (A4), the pseudo-likelihood estimator $\tilde{\boldsymbol{\theta}}_N$, defined as the maximizer of (3.1), has the following properties:*

1. *$\tilde{\boldsymbol{\theta}}_N$ is consistent for estimating $\boldsymbol{\theta}_0$.*
2. *$\sqrt{N}(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix*

$$\Lambda = J(\boldsymbol{\theta}_0)^{-1} K(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \tag{3.6}$$

where $J(\boldsymbol{\theta})$ is defined by (3.5) and $K(\boldsymbol{\theta})$ by

$$K_{kl}(\boldsymbol{\theta}) = \sum_{s,t\in S} \delta_s \delta_t E_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial\theta_k} \log f_s(\boldsymbol{y}^{(s)};\boldsymbol{\theta}) \frac{\partial}{\partial\theta_l} \log f_t(\boldsymbol{y}^{(t)};\boldsymbol{\theta}) \right]. \tag{3.7}$$

The proofs are closely related to the classical proofs for maximum likelihood estimators (see Lehmann (1983, p. 429–434) for example).

Proof of Consistency

To prove the existence, with probability tending to 1, of a sequence of solutions of the pseudo-score equations which is consistent, we shall consider the behavior of the log pseudo-likelihood on a sphere Q_a with center at the true point $\boldsymbol{\theta}_0$ and radius a. If it can be shown that for any sufficiently small $a > 0$ the probability tends to 1 that

$$\log PL(\boldsymbol{\theta}) < \log PL(\boldsymbol{\theta}_0)$$

for all points $\boldsymbol{\theta}$ on the surface of Q_a, then the function $\log PL(\boldsymbol{\theta})$ has a local maximum in the interior of Q_a. Since at this local maximum the pseudo-likelihood equations are satisfied, it will follow that for any sufficiently small $a > 0$, with probability tending to 1 as $N \to \infty$, the pseudo-likelihood equations have a solution $\tilde{\boldsymbol{\theta}}_N(a)$ within Q_a. To ensure the existence of a consistent root that does not depend on a, we can take $\tilde{\boldsymbol{\theta}}^*_N$ defined as the root closest to $\boldsymbol{\theta}_0$. Such a point exists since the limit of a sequence of roots is again a root by continuity.

Let $a > 0$ such that Q_a is included in ω. We first expand the log pseudo-likelihood around the true point $\boldsymbol{\theta}_0$. After division by N, we have

$$\frac{1}{N}\log PL(\boldsymbol{\theta}) - \frac{1}{N}\log PL(\boldsymbol{\theta}_0) = \frac{1}{N}\sum_{k=1}^{p} P_k(\boldsymbol{y})(\theta_k - \theta_{0k})$$

$$+\frac{1}{2N}\sum_{k=1}^{p}\sum_{l=1}^{p} Q_{kl}(\boldsymbol{y})(\theta_k - \theta_{0k})(\theta_l - \theta_{0l}) \tag{3.8}$$

$$+\frac{1}{6N}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{m=1}^{p}(\theta_k - \theta_{0k})(\theta_l - \theta_{0l})(\theta_m - \theta_{0m})\sum_{j=1}^{N}\gamma_{klm}(\boldsymbol{y}_j)M_{klm}(\boldsymbol{y}_j),$$

where

$$P_k(\boldsymbol{y}) = \frac{\partial}{\partial\theta_k}\log PL(\boldsymbol{\theta}_0),$$
$$Q_{kl}(\boldsymbol{y}) = \frac{\partial^2}{\partial\theta_k\partial\theta_l}\log PL(\boldsymbol{\theta}_0),$$

and

$$0 \le |\gamma_{klm}(\boldsymbol{y})| \le 1.$$

Next it can be noted, by (A2) and the law of large numbers, that

$$\frac{1}{N}P_k(\boldsymbol{y}) = \frac{1}{N}\sum_{j=1}^{N}\sum_{\boldsymbol{s}\in S}\delta_{\boldsymbol{s}}\frac{\partial}{\partial\theta_k}\log f_{\boldsymbol{s}}(\boldsymbol{y}_j^{(\boldsymbol{s})};\boldsymbol{\theta}_0) \xrightarrow{P} 0, \tag{3.9}$$

where $\xrightarrow{P}$ denotes convergence in probability. Likewise, we have

$$\frac{1}{N}Q_{kl}(\boldsymbol{y}) = \frac{1}{N}\sum_{j=1}^{N}\sum_{\boldsymbol{s}\in S}\delta_{\boldsymbol{s}}\frac{\partial^2}{\partial\theta_k\theta_l}\log f_{\boldsymbol{s}}(\boldsymbol{y}_j^{(\boldsymbol{s})};\boldsymbol{\theta}_0) \xrightarrow{P} -J_{kl}(\boldsymbol{\theta}_0). \tag{3.10}$$

Let us write the three terms on the right-hand side of (3.8) S_1, S_2 and S_3 respectively. Then on Q_a we have

$$|S_1| \le \frac{a}{N}\sum_{k=1}^{p}|P_k(\boldsymbol{y})|.$$

It follows from (3.9) that $\frac{1}{N}|P_k(\boldsymbol{y})| < a^2$ and hence $|S_1| < pa^3$ with probability tending to 1.

Next, we consider

$$2S_2 = \sum_{k=1}^{p}\sum_{l=1}^{p}\big[-J_{kl}(\boldsymbol{\theta}_0)(\theta_k-\theta_{0k})(\theta_l-\theta_{0l})\big] + \sum_{k=1}^{p}\sum_{l=1}^{p}\Big\{\frac{1}{N}Q_{kl}(\boldsymbol{y}) - \big[-J_{kl}(\boldsymbol{\theta}_0)\big]\Big\}(\theta_k-\theta_{0k})(\theta_l-\theta_{0l}).$$

From an argument analogous to that for S_1, but based on (3.10), it follows that the absolute value of the second term is less than p^2a^3 with probability tending to 1. The first term is a quadratic form in $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$. For $\boldsymbol{\theta}$ on Q_a, this can be reduced, by an appropriate orthogonal transformation, to a diagonal form $\sum\lambda_k\zeta_k^2$ with $\sum\zeta_k^2 = a^2$, where the λ_k's are the (negative) eigenvalues of the matrix $-J(\boldsymbol{\theta}_0)$. Suppose that the λ_k's are numbered so that $\lambda_p \le \lambda_{p-1} \le \ldots \le \lambda_1 < 0$. Then $\sum\lambda_k\zeta_k^2 \le \lambda_1 a^2$. Combining the above two arguments, we see that there exists $c > 0$ such that for any a sufficiently small,

$$S_2 < -ca^2$$

with probability tending to 1.

Finally, by (A4) we have

$$\frac{1}{N}\sum_{j=1}^{N}M_{klm}(\boldsymbol{y}_j) \xrightarrow{P} \mu_{klm},$$

which implies, with probability tending to 1,

$$\frac{1}{N}\sum_{j=1}^{N}M_{klm}(\boldsymbol{y}_j) < 2\mu_{klm}$$

and hence $S_3 < ba^3$ where

$$b = \frac{1}{3}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{m=1}^{p}\mu_{klm}.$$

Combining the three inequalities, we conclude that

$$\max_{Q_a}(S_1+S_2+S_3) < -ca^2 + (b+p)a^3$$

with probability tending to 1. This completes the proof since the right-hand expression is smaller than zero if $a < c/(b+p)$.

Proof of Asymptotic Normality

We start by expanding $\partial \log PL(\boldsymbol{\theta})/\partial\theta_k$ about $\boldsymbol{\theta}_0$:

$$\frac{\partial}{\partial\theta_k}\log PL(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_k}\log PL(\boldsymbol{\theta}_0) + \sum_{l=1}^{p}(\theta_l-\theta_{0l})\frac{\partial^2}{\partial\theta_k\partial\theta_l}\log PL(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{l=1}^{p}\sum_{m=1}^{p}(\theta_l-\theta_{0l})(\theta_m-\theta_{0m})\frac{\partial^3}{\partial\theta_k\partial\theta_l\partial\theta_m}\log PL(\boldsymbol{\theta}^*), \tag{3.11}$$

where $\boldsymbol{\theta}^*$ is a point on the line segment connecting $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}$. We know, by the first part of the proof, that there exists, with probability tending to 1, a solution $\tilde{\boldsymbol{\theta}}_N$ of the pseudo-likelihood equations which is consistent. At $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_N$, the left side of (3.11) is zero and the resulting equations can be rewritten

$$\sqrt{N}\sum_{l=1}^{p}(\tilde{\theta}_{Nl}-\theta_{0l})\left[\frac{-1}{N}\frac{\partial^2}{\partial\theta_k\partial\theta_l}\log PL(\boldsymbol{\theta}_0) - \frac{1}{2N}\sum_{m=1}^{p}(\tilde{\theta}_{Nm}-\theta_{0m})\frac{\partial^3}{\partial\theta_k\partial\theta_l\partial\theta_m}\log PL(\boldsymbol{\theta}^*)\right] = \frac{1}{\sqrt{N}}\frac{\partial}{\partial\theta_k}\log PL(\boldsymbol{\theta}_0).$$

This set of equations has the form (3.12) of the lemma stated below, if we define:

$$\begin{aligned} T_{Nk} &= \frac{1}{\sqrt{N}}\frac{\partial}{\partial\theta_k}\log PL(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\frac{\partial}{\partial\theta_k}p\ell_j(\boldsymbol{y}_j;\boldsymbol{\theta}_0),\\ Y_{Nl} &= \sqrt{N}(\tilde{\theta}_{Nl}-\theta_{0l}),\\ A_{Nkl} &= \frac{-1}{N}\frac{\partial^2}{\partial\theta_k\partial\theta_l}\log PL(\boldsymbol{\theta}_0) - \frac{1}{2N}\sum_{m=1}^{p}(\tilde{\theta}_{Nm}-\theta_{0m})\frac{\partial^3}{\partial\theta_k\partial\theta_l\partial\theta_m}\log PL(\boldsymbol{\theta}^*). \end{aligned}$$

Since $E_{\boldsymbol{\theta}_0}[\partial p\ell_j(\boldsymbol{y}_j;\boldsymbol{\theta}_0)/\partial\theta_k] = 0$ by (A2), the multivariate central limit theorem implies that $\boldsymbol{T}_N = (T_{N1},\ldots,T_{Np})$ has a multivariate normal limit distribution $\boldsymbol{T}$ with mean vector $\mathbf{0}$ and variance-covariance matrix $K(\boldsymbol{\theta}_0)$.

We also know, by (3.10), that

$$\frac{-1}{N}\frac{\partial^2}{\partial\theta_k\partial\theta_l}\log PL(\boldsymbol{\theta}_0) \xrightarrow{P} J_{kl}(\boldsymbol{\theta}_0).$$

Since by (A4)

$$\frac{\partial^3}{\partial\theta_k\partial\theta_l\partial\theta_m}\log PL(\boldsymbol{\theta}^*)$$

is bounded, we have that

$$A_{Nkl} \xrightarrow{P} a_{kl} = J_{kl}(\boldsymbol{\theta}_0).$$

By the lemma, we therefore conclude that the limit distribution of $(Y_{N1}, \ldots, Y_{Np})$ is that of

$$J(\boldsymbol{\theta}_0)^{-1}\boldsymbol{T},$$

which is multivariate normal with mean zero and covariance matrix

$$\Lambda = J(\boldsymbol{\theta}_0)^{-1}K(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}.$$

Lemma. *Consider the following set of random linear equations in p unknowns*

$$T_{Nk} = \sum_{l=1}^{p} A_{Nkl}Y_{Nl}, \qquad (k = 1, \ldots, p). \tag{3.12}$$

Suppose that $(T_{N1}, \ldots, T_{Np})$ is a sequence of random vectors converging weakly to $(T_1, \ldots, T_p)$ and that for each fixed k and l, A_{Nkl} is a sequence of random variables converging in probability to constants a_{kl} for which the matrix $A = (a_{kl})$ is nonsingular. Let $B = A^{-1}$. Then $(Y_{N1}, \ldots, Y_{Np})$ converges in probability to the solutions $(Y_1, \ldots, Y_p)$ of

$$\sum_{l=1}^{p} a_{kl}Y_l = T_k \qquad (k = 1, \ldots, p)$$

given by

$$Y_k = \sum_{l=1}^{p} b_{kl}T_l.$$

A proof of this lemma can be found in Lehmann (1983, p. 432–433).

It is interesting to highlight the connection between pseudo-likelihood and estimating equations. To model correlated data, many researchers have proposed the use of estimation techniques which do not require knowledge of the whole distribution of the response vector, thereby avoiding the need of prohibitive computations for the likelihood. The classical score equations are then replaced by so-called *estimating equations* which are easier to solve. If $g(\boldsymbol{y};\theta)$ is an *estimating function*, an estimator can be defined as the solution of the estimating equation $g(\boldsymbol{y};\theta) = 0$. A well-known example of this approach to modeling of correlated discrete data are the generalized estimating equations of Liang and Zeger (1986), a multivariate extension of quasilikelihood (Wedderburn, 1974)

The theory of estimating equations is in fact quite general and has been in development for a long time now (Godambe, 1991). A special feature of estimating functions is unbiasedness. An estimating function $g(\boldsymbol{y};\theta)$, and the associated estimating equation $g(\boldsymbol{y};\theta) = 0$, are said to be *unbiased* if $E_\theta[g(\boldsymbol{y};\theta)] = 0$ for any θ. Thus, the classical score equations are a trivial example of unbiased estimating equations. The pseudo-score equations (3.2) are another one since a pseudo-likelihood is nothing else but a sum of likelihoods. The reason why unbiased estimating equations are special is that general results can be derived about the consistency and asymptotic normality of estimators obtained from such equations, and the above theorem can be framed in this general theory.

A major advantage of pseudo-likelihood over other estimating equation approaches is that we face an optimization problem, i.e. maximizing the pseudo-likelihood function. As a result, the value of the pseudo-likelihood can be used to distinguish between multiple roots of the pseudo-score equations, or to discriminate between different non-nested models with the same number of parameters. Also, in addition to pseudo-score test statistics such as those available for GEE (Rotnitzky and Jewell, 1990), pseudo-likelihood ratio statistics can be defined, as discussed by Geys (1999).

In the maximum likelihood case, each of the matrices $J(\boldsymbol{\theta}_0)$ and $K(\boldsymbol{\theta}_0)$ reduces to Fisher's information $I(\boldsymbol{\theta}_0)$ and the covariance matrix (3.6) simplifies to $I(\boldsymbol{\theta}_0)^{-1}$. As pointed out by Arnold and Strauss (1991), the Cramèr-Rao inequality implies that $J(\boldsymbol{\theta}_0)^{-1}K(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}$ is greater than $I(\boldsymbol{\theta}_0)^{-1}$, in the sense that $J(\boldsymbol{\theta}_0)^{-1}K(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1} - I(\boldsymbol{\theta}_0)^{-1}$ is positive semi-definite, with strict inequality if $\tilde{\boldsymbol{\theta}}_N$ is not a function of a minimal sufficient statistic. In other words, maximum likelihood estimators will be, in general, more efficient than maximum pseudo-likelihood estimators. Sacrificing some efficiency is therefore the price we pay for computational simplicity.

In practice, the covariance matrix of the pseudo-likelihood estimator $\tilde{\boldsymbol{\theta}}_N$ can be (consistently) estimated by

$$\tilde{\Lambda}_N = J_N^{-1} K_N J_N^{-1} \tag{3.13}$$

with

$$J_N = -\sum_{j=1}^{N} \sum_{\boldsymbol{s}\in S_j} \delta_{\boldsymbol{s}} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \log f_{\boldsymbol{s}}(\boldsymbol{y}_j^{(\boldsymbol{s})};\tilde{\boldsymbol{\theta}}_N) \tag{3.14}$$

and

$$K_N = \sum_{j=1}^{N} \sum_{s,t \in S_j} \delta_s \delta_t \frac{\partial}{\partial \boldsymbol{\theta}} \log f_s(\boldsymbol{y}_j^{(s)}; \tilde{\boldsymbol{\theta}}_N) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log f_t(\boldsymbol{y}_j^{(t)}; \tilde{\boldsymbol{\theta}}_N). \tag{3.15}$$

The above expression can be recognized as a 'sandwich' estimator, similar in spirit to the robust variance estimate of Liang and Zeger (1986). Royall (1986) discussed general properties of this type of estimator and gave some applications. Based on (3.4) a simpler estimate of $J(\boldsymbol{\theta}_0)$, which does not require evaluation of second-order derivatives, is given by

$$J_N = \sum_{j=1}^{N} \sum_{s \in S_j} \delta_s \frac{\partial}{\partial \boldsymbol{\theta}} \log f_s(\boldsymbol{y}_j^{(s)}; \tilde{\boldsymbol{\theta}}_N) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log f_s(\boldsymbol{y}_j^{(s)}; \tilde{\boldsymbol{\theta}}_N). \tag{3.16}$$

3.3 Pairwise Likelihood in the Multilevel Probit Model

In the previous section, a general pseudo (or composite) likelihood method of estimation was introduced. We now restrict attention to a specific form of pseudo-likelihood, namely, pairwise likelihood (PL). The technique does not seem to have been used very much in applications until quite recently. It has been most advantageously used in the spatial data context, where the full likelihood distribution is typically cumbersome. Hjort (1993), for example, applies the method to covariance estimation in geostatistics, while Heagerty and Lele (1998) use it to model binary spatial data. Another broad class of applications is related to marginal modeling of correlated binary data. Thus, Le Cessie and Van Houwelingen (1994) developped a model with logistic marginal response probabilities, using the odds ratio or the tetrachoric correlation as a measure of association. Geys, Molenberghs and Ryan (1997) investigate the performance of the maximum pairwise likelihood estimator in a model proposed by Molenberghs and Ryan (1999), based on a multivariate exponential family model. Geys, Molenberghs and Lipsitz (1998) compare pairwise likelihood with other estimating equations approaches (GEE1 and GEE2) in marginally specified odds ratio models with exchangeable association structure, whereas Kuk and Nott (2000) examine pairwise likelihood in a model with a more general specification for the association structure, similar to the alternating logistic regression approach of Carey, Zeger and Diggle (1993). We emphasize that the above references were centered on population-averaged models, as opposed to the cluster-specific modeling approach endorsed here.

3.3.1 The Multilevel Probit Model

Methodological developments in the remainder of this chapter assume a two-level structure for simplicity. The procedure described hereafter can, in principle, be applied to general multilevel structures, although practical limitations at the computational level arise (see Section 3.8 for further discussion).

As it will become apparent, a probit link must be postulated for full computational gains. The logit link, on the other hand, has been quite popular in marginal models because parameters are interpretable in terms of (log) odds ratios. By extension, it has been mostly studied in generalized linear multilevel models as well. In addition to the fact that both links tend to provide similar model fits in practice whatever modeling approach is followed, we do believe that the logit link function is less appealing with random-effects modeling because of the cluster-specific interpretation of the parameters. Moreover, the probit link is more convenient to make population-averaged inference since unconditional (i.e. marginal) probabilities are simple probit functions.

We now introduce the multilevel probit model from a latent variable perspective. Let $\boldsymbol{y}_j = (y_{ij} : i = 1, \ldots, n_j)^T$ denote the vector of binary measurements on unit j ($j = 1, \ldots N$). We assume that each y_{ij} can take on a value of 0 or 1. The model for the probability that, conditionally on a set of random effects $\boldsymbol{u}_j$, a positive response be observed on the ith unit in the jth cluster is specified as

$$\Phi^{-1}(P[y_{ij} = 1|\boldsymbol{u}_j]) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_j, \tag{3.17}$$

where $\boldsymbol{x}_{ij}$ is a vector of covariates having fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{z}_{ij}$ denotes a vector of covariates, possibly overlapping with $\boldsymbol{x}_{ij}$, having random effects $\boldsymbol{u}_j \sim N(0, \Omega_u)$ of dimension q.

We posit the existence of a latent variable $\tilde{y}_{ij}$ that is continuously distributed and related to the actual response through a certain threshold. In the context of iid binary data, this approach motivates a wide class of models, of which the standard logistic and probit regression models are special cases (Cox and Snell, 1989). We assume that the *observed* binary response is actually obtained by dichotamizing an unobserved continuous variable, hence the need for a certain threshold or 'cut-off' value. This threshold can be chosen to be 0 without loss of generality, provided an intercept term is included in the model. In other words, it is assumed that a positive response is recorded ($y_{ij} = 1$) if $\tilde{y}_{ij} > 0$ and a negative response ($y_{ij} = 0$) otherwise. If we further assume that $\tilde{y}_{ij}$ is normally distributed, then the random-effects regression model for

the latent response variable can be written as follows:

$$\tilde{y}_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_j + \tilde{\varepsilon}_{ij}, \tag{3.18}$$

where the residual error terms $\tilde{\varepsilon}_{ij}$ are assumed to be normally distributed with mean zero and variance σ_e^2.

We note that the parameter σ_e^2 is not identifiable in the model. For identifiability, its value is fixed, without loss of generality, to 1. With this choice, the derived model for the binary response variable y_{ij} is exactly (3.17).

3.3.2 Pairwise Likelihood

As the name suggests, with pairwise likelihood we aim to replace the likelihood contribution $P(y_{1j}, \ldots, y_{n_j j})$ by the product of all possible pairwise probabilities. The log pairwise likelihood for the response vector $\boldsymbol{y}_j$ can be written

$$p\ell_j(\boldsymbol{\beta}, \Omega_u) = \sum_{i=1}^{n_j}\sum_{i'>i}\sum_{k,l=0}^{1} \delta_{ii'jkl} \log P[y_{ij} = k, y_{i'j} = l], \tag{3.19}$$

with

$$\delta_{ii'jkl} = \begin{cases} 1 & \text{if } y_{ij} = k \text{ and } y_{i'j} = l, \\ 0 & \text{otherwise.} \end{cases}$$

A few remarks are in place here. Clearly, the pairwise likelihood and the classical likelihood functions coincide when clusters are of size 2, which makes the two approaches equivalent in this case. Next, we emphasize that pairwise probabilities in (3.19) are *marginal*, not conditional, probabilities. Using the latent variable formulation of the model, these marginal pairwise probabilities are straightforward to calculate in terms of univariate and bivariate probits. Thus, if we write

$$\xi_{ij} = \frac{-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{\text{var}[\tilde{y}_{ij}]}}, \qquad j = 1, \ldots, N;\ i = 1 \ldots, n_j,$$

and

$$\rho_{ii'j} = \text{corr}[\tilde{y}_{ij}, \tilde{y}_{i'j}], \qquad j = 1, \ldots, N;\ i, i' = 1 \ldots, n_j,$$

then we have:

$$\begin{aligned}
P[y_{ij}=0, y_{i'j}=0] &= P[\tilde{y}_{ij}<0, \tilde{y}_{i'j}<0] \\
&= \int_{-\infty}^{\xi_{ij}} \int_{-\infty}^{\xi_{i'j}} \phi_2(u,v;\rho_{ii'j}) dv du \\
&= \Phi_2\left(\xi_{ij}, \xi_{i'j}; \rho_{ii'j}\right), \\
P[y_{ij}=0, y_{i'j}=1] &= \int_{-\infty}^{\xi_{ij}} \int_{\xi_{i'j}}^{+\infty} \phi_2(u,v;\rho_{ii'j}) dv du \\
&= \int_{-\infty}^{\xi_{ij}} \Big[\int_{-\infty}^{+\infty} \phi_2(u,v;\rho_{ii'j}) dv - \int_{-\infty}^{\xi_{i'j}} \phi_2(u,v;\rho_{ii'j}) dv \Big] du \\
&= \Phi\left(\xi_{ij}\right) - \Phi_2\left(\xi_{ij}, \xi_{i'j}; \rho_{ii'j}\right), \\
P[y_{ij}=1, y_{i'j}=0] &= \Phi\left(\xi_{i'j}\right) - \Phi_2\left(\xi_{ij}, \xi_{i'j}; \rho_{ii'j}\right), \\
P[y_{ij}=1, y_{i'j}=1] &= 1 - \Phi\left(\xi_{ij}\right) - \Phi\left(\xi_{i'j}\right) + \Phi_2\left(\xi_{ij}, \xi_{i'j}; \rho_{ii'j}\right),
\end{aligned}$$

where the function $\phi_2(u,v;\rho)$ denotes the standardized bivariate Gaussian density function with correlation coefficient ρ, and the functions Φ and Φ_2 denote the standardized univariate and bivariate Gaussian distribution functions, respectively. In the above expressions, $\text{var}[\tilde{y}_{ij}]$, $\text{var}[\tilde{y}_{i'j}]$ and $\rho_{ii'j}$ are obtained by selecting the appropriate 2×2 submatrix of the (marginal) covariance matrix of $\tilde{\boldsymbol{y}}_j$,

$$V_j = Z_j \Omega_u Z_j^T + I_{n_j},$$

where $Z_j = (\boldsymbol{z}_{1j}, \ldots, \boldsymbol{z}_{n_j j})^T$ and I_{n_j} denotes the identity matrix of dimension n_j.

Closer inspection of expression (3.19) reveals that cluster size is the determining factor of the computational cost to evaluate a cluster's contribution to the log PL. Indeed, (3.19) entails calculating $C_2^{n_j} = n_j(n_j-1)/2$ pairwise probabilities, a number which rises rapidly with n_j. The computational cost can be greatly reduced, for example when all covariates are discrete/categorical, since one actually needs to calculate as many pairwise probabilities as there are of possible pairs of covariate values. The main advantage of (3.19), of course, is that it implies evaluation of simple probit functions which can be calculated using highly effective methods and with high ac-

curacy, and which are generally available in most statistical software packages[†]. It is clear that in some circumstances, the benefit of simple calculations will be outweighed by the sheer number of terms that contribute; hence, MPL will not always be more advantageous than ML in computational terms.

In (3.19), we have implicitly assumed that all clusters have at least two observations to make pairing possible, but often in practice, some clusters may have a single observation. Although such clusters provide no information about association parameters, they do contribute information about fixed-effects parameters. When a cluster is of size one, its contribution is simply the log likelihood based on the single observation.

To conclude, we briefly comment on the implementation of the algorithm. To remove constraints on the matrix Ω_u which should be positive-definite, and thereby improve the convergence properties of the algorithm, a Cholesky decomposition $\Sigma^T\Sigma = \Omega_u$ is used and the PL function maximized with respect to the elements of the Cholesky factor Σ. The algorithm was implemented in SAS IML (SAS Institute Inc., 1995) and maximization of the log PL was performed using the NLPDD (Double-Dogleg) optimization routine. This optimization procedure requires only function and gradient calls which are less expensive to evaluate and much easier to obtain than the hessian matrix. Upon convergence of the algorithm, estimates of the standard errors of $(\tilde{\boldsymbol{\beta}}, \tilde{\Omega}_u)$ can be obtained via (3.13), with J_N estimated using (3.15) or, more conveniently, (3.16). In the former case, the final hessian matrix can be computed using numerical second-order derivatives by forward difference approximations. Of course, the more expensive but accurate central difference approximation can be used instead.

3.4 Asymptotic Relative Efficiency

In this section we compare the asymptotic efficiency of the MPLE relative to the MLE in specific models. For simplicity, we suppose that clusters are of equal size (i.e. $n_j = n$ for all j). To evaluate asymptotic variances of the MLE and the MPLE, we must calculate the probability of each of the 2^n possible configurations of the outcome vector. For this reason, asymptotic calculations rapidly become prohibitive and we restrict attention to the case $n = 5$ here.

[†]The SAS functions PROBNORM and PROBBNRM were used to calculate univariate and bivariate probits, respectively.

Let $\boldsymbol{\theta}$ be the vector of parameters in the model (thus, $\boldsymbol{\theta}$ contains the fixed and the random parameters), and let $\hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\boldsymbol{\theta}}_{MPL}$ denote the MLE and MPLE of $\boldsymbol{\theta}$, respectively. Let $\boldsymbol{\theta}_0$ be the true parameter value and $\ell(\boldsymbol{y};\boldsymbol{\theta})$ and $p\ell(\boldsymbol{y};\boldsymbol{\theta})$ denote the (marginal) likelihood and pairwise likelihood of a cluster with outcome vector $\boldsymbol{y}$. Omitting all functional dependencies on $\boldsymbol{\theta}_0$ below, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{ML}$ can be calculated as follows:

$$\begin{aligned} V_{ML}^{-1} &= E_{\boldsymbol{\theta}_0}[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log\ell(\boldsymbol{y})] \\ &= E_{\boldsymbol{\theta}_0}[\frac{\partial}{\partial\boldsymbol{\theta}}\log\ell(\boldsymbol{y})\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\ell(\boldsymbol{y})] \\ &= \sum_{y_1=0}^{1}\cdots\sum_{y_n=0}^{1}\ell(\boldsymbol{y})\frac{\partial}{\partial\boldsymbol{\theta}}\log\ell(\boldsymbol{y})\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\ell(\boldsymbol{y}). \end{aligned} \tag{3.20}$$

The marginal probabilities $\ell(\boldsymbol{y}) = P(y_1,\ldots,y_n)$ and derivatives $\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{y})$ are obtained after integration of random effects (assuming that derivation can be passed under the integral sign in the latter case).

The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{MPL}$ is given by

$$V_{MPL} = J(\boldsymbol{\theta}_0)^{-1}K(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}, \tag{3.21}$$

with

$$\begin{aligned} J(\boldsymbol{\theta}_0) &= E_{\boldsymbol{\theta}_0}[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log p\ell(\boldsymbol{y})] \\ &= \sum_{i=1}^{n}\sum_{i'>i}E_{\boldsymbol{\theta}_0}[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log P(y_i,y_{i'})] \\ &= \sum_{i=1}^{n}\sum_{i'>i}E_{\boldsymbol{\theta}_0}[\frac{\partial}{\partial\boldsymbol{\theta}}\log P(y_i,y_{i'})\frac{\partial}{\partial\boldsymbol{\theta}^T}\log P(y_i,y_{i'})] \\ &= \sum_{i=1}^{n}\sum_{i'>i}\sum_{k,l=0}^{1}P(y_i=k,y_{i'}=l)\frac{\partial}{\partial\boldsymbol{\theta}}\log P(y_i=k,y_{i'}=l) \\ &\qquad\qquad \times\frac{\partial}{\partial\boldsymbol{\theta}^T}\log P(y_i=k,y_{i'}=l), \end{aligned}$$

and

$$\begin{aligned} K(\boldsymbol{\theta}_0) &= E_{\boldsymbol{\theta}_0}[\frac{\partial}{\partial \boldsymbol{\theta}} \log p\ell(\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log p\ell(\boldsymbol{y})] \\ &= \sum_{y_1=0}^{1} \cdots \sum_{y_n=0}^{1} P(y_1, \ldots, y_n) \sum_{i,k=1}^{n} \sum_{\substack{i' > i \\ k' > k}} \frac{\partial}{\partial \boldsymbol{\theta}} \log P(y_i, y_{i'}) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log P(y_k, y_{k'}). \end{aligned}$$

The asymptotic relative efficiency of the MPLE relative to the MLE is then defined as the ratio V_{ML}/V_{MPL}.

We consider two models for the efficiency calculations. The first one assumes a random intercept and is specified as

$$\Phi^{-1}(P[y_i = 1|u]) = \beta_0 + \beta_1 t_i + u,$$

where $\beta_0 = -1$, $\beta_1 = 0.5$ and $u \sim N(0, \sigma_u^2)$ with $\sigma_u^2 = 0.5, 1, 4$. Two regression designs are considered:

- *Two-sample design*: the vector $\boldsymbol{t} = (t_1, \ldots, t_5)$ is made up entirely of zeroes or ones, representing membership to two groups. Each group is given equal probability.
- *Trend design*: the vector $\boldsymbol{t}$ is taken to be $(0, 1, 2, 3, 4)$.

The second model examined here is a random-intercept-and-slope model specified as

$$\Phi^{-1}(P[y_i = 1|u, v]) = (\beta_0 + u) + (\beta_1 + v)t_i,$$

with the same values of β_0 and β_1 as above, and

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \right).$$

Values for the parameters σ_u^2, σ_v^2 and σ_{uv} were chosen such that $\sigma_u^2 = \sigma_v^2 = 0.5, 1, 4$ and the correlation between u and v is equal to 0, 0.5 or 0.9. Note that only the trend design can be considered in this case.

Table 3.1 presents asymptotic relative efficiencies of $(\beta_0, \beta_1, \sigma_u^2)$ in the random-intercept model. Numerical integration required to calculate (3.20) was performed using the QUAD subroutine in SAS/IML (SAS Institute Inc., 1995) which is a numerical integrator based on adaptive Romberg-type integration techniques and can achieve any prescribed accuracy (10^{-7} by default). Table 3.1 indicates that for both

Table 3.1. *Asymptotic efficiency of MPL versus ML in the random-intercept model.*

	Two-sample design			**Trend design**		
	σ_u^2			σ_u^2		
Parameter	0.5	1	4	0.5	1	4
β_0	99.0	98.3	97.4	99.9	99.8	98.7
β_1	99.3	98.8	97.8	100.0	99.9	99.5
σ_u^2	98.3	97.8	96.7	99.8	99.2	97.0

Cell entries are asymptotic relative efficiencies given as percentages.

designs, asymptotic efficiency is close to optimal for all three parameter estimates. Obviously, this could be expected since a random-intercept model assumes that the correlation between any two observations (with fixed values of the covariates) is constant. Therefore, most of the information about the correlation structure is likely to be recovered by pairwise likelihood. Finally, we note that the MPLE is slightly less efficient with increasing values of σ_u^2, that is, with stronger correlation between responses.

Table 3.2 reports asymptotic relative efficiencies of $(\beta_0, \beta_1, \sigma_u^2, \sigma_{uv}, \sigma_v^2)$ in the random-intercept-and-slope model. To ensure that numerical integration be sufficiently precise, a basic Monte Carlo estimator was used. The technique of antithetic variates (Lange, 1998) was utilized to gain in accuracy. A number of 500000 random values were generated to estimate the integrals. Asymptotic relative efficiencies are presented for each combination of values of $\sigma_u^2 = \sigma_v^2$ and

$$\rho = \frac{\sigma_{uv}}{\sqrt{\sigma_u^2 \sigma_v^2}}.$$

As can be seen in Table 3.2, asymptotic relative efficiencies for the fixed parameters are relatively high, with an efficiency loss inferior to 20% for β_0 and inferior to 10% for β_1. For random parameters, the loss of efficiency is most severe for the σ_u^2 parameter and attains a level of about 35%. The other two random parameters have higher relative efficiency (generally larger than 80%).

Table 3.2. *Asymptotic efficiency of MPL versus ML in the random-intercept-and-slope model.*

	$\sigma_u^2 = \sigma_v^2 = 0.5$			$\sigma_u^2 = \sigma_v^2 = 1$			$\sigma_u^2 = \sigma_v^2 = 4$		
	ρ			ρ			ρ		
Parameter	0	0.5	0.9	0	0.5	0.9	0	0.5	0.9
β_0	86.0	84.9	83.2	84.6	83.0	81.4	81.0	81.2	81.6
β_1	92.4	92.0	91.9	92.1	91.7	91.2	90.4	91.6	93.0
σ_u^2	66.1	64.2	63.6	66.7	64.9	64.2	64.8	65.8	65.1
σ_{uv}	74.7	77.0	84.1	80.6	89.8	96.6	87.8	94.5	90.6
σ_v^2	92.4	94.2	95.3	92.0	95.2	96.4	82.3	91.9	97.3

Cell entries are asymptotic relative efficiencies given as percentages.

3.5 Weighted Pairwise Likelihood

Expression (3.19) showed how the log pairwise likelihood function can be constructed. Interestingly, each response y_{ij} occurs $(n_j - 1)$ times in this expression. In consequence, when cluster sizes are unequal, an observation coming from a large cluster will contribute more information to the log pairwise likelihood than one taken from a small cluster and the issue arises as to whether one should weight or not each term in the log pairwise likelihood.

Le Cessie and Van Houwelingen (1994) argued that $p\ell_j$ should be inversely weighted by a factor $(n_j - 1)$, which led them to define the weighted (log) pairwise likelihood (WPL):

$$\log PL^*(\boldsymbol{\beta}, \Omega_u) = \sum_{j=1}^{N} \frac{p\ell_j(\boldsymbol{\beta}, \Omega_u)}{n_j - 1}.$$

Their argument is that if the observations are in fact independent, the contribution of each observation is counted $(n_j - 1)$ times in $p\ell_j$ and so observations in the large clusters are given more weight than observations in the small clusters whereas they should be treated equally under independence. Therefore, the log WPL reduces to the log likelihood under independence and can be expected to be nearly efficient under weak dependence.

In marginal models with exchangeable correlation structure, Geys, Molenberghs

and Lipsitz (1998) show that PL and GEE2 approaches have strong common bases. More specifically, the set of GEE2 equations for estimating marginal parameters is closely related to the corresponding set of WPL equations, whereas the set of GEE2 equations for estimating association parameters is closely related to the corresponding set of PL equations. The authors conclude that when main interest lies in the marginal parameters, WPL should be preferred whereas if main interest lies in the association parameters, use of PL should be recommended.

Kuk and Nott (2000) support this conclusion using an argument based on the theory of optimal estimating functions. They show that for a simple model with marginal probability π (no covariates) and common pairwise correlation ρ, the optimally weighted score equations are

$$\sum_{j=1}^{N} \frac{1}{n_j - 1} \frac{\partial p\ell_j^{(m)}}{\partial \pi} = 0 \tag{3.22}$$

and

$$\sum_{j=1}^{N} \frac{\partial p\ell_j^{(m)}}{\partial \rho} = 0, \tag{3.23}$$

where $p\ell_j^{(m)}$ is the log pairwise likelihood from the marginal model for cluster j. In other words, derivatives with respect to the marginal parameter π should be weighted but not those with respect to the association parameter ρ. Hence to solve this pair of equations, one should alternate between maximizing WPL with respect to π for a given ρ and maximizing PL with respect to ρ for a given π.

We shall now apply the argument of Kuk and Nott (2000) to the multilevel probit model of interest here. To this end, we consider a model with a single intercept term β (no covariates) and a random intercept with variance $\theta = \sigma_u^2$.

We first introduce some notation, by defining $\ell_{ii'j} = P(y_{ij}, y_{i'j})$ and letting

$$p\ell_j' = \left(\frac{\partial p\ell_j}{\partial \beta}, \frac{\partial p\ell_j}{\partial \theta} \right)^T$$

denote the vector of first derivatives of $p\ell_j$, and

$$p\ell_j'' = \begin{pmatrix} \frac{\partial^2 p\ell_j}{\partial \beta^2} & \frac{\partial^2 p\ell_j}{\partial \beta \partial \theta} \\ \frac{\partial^2 p\ell_j}{\partial \beta \partial \theta} & \frac{\partial^2 p\ell_j}{\partial \theta^2} \end{pmatrix}$$

denote the matrix of second derivatives.

The argument of Kuk and Nott (2000) proceeds by noting that the optimal weighting of $p\ell_j$ in the score equation is given by

$$\sum_{j=1}^{N} J_j K_j^{-1} p\ell'_j = 0, \tag{3.24}$$

where $J_j = E[-p\ell''_j]$ and $K_j = \text{cov}[p\ell_j] = E[p\ell'_j p\ell'^{\,T}_j]$. They go on by evaluating J_j and K_j under the simplifying assumption of independence. After some straightforward but tedious calculations, we obtain

$$J_j = C_2^{n_j} E[-\ell''_{12j}] = C_2^{n_j} E[\ell'_{12j}\ell'^{\,T}_{12j}] = C_2^{n_j}\,\zeta \begin{pmatrix} 2 & -\beta \\ -\beta & \frac{\beta^2}{2}+\zeta \end{pmatrix}, \tag{3.25}$$

where

$$\zeta = \frac{\phi^2(\beta)}{\Phi(\beta)\,[1-\Phi(\beta)]}.$$

Similarly, it can be shown that

$$K_j = \text{cov}[\sum_{i'>i} \ell'_{ii'j}] = C_2^{n_j}\text{cov}[\ell'_{12j}] + 2C_2^{n_j}\, n_j \text{cov}[\ell'_{12j}, \ell'_{13j}], \tag{3.26}$$

where we have used the fact that $\text{cov}[\ell'_{ikj}, \ell'_{i'k'j}] = 0$ under the independence assumption. Now $\text{cov}[\ell'_{12j}] = E[-\ell''_{12j}]$ is given above and

$$\text{cov}[\ell'_{12j}, \ell'_{13j}] = E[\ell'_{12j}\ell'^{\,T}_{13j}] = \begin{pmatrix} 1 & \frac{-\beta}{2} \\ \frac{-\beta}{2} & \frac{\beta^2}{4} \end{pmatrix}.$$

After substitution into (3.26) and some algebra, we find

$$K_j = (n_j-1)C_2^{n_j}\,\zeta \begin{pmatrix} 2 & -\beta \\ -\beta & \frac{\beta^2}{2}+\frac{\zeta}{n_j-1} \end{pmatrix}$$

and

$$J_j K_j^{-1} = \begin{pmatrix} \frac{1}{n_j-1} & 0 \\ \frac{(n_j-2)\beta}{n_j-1} & 1 \end{pmatrix}.$$

In comparison, Kuk and Nott (2000) found

$$J_j K_j^{-1} = \begin{pmatrix} \frac{1}{n_j-1} & 0 \\ 0 & 1 \end{pmatrix},$$

which led them to the pair of equations (3.22)–(3.23).

From the above argument we obtain a weighted equation for the fixed parameter but a weighted sum of pseudo-score functions for the random parameter. It can be noted, however, that the weights will not be strongly dependent on cluster size – at least when clusters are large since $(n_j-2)/(n_j-1) \simeq 1$ in this case – but the resulting equation will depend on the fixed parameters, which prevents us from drawing the same conclusions as Kuk and Nott (2000).

To supplement this result and gain further insight into the problem, we report some asymptotic efficiency calculations to compare the weighted and unweighted MPL estimators. We consider the same model as in the above discussion, with $\beta = 0, 1, 2, 3$ and $\theta = 0, 0.25, 0.5, 1, 2, 4$. Clusters are either of size 3, 4, or 5 and assumed each to occur with equal probability. For fixed cluster size, expression (3.21) can be used to calculate the asymptotic variance matrix of the MPL estimators and appropriate weighting can be introduced to calculate the asymptotic variance matrix of the MWPL estimators. The results for each estimator are then combined to calculate the required expectations.

Table 3.3 reports asymptotic efficiencies of the MPLE relative to the MWPLE for each combination of parameter values. When there is none or little intra-cluster correlation (i.e., θ is close to 0), the MPLE appears to be more efficient for the random parameter but for the fixed parameter the conclusion depends on its true assumed value. If the success probability evaluated at the mean of the random-effect distribution is large (i.e., β is small), the MWPLE is more efficient than the MPLE, whereas the reverse holds true for more extreme success probabilities. As the association among responses becomes stronger, the MWPLE tends to be more efficient than the MPLE for both parameters. In conclusion, none of the MPLE or MWPLE is uniformly more efficient than the other and losses of efficiency between the two estimators are modest (less than 10%). In view of Table 3.3, the MWPLE tends to perform slightly better than the MPLE overall, especially when $\theta > 0.5$, that is, for values of θ that we have found to be relatively common in practice.

Table 3.3. *Asymptotic efficiency of MPL versus MWPL in the random-intercept model. Cell entries are asymptotic relative efficiencies (percentages) for $\hat{\beta}$ (first row) and $\hat{\theta}$ (second row).*

	θ					
β	0	0.25	0.5	1	2	4
0	94.0	91.0	89.8	88.8	88.0	87.5
	107.8	100.1	97.4	95.1	93.4	92.3
1	97.5	94.3	92.7	88.8	89.7	87.5
	107.8	100.8	97.8	95.1	93.4	92.3
2	105.3	101.2	98.3	95.2	92.8	91.0
	107.8	102.9	99.2	95.7	93.5	92.3
3	107.6	105.2	101.6	97.4	94.4	92.4
	107.8	105.5	101.5	96.8	93.8	92.3

3.6 Example: a Meta-Analysis of Trials in Schizophrenic Subjects

To illustrate MPL estimation on a real set of data, we consider a group of five randomized clinical trials comparing the effect of risperidone to conventional antipsychotic agents (or placebo) for the treatment of chronic schizophrenia. Schizophrenia has long been recognized as a heterogeneous disorder with patients suffering from both "negative" and "positive" symptoms. Negative symptoms are characterized by deficits in social functions (poverty of speech, apathy, emotional withdrawal, etc.). Positive symptoms entail more florid symptoms such as delusions, hallucinations and disorganized thinking, which are superrimposed on the mental status (Kay, Fiszbein and Opler, 1987).

Several measures can be considered to assess a patient's global condition. The *Positive and Negative Syndrome Scale* (PANSS) (Kay, Fiszbein and Opler, 1987) consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia. Another useful scale is the Clinician's Global Impression (CGI), which is generally

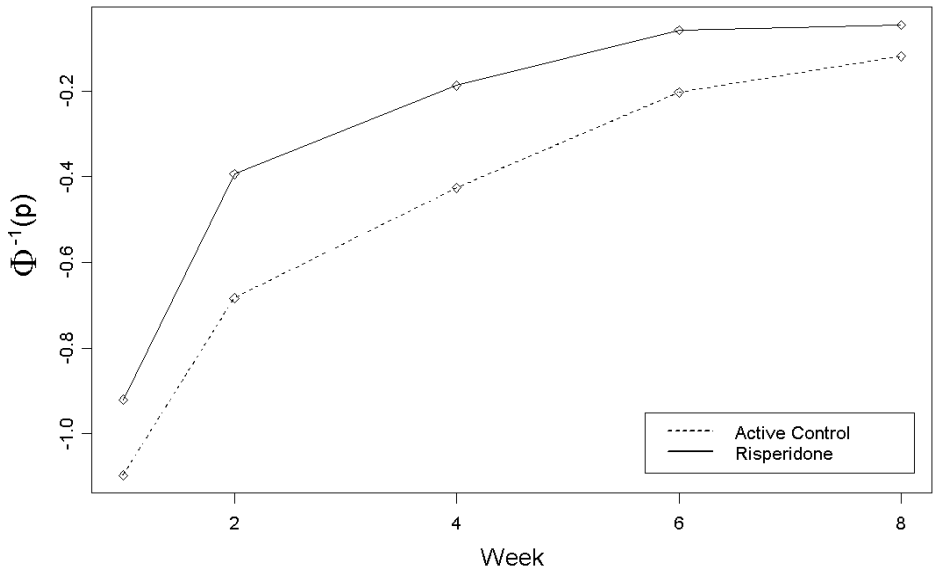


Figure 3.1. *Schizophrenia data: Proportion (on the probit scale) of patients who have experienced clinical improvement since baseline as a function of time.*

accepted as a subjective clinical measure of change. In the five trials in question, we have measures of the CGI overall change versus baseline. This is a 7-grade scale which ranges from 1='very much improved' to 7='very much worsened' and is used by the treating physician to assess a subject's overall clinical improvement compared to baseline.

Only subjects who received optimal doses of risperidone (4-6 mg/day) or an active control (haloperidol, perphenazine, zuclopenthixol) are included in the analysis, to provide a total number of 805 patients. We transform the CGI overall change versus baseline into a binary outcome by defining a success ($Y_{ij} = 1$) as clinical improvement since baseline (CGI grade of 1 or 2) and a failure otherwise.

Depending on the trial, treatment was administered for a period of 4 to 8 weeks and overall, we have scores for the CGI overall change versus baseline at weeks 1, 2, 4, 6 and 8. Figure 3.1 shows the proportion (on the probit scale) of patients who experienced clinical improvement since baseline as a function of time for each

Table 3.4. *Schizophrenia data: ML (with 20 quadrature nodes), MPL and MWPL parameter estimates and their estimated standard errors.*

	ML		MPL		MWPL	
Parameter	**Estimate**	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**
Fixed-Effects Structure:						
Intercept	-0.665	0.327	-0.504	0.333	-0.581	0.346
Week1	-1.833	0.375	-1.827	0.362	-1.796	0.392
Week2	-1.055	0.321	-1.062	0.316	-1.075	0.336
Week4	-0.725	0.268	-0.778	0.270	-0.753	0.284
Week6	-0.216	0.230	0.219	0.231	-0.238	0.239
Treat×Week1	0.343	0.248	0.250	0.253	0.308	0.243
Treat×Week2	0.708	0.241	0.571	0.250	0.670	0.242
Treat×Week4	0.814	0.281	0.689	0.291	0.743	0.288
Treat×Week6	0.641	0.358	0.477	0.371	0.574	0.378
Treat×Week8	0.575	0.457	0.384	0.468	0.470	0.485
Random-Effects Structure:						
σ_{11}	2.124	0.243	1.994	0.253	2.022	0.285
σ_{12}	-0.068	0.050	-0.040	0.056	-0.063	0.060
σ_{22}	0.491	0.064	0.484	0.060	0.526	0.063

Coding for Treat: 0 = active control, 1 = risperidone.

treatment group.

A saturated 'treatment by time' model, with a distinct parameter for each treatment and time combination, is considered for the proportion of patients who experienced clinical improvement since baseline. We further assume a random intercept and a random slope in the model. ML, MPL and MWPL parameter estimates along with their estimated standard errors are reported in Table 3.4. Note that for the random-effects structure, the parameters of the upper triangular Cholesky factor of $\Omega_u = \Sigma^T\Sigma$ are given. For the ML procedure, adaptive Gaussian quadrature was used to evaluate the likelihood. Different numbers ($q = 10, 20, 50$) of quadrature nodes were tested and while there were sizeable differences between $q = 10$ and $q = 20$ for

some of the parameters, estimates obtained with $q = 20$ and $q = 50$ did not differ much. These models were fitted using the SAS NLMIXED procedure (SAS Institute Inc., 2000) and the corresponding computing times on a Personal Computer (PC) with Pentium 450MHz processor were about 99 min for $q = 10$, 340 min for $q = 20$ and 1705 min for $q = 50$. The MPL and MWPL procedures, on the other hand, took about 4 min to run to completion, thus providing a substantial improvement.

Table 3.4 shows that parameter estimates do not differ much accross all three procedures, at least in view of the variability attached to these estimates. Interestingly, variance estimates, which were calculated based on (3.16), are smaller with MWPL than MPL here, although it can be observed that most of the parameter estimates are smaller in size as well.

3.7 Simulation Study

A simulation study was conducted to further evaluate the performances of the MPLE in finite samples and compare it with the ML and PQL2 estimators in relatively simple two-level models for binary responses. Of particular interest are the loss of efficiency induced by the MPLE and the overall bias/variance trade-off.

The underlying model from which data were simulated was taken to be:

$$\Phi^{-1}(P[y_{ij} = 1|\boldsymbol{u}_i]) = \beta_0 + \beta_1 x_{ij} + \overset{T}{z_{ij}}\boldsymbol{u}_i. \tag{3.27}$$

with $\beta_0 = 0.5$ and $\beta_1 = 1$. The results reported here are for a covariate x with values generated from a uniform distribution over $[-1, 1]$ for each observation in a cluster. The case of a cluster-specific covariate was also considered and conclusions are briefly discussed below. In model (3.27) we assume either a random intercept ($z_{ij} = 1$) or a random intercept and slope ($z_{ij} = (1, x_{ij})^T$). The random effects were generated from independent normal distributions with identical variances (0.5 or 1). To form the binary outcome y_{ij}, we further generated individual random components $N(0, 1)$ so that the binary outcome was set to 1 when the linear predictor plus the sum of the random effects and the individual random component was greater than 0. The number of clusters was fixed to either 20 or 50, and all clusters had sizes taken from a uniform distribution on $[10, 30]$. The results were based on 500 replicates under each scenario.

For each generated data set, parameters were estimated using ML, PQL2 and MPL procedures. For ML, adaptive Gaussian quadrature with 10 quadrature nodes was

utilized to evaluate the marginal likelihood function. The Double-Dogleg routine was employed to maximize the marginal likelihood function and the sample information matrix was estimated using numerical second-order derivatives (with forward difference approximation) upon convergence. To obtain PQL2 estimates, the approach of Goldstein and Rasbash (1996), and its implementation in MLwiN (Goldstein *et al.*, 1998), were followed. Finally, the MPL procedure was implemented as outlined in Section 3.3. Results reported hereafter are for the weighted estimator as it was found more efficient than the unweighted estimator under the chosen simulation settings. Note that for ML and MWPL, an absolute gradient convergence criterion of 0.001 was specified, whereas for PQL2 a relative parameter criterion of 0.001 was used.

Results for the random-intercept model are summarized in Figures 3.2 and 3.3. These figures present box and whisker plots of the simulated estimates for each parameter. The whiskers extend from the interquartile range (the box) as far as the data extend to a distance not exceeding 1.5 interquartile range. The thick mark in the box represents the median. Also included as horizontal lines are the mean ± 1 root mean square error (MSE). Note that in these figures and subsequent ones, we have occasionally removed a few extreme values from the plots (but not from the calculations of summary statistics) for better readibility.

In both figures estimates of the fixed parameters show no systematic bias, except for the slope parameter which seems to be slightly biased upward under PQL2. Estimates of σ^2_{u0} were located, on average, on the true value of the parameter and their distribution is, as can be expected, slightly skewed to the right, this being more pronounced with smaller sample sizes (Figure 3.2). When comparing MSEs of β_0 and β_1, there is little difference across all three estimation methods, with a slight advantage to ML. For σ^2_{u0}, MSEs were also smaller with ML when $\sigma^2_{u0} = 0.5$ but were, surprisingly, in favor of PQL2 when $\sigma^2_{u0} = 1$. When looking at the relative efficiency of the MWPL estimator to the ML estimator (which was estimated as the ratio of the variance of the ML parameter estimates over that of the MWPL estimates), efficiency losses were around 10% (ranging from 5 to 18%) for all parameters accross the four simulation settings.

We found the good performance of the PQL2 estimator when $\sigma^2_{u0} = 1$ quite surprising as this goes against the preconceived idea that PQL2 is biased downwards for large variance components. It should be stressed, however, that PQL2 has been studied mostly in models with logit link, which suggests that the link specification may be an important consideration when fitting multilevel models to binary data

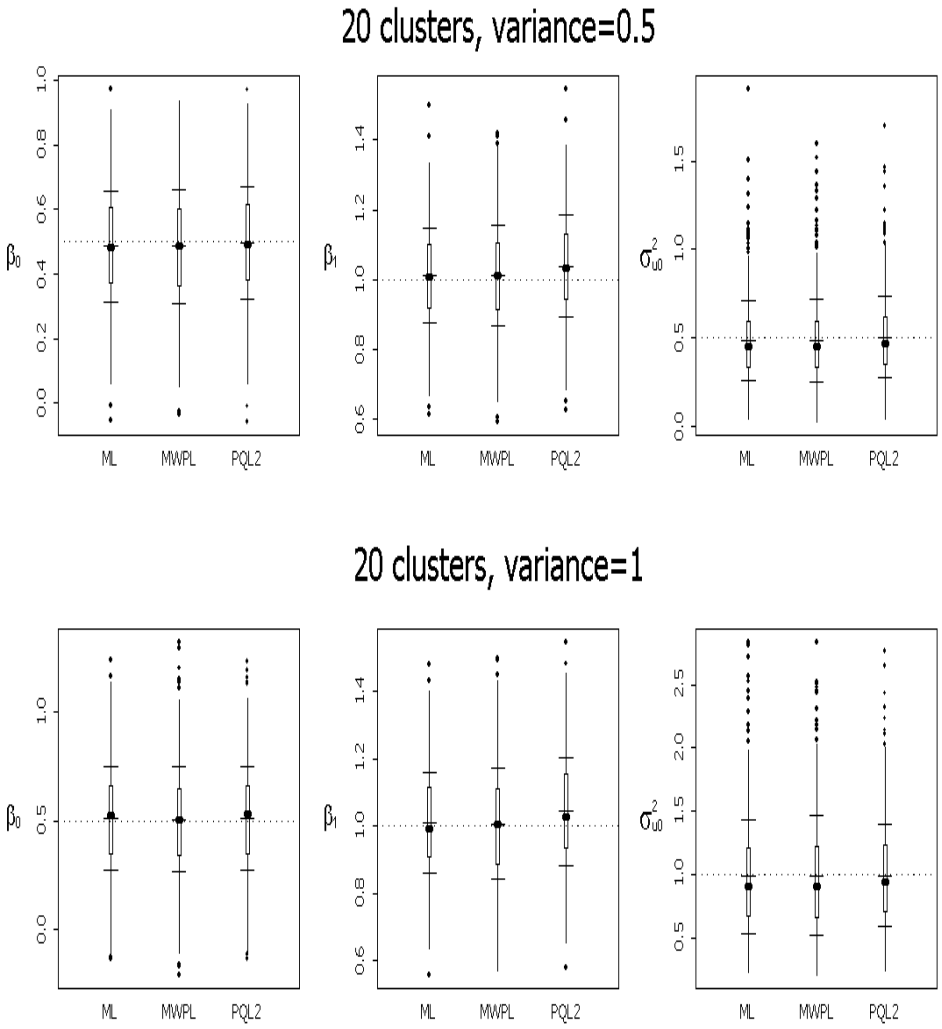


Figure 3.2. *Boxplots of ML, MWPL and PQL2 simulated parameter estimates under Model (3.27) with random intercept* $\sim N(0, \sigma^2_{u0})$*. Top panel: 20 clusters with* $\sigma^2_{u0} = 0.5$*; Bottom panel: 20 clusters with* $\sigma^2_{u0} = 1$*.*

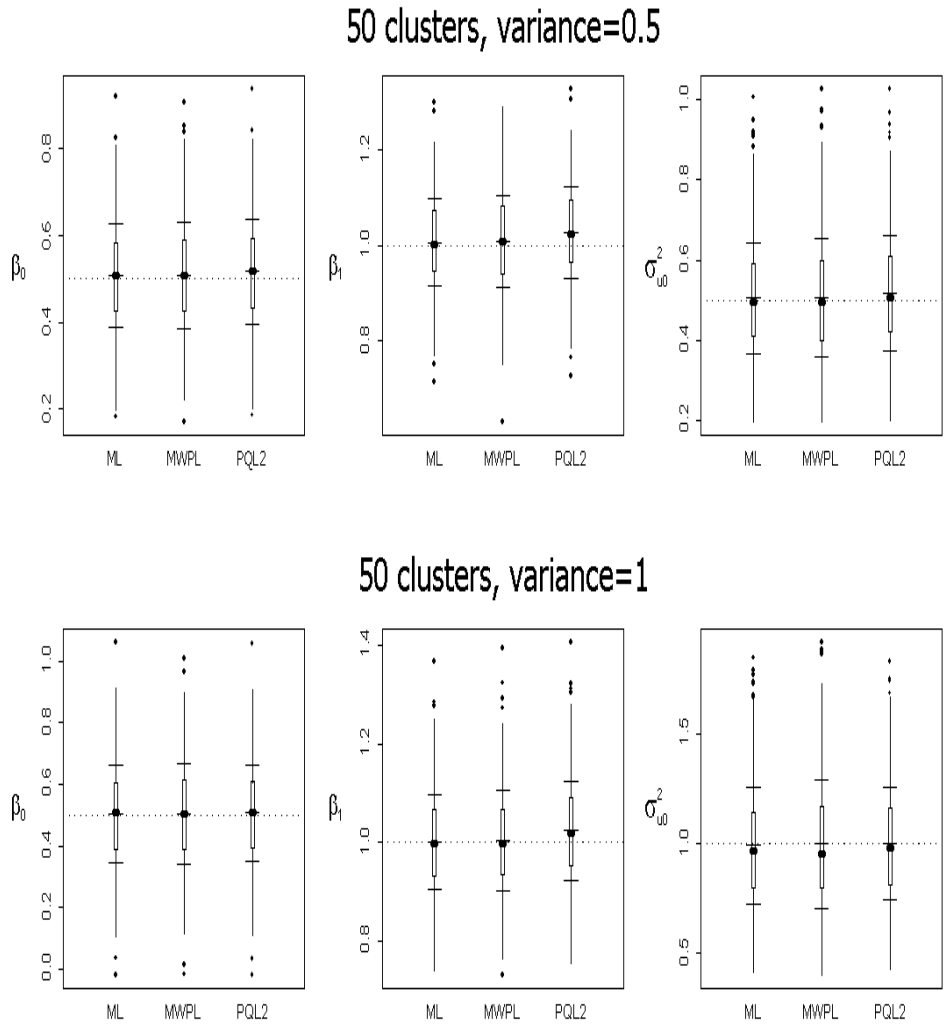


Figure 3.3. *Boxplots of ML, MWPL and PQL2 simulated parameter estimates under Model (3.27) with random intercept* $\sim N(0, \sigma^2_{u0})$*. Top panel: 50 clusters with* $\sigma^2_{u0} = 0.5$*; Bottom panel: 50 clusters with* $\sigma^2_{u0} = 1$*.*

Table 3.5. *Simulations results (1000 replicates) to compare PQL2 estimates in the random-intercept model with logit and probit link. Means are reported with Monte Carlo error given between parentheses.*

Parameter	Probit	Logit
β_0	0.506 (0.005)	0.499 (0.005)
β_1	1.028 (0.003)	1.008 (0.004)
σ^2_{u0}	0.991 (0.008)	0.963 (0.009)

using the PQL algorithm. To further investigate this hypothesis, 1000 data sets of 50 clusters were generated from the random-intercept model, with the value of the variance parameter equal to 1. This was done using logit and probit links and results are summarized in Table 3.5. Note that the simulation setting is comparable to the one used by Goldstein (1995, p. 100). Results for the logit link are indeed very close to those reported by Goldstein and show the downward bias present in σ^2_{u0}, with the values of the fixed parameters close to their true value. For the probit link the bias in the variance parameter is essentially eliminated, but there apparently is a bias transfer to the slope parameter, now with upward tendency. The reason behind this is not entirely clear. It can be recalled, however, that the PQL algorithm entails iterative fitting of linear models that assume normality of the level 1 residuals. The latter assumption seems, in fact, better satisfied with the probit link in the random-intercept model. This, in turn, might lead to improved estimation of the variance parameter. We will see below, however, that this is not always so.

Results for the random-intercept-and-slope model are summarized in Figures 3.4 and 3.5. The three estimation procedures yielded, on average, similar results for the fixed parameters β_0 and β_1. No clear bias tendency could be detected in PQL2 estimates of β_1 under this scenario. MSEs for β_0 were comparable under ML and PQL2 when $\sigma^2_{u0} = \sigma^2_{u1} = 0.5$ and were smaller under PQL2 when $\sigma^2_{u0} = \sigma^2_{u1} = 1$. Variance parameters exhibit substantially more variability here. In Figure 3.4, the distribution of ML and MWPL estimates for σ^2_{u0} and σ^2_{u1} is shifted towards larger values, whereas the opposite happens with PQL2. With larger sample sizes (Figure 3.5), ML and MWPL estimators were located, on average, on their true values but this was less clear with PQL2 (see bottom panel). Interestingly, though, MSEs of σ^2_{u0} and σ^2_{u1}

were lower with PQL2. The covariance parameter σ_{u01} was, on average, correctly estimated with all three estimation methods. Finally, if we look at the relative efficiency of MWPL to ML, the loss was typically less than 15% for the fixed-effects parameters and σ^2_{u0}, and about 40% for σ_{u01} and σ^2_{u1}.

When the covariate x was cluster-specific (i.e., taking the same value for all observations within a cluster), conclusions remained essentially unchanged in the random-intercept model. In the random-intercept-and-slope model, convergence was an issue with all three procedures. With the smaller sample sizes (20 clusters), convergence rates were about 60% with ML and MWPL and dropped to less than 30% with PQL2. Fixed-effects parameters were reasonably well estimated but variance parameters were relatively biased (downwards for σ^2_{u0} and upwards for σ^2_{u1}) and their distribution markedly skewed with all three procedures. With larger sample sizes, convergence rates were about 80% with ML and MWPL and about 50% with PQL2. The bias in the variance parameters was still present albeit to a lesser extent. The efficiency loss of the MWPL estimator compared to the ML estimator was moderate ($< 20\%$) in both settings.

To further evaluate the performance of MWPL in data sets with clusters of smaller size, an additional set of simulations was performed. The intent was to match a longitudinal study design, so these simulations were based on the following model:

$$\Phi^{-1}(P[y_{ij} = 1|\boldsymbol{u}_i]) = \beta_0 + \beta_1 t_{ij} + z_{ij}^T \boldsymbol{u}_i. \tag{3.28}$$

with $\beta_0 = -1$, $\beta_1 = 0.5$ and $t_{ij} = 0, \ldots, 4$. The random effects had variance equal to 1. The number of subjects was taken to be 100, each subject having between 1 and 5 observations (randomly determined).

Results for the random-intercept model are summarized in Figure 3.6. ML and MWPL procedures behave comparably, with a small loss of efficiency for MWPL compared to ML ($< 10\%$). Interestingly, PQL2 seemed to perform worse under this scenario: all parameters appear to be biased and to exhibit more variability.

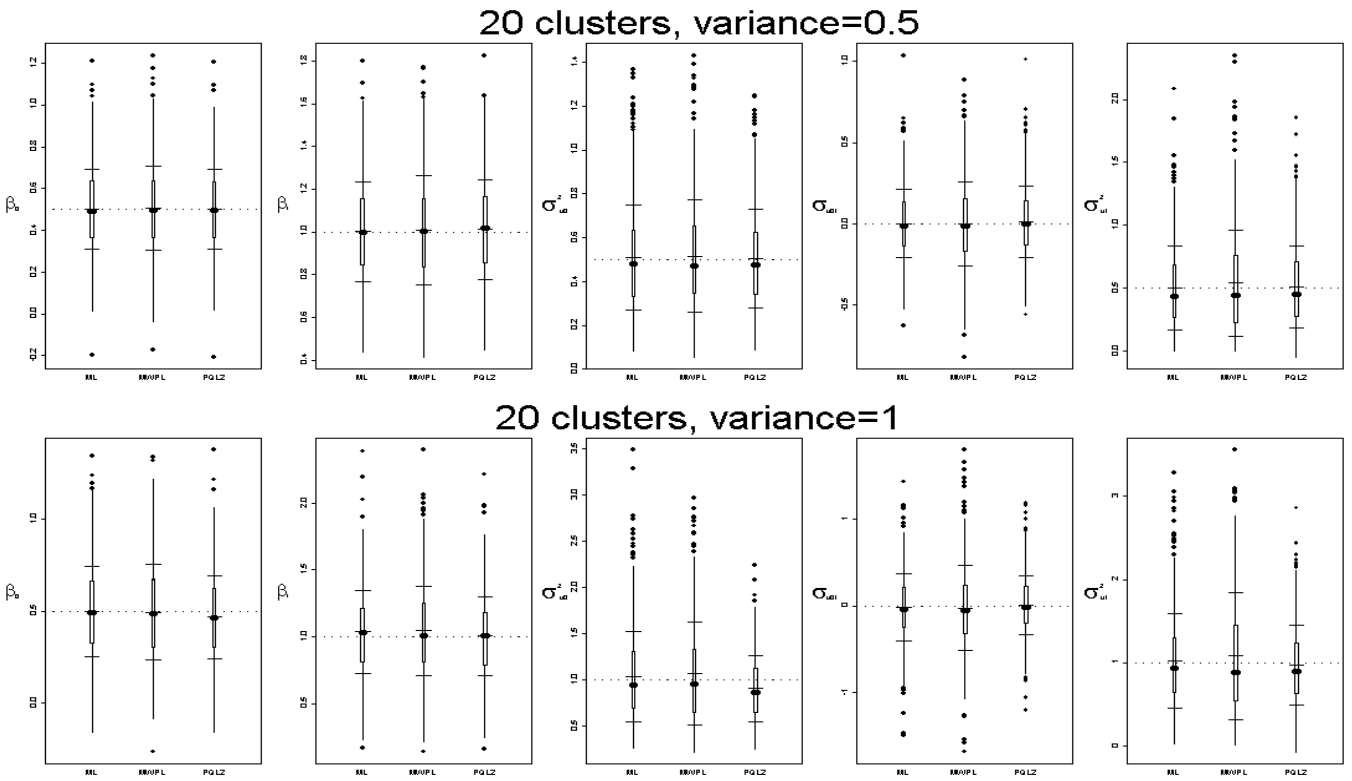


Figure 3.4. *Boxplots of ML, MWPL and PQL2 simulated parameter estimates under Model (3.27) with random intercept and random slope* $\sim N(0, \Omega_u)$*. Top panel: 20 clusters with* $\sigma^2_{u0} = 0.5 = \sigma^2_{u1}$ *and* $\sigma_{u01} = 0$*; Bottom panel: 20 clusters with* $\sigma^2_{u0} = 1 = \sigma^2_{u1}$ *and* $\sigma_{u01} = 0$*.*

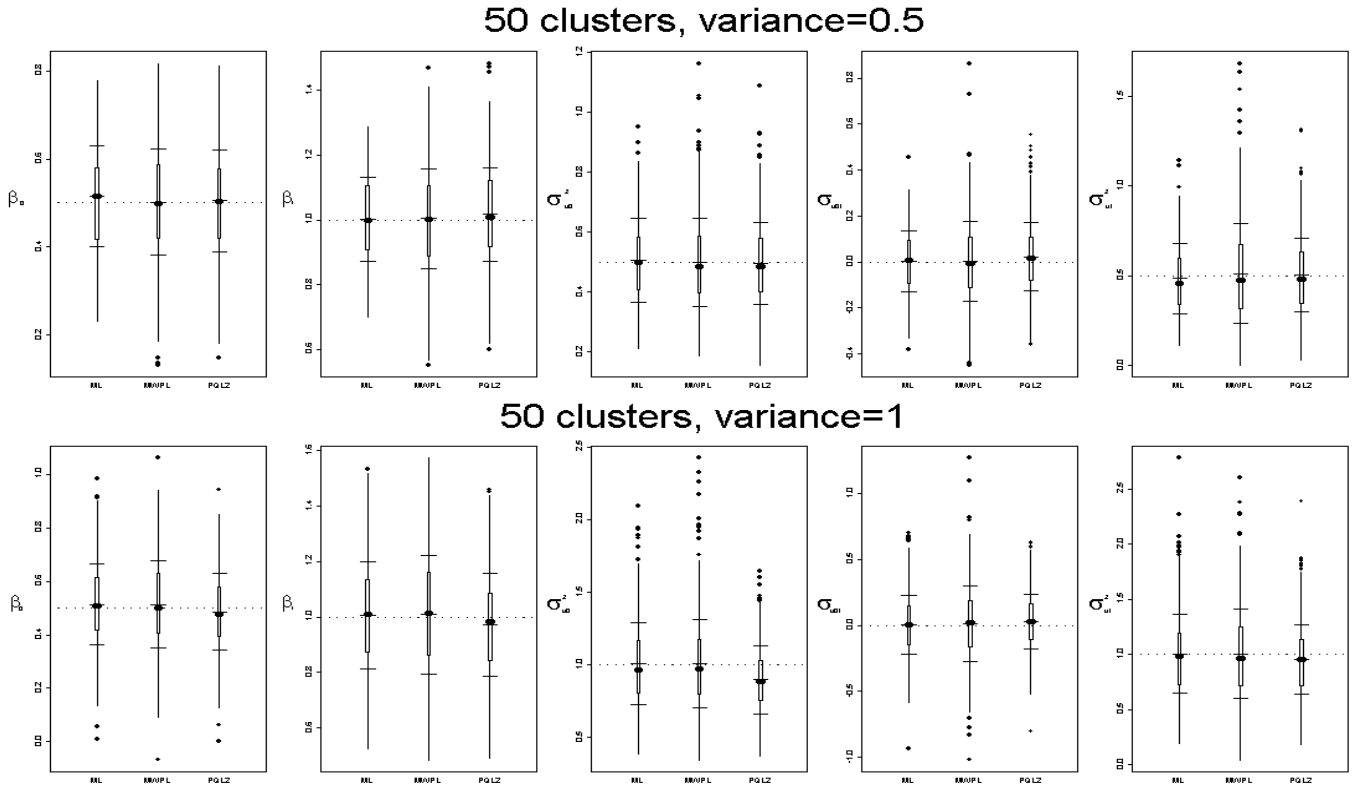


Figure 3.5. *Boxplots of ML, MWPL and PQL2 simulated parameter estimates under Model (3.27) with random intercept and random slope* $\sim N(0, \Omega_u)$*. Top panel: 50 clusters with* $\sigma^2_{u0} = 0.5 = \sigma^2_{u1}$ *and* $\sigma_{u01} = 0$*; Bottom panel: 50 clusters with* $\sigma^2_{u0} = 1 = \sigma^2_{u1}$ *and* $\sigma_{u01} = 0$*.*

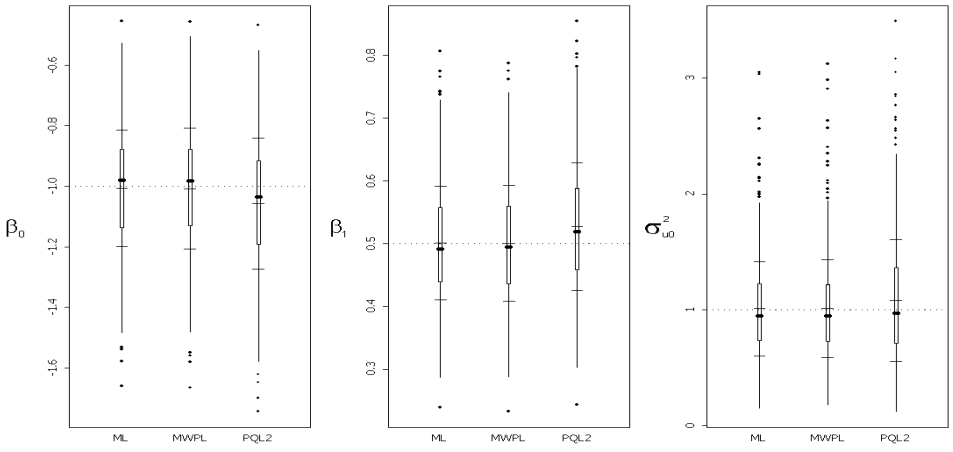


Figure 3.6. *Boxplots of ML, MWPL and PQL2 simulated parameter estimates under Model (3.28) with random intercept.*

For the random-intercept-and-slope model (Figure 3.7), convergence difficulties were more frequent. With PQL2, a solution could be obtained in approximately 75% of the cases. With ML and MWPL procedures, convergence was easier to achieve ($\approx$ 95% of the cases) although some solutions had unusually large values for some of the parameters, especially (co)variance parameters. After deleting these cases, we were left with about 90% of the observations. Note that the convergence criterion was lowered in order to detect solutions converging towards a non positive-definite covariance matrix Ω_u, which partly caused the aforementioned problem. Figure 3.7 shows that ML and MWPL behave comparably. The efficiency loss of MWPL over ML was about 15% for fixed parameters, about 30% for the parameters σ^2_{u0} and σ_{u01}, whereas it was negligeable for σ^2_{u1}. This is in line with the asymptotic results reported in Section 3.4 (see Table 3.2) where a similar setting was considered. Finally, even though PQL2 exhibits less variability, it seems to fail quite dramatically in this setting and variance parameters are seriously biased downwards.

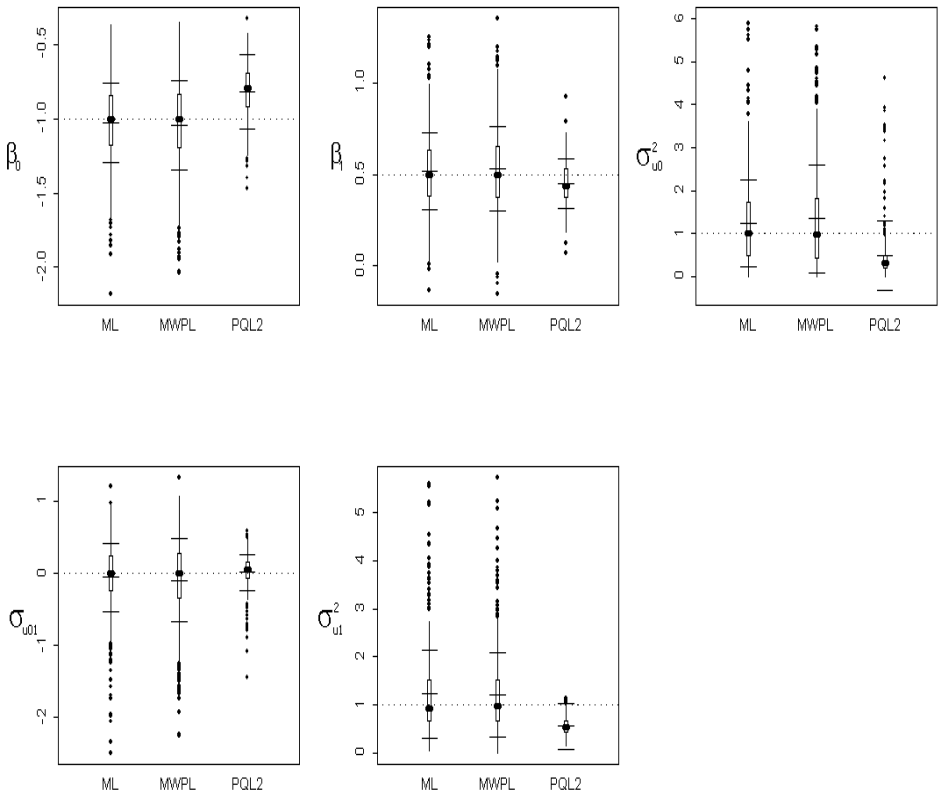


Figure 3.7. *Boxplots of ML, MWPL and PQL2 simulated parameter estimates under Model (3.28) with random slope.*

3.8 Discussion

To conclude, we attempt to better delineate advantages and disadvantages of fitting multilevel models with binary responses using MPL for estimation purposes.

One of the advantages of MPL over ML stands at the computational level since the PL function involves evaluation of univariate and bivariate probits only, regardless of

Table 3.6. *Median computing times (seconds) for fitting random-effects model (3.27) to 100 simulated data sets using ML and MPL.*

	Dimension of random effects		
	1	**2**	**3**
ML	3.0	57.5	408.2
MPL	86.4	116.0	185.8

Simulation settings:
- clusters of size between 10 and 30
- random-effects variances equal to 1

the number of specified random effects. Obviously, its use will not be very appealing when the number of random effects is small since ML estimation could be effectively used instead. It should be recalled that the computational cost to evaluate the marginal likelihood using a quadrature-based method increases exponentially with the dimension of integration. With MPL, on the other hand, the complexity increases roughly as a quadratic function of cluster size.

To further illustrate this point, we report in Table 3.6 the median computing times for fitting model (3.27) to 100 simulated data sets for random effects of dimension 1 to 3. A quadratic term x_{ij}^2 was added to get random effects of dimension 3. As can be seen, MPL is clearly not advantageous with random effects of dimension less than 3, at least in data sets with similarly large cluster sizes, but it becomes attractive with three-dimensional random effects.

Another facet of the problem was illustrated by the analysis of the schizophrenia data in Section 3.6 which showed how efficient MPL can be from a computational standpoint when clusters are of small size, even in a situation where only two random effects are specified. Based on these considerations, we can tentatively recommend the use of MPL estimation in multilevel probit models when:

- clusters are of small sizes, such as typically is the case in many longitudinal studies;
- and/or the random structure of the model to be fitted is complex.

An additional advantage of MPL is its numerical stability since it does not involve

possibly unstable operations like matrix inversions. Convergence rates for MPL were found comparable to those for ML in our simulation study. Approximate methods such as PQL, on the other hand, can sustain a great deal of numerical problems and from our personal experience, it is not uncommon that the algorithm fails to converge in practical applications, the problem being worsened as more complex models are fitted. It was clearly seen from our simulation study that the algorithm can behave well in some settings and fail miserably in others. Furthermore, the PQL2 algorithm appears to be sensitive to extreme success probabilities (i.e. ≈ 0 or ≈ 1). Thus, convergence rates in the first set of simulations were high because values chosen for the fixed parameters correspond to success probabilities in the mid-range (centered around 0.6) of the unit interval. However, they drop off significantly as soon as one departs from this setting. For instance, a value of $\beta_0 = -2$ in the random-intercept model (3.27) gives convergence rates of about 90% when $\sigma^2_{u0} = 0.5$ and about 60% when $\sigma^2_{u0} = 1$. This is even worse in the random-intercept-and-slope model, where our implementation of the procedure becomes highly unstable with convergence rates as low as 10%. If we add to this the fact that the PQL algorithm produces biased estimates and that the bias is dependent on the link specification and affects more or less severely different parameters, it becomes difficult to make definite recommendations about the use of this algorithm in general. It should be noted, however, that PQL runs much faster than either PL and ML so it might happen to be the only procedure worth of consideration.

While MPL estimation was advocated on the basis of its computational ease, this comes with a price, namely, loss of efficiency. The asymptotic calculations of Section 3.4 and our simulation study showed a generally moderate loss (less than 20%), although it can attain up to 35% for some parameters.

Another drawback, clearly, is that MPL will not be applicable to any hierarchically structured data set and practical limitations on the number of levels will arise. For instance, in a three-level model all possible pairs within and between level 2 units pertaining to the same level 3 unit should be considered. This will become computationally prohibitive as the number of levels and the cluster size increase. Note that an application of MPL estimation in a three-level model will be presented in the next chapter.

An interesting practical situation that can reduce the computational burden of the procedure is when covariates are discrete. In this case it suffices to consider solely contributions of (within-cluster) pairs of observations with different covariate

values, while keeping track of the multiplicities for each combination. If this is coupled with the ordinary (unweighted) MPL estimation procedure, then only observations with different covariate values across the whole data set need to be kept, thereby making evaluation of the PL function inexpensive. It should be observed that cluster membership information is still required to calculate the covariance matrix of the parameter estimates *via* the sandwich estimator.

In this chapter we have focused exclusively on models dealing with a probit link specification. Even though a logit link would be straightforward to specify by assuming a standard logistic rather than normal distribution for $\tilde{\epsilon}_{ij}$ in (3.18), this yields intractable integrals for pairwise probabilities and thus, MPL estimation cannot be applied in a beneficial manner. As discussed in Section 3.3, we do not see this as a strong limitation. One could also view the latent variable assumption as a restriction of the present approach. While this assumption may be sensible in many applications, it will not always be so. Moreover, the existence of the latent variable is usually unverifiable in practice. We emphasize, however, that this assumption was made only to facilitate mathematical developments (through probabilistic identities) and we do not necessarily need to believe that the threshold model holds.

Finally, when the response variable is measured on an ordinal rather than binary scale, Hedeker and Gibbons (1994) show how model (3.17) can be extended. In this case a number of thresholds or 'cut-off' points must be considered to define the different response categories, which then become additional parameters to be estimated in the model. The MPL approach can be easily extended to deal with this kind of situations.

Chapter 4

Validation of Surrogate Endpoints in Multiple Randomized Clinical Trials with Discrete Outcomes

4.1 Introduction

The evaluation of a treatment (Z) is based on the observation of a clinically meaningful endpoint which is referred to as the "true" or "final" endpoint (T). Often the true endpoint upon which treatment benefits will ultimately be assessed is distant in time or measured at high expense, making it worthwhile to consider an intermediate or surrogate endpoint (S) that can be measured earlier, more conveniently, or more frequently than the endpoint of interest.

The validation of surrogate endpoints in clinical trials is a controversial issue (Ellenberg and Hamilton, 1989; Boissel *et al.*, 1992; Lagakos and Hoth, 1992; Fleming *et al.*, 1994; Fleming and DeMets, 1996; De Gruttola *et al.*, 1997; Chuang-Stein and DeMasi, 1998) and should be rigorously established. In a landmark paper, Prentice (1989) proposed a definition as well as a set of operational criteria to validate surrogate endpoints, but they are equivalent solely if the surrogate and true endpoints are binary

(Buyse and Molenberghs, 1998). Freedman, Graubard, and Schatzkin (1992) supplemented these criteria with the so-called *proportion of treatment explained* (PTE), which quantifies the proportion of treatment effect on the true endpoint that is mediated through the surrogate endpoint. This quantity has some drawbacks, however. First, it is not a genuine proportion in the strict sense and can take values on the whole real line. Also, confidence limits for this quantity tend to be wide in general, unless the sample size is large. Flandre and Saidi (1999) and Molenberghs *et al.* (2002) further discuss difficulties associated with PTE.

Buyse and Molenberghs (1998) suggest to replace the proportion explained by two quantities: the *relative effect*, linking the effects of treatment on both endpoints at the population level, and the *adjusted association*, an individual-level measure of agreement between the two endpoints after accounting for the effect of treatment. They focused on the case where both the surrogate and true endpoints are either binary or normally distributed. Technically, a joint model for the two endpoints is required. The relative effect is defined as $RE = \beta/\alpha$, where α and β denote the effects of Z on S and T respectively. For normally distributed endpoints the adjusted association γ_z is the correlation between S and T after correcting for treatment, whereas for binary endpoints γ_z can take the form of a log odds ratio for example. Buyse and Molenberghs (1998) show that when the true and surrogate endpoints are normally distributed, PTE is essentially the product of the relative effect and the adjusted association. This suggests that PTE is, in effect, a composite quantity, a mixture of two aspects of the model: the fixed effects (population-averaged level) and the random component (individual level).

In order to be informative and of practical value, the validation of a surrogate endpoint will typically require a large number of observations. It is therefore useful to consider situations where data are available from multiple randomized experiments. Buyse *et al.* (2000) show how the relative effect and the adjusted association can be extended in the presence of multiple grouping units (e.g. trials in a meta-analytic setting). With this approach the surrogate endpoint validation issue is examined at each of the two levels of interest, that is, at the trial level and at the individual level.

Whereas the paper of Buyse *et al.* (2000) treats the methodologically appealing case of two normally distributed endpoints, situations abound in practice where the endpoints are of a different nature, such as failure-time or binary outcomes. The question then arises as to which model should be used for specifying the joint distribution of the surrogate and true endpoints. The answer is rarely trivial, mainly because of

the absence of flexible tools like the multivariate normal distribution and the linear mixed model. For example, when both endpoints are time-to-event variables, an extension of the method can be based on the use of copula models (Burzykowski *et al.*, 2001). Another interesting situation is when the endpoints are of a mixed continuous and discrete nature, in which case a latent variable approach based on a probit-linear or Plackett-Dale model can be used (Molenberghs, Geys and Buyse, 2001). In the present chapter we describe an extension of the method to deal with binary outcomes. Most of the results presented here can be found in Renard *et al.* (2002a).

In Section 4.2 we describe the meta-analytic approach to surrogate endpoint validation proposed by Buyse *et al.* (2000). Computational difficulties, causing frequent failures of the algorithm to converge, are also discussed. We show, in Section 4.3, how the model can be extended to the case of two binary endpoints. The task amounts to fit a three-level model with a four-dimensional random-effects structure, which we propose to do using the maximum pairwise likelihood (MPL) estimation method described in the previous chapter. Some simulations are reported in Section 4.4 to assess the impact of such factors as the number of trials and the trial size on the two key surrogacy measures used in the validation process. Finally, in Section 4.5 the method is illustrated on the schizophrenia data introduced in the previous chapter (Section 3.6).

4.2 Surrogate Endpoint Validation: Two Normally Distributed Endpoints

We first describe the meta-analytic approach to surrogate endpoint validation developed by Buyse *et al.* (2000) in the case of two normally distributed endpoints and then turn to computational difficulties associated with the random-effects modeling methodology.

4.2.1 A Hierarchical Model

Two distinct modeling strategies can be followed, based on a two-stage fixed effects representation on the one hand and random effects on the other hand.

We start by describing the two-stage model. The first stage is based upon a joint

regression model for S and T:

$$\begin{cases} S_{ij} &= \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} &= \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \end{cases} \tag{4.1}$$

where the indices i and j refer to trials and subjects within trials respectively; μ_{Si} and μ_{Ti} are trial-specific intercepts; and α_i and β_i are the trial-specific effects of Z on the two endpoints in trial $i = 1, \ldots, N$. Finally, ε_{Sij} and ε_{Tij} are correlated error terms, assumed to be normally distributed with mean zero and covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}. \tag{4.2}$$

At the second stage, we assume that

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{4.3}$$

where the second term on the right-hand side is assumed to follow a zero-mean normal distribution with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Tb} & d_{ab} & d_{bb} \end{pmatrix}. \tag{4.4}$$

The random-effects representation is obtained by combining the two steps above:

$$\begin{cases} S_{ij} &= \mu_S + m_{Si} + (\alpha + a_i) Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} &= \mu_T + m_{Ti} + (\beta + b_i) Z_{ij} + \varepsilon_{Tij}. \end{cases} \tag{4.5}$$

4.2.2 Trial-Level Surrogacy

Since both the individual- and trial-level associations are of interest, the surrogate endpoint validation issue is examined at each of these levels. A key motivation for validating a surrogate endpoint is to be able to predict the effect of treatment on the true endpoint, based on the observed effect of treatment on the surrogate endpoint. It is therefore essential to explore the quality of the prediction of the treatment effect on

the true endpoint by (a) information obtained in the validation process based on trials $i = 1, \ldots, N$ and (b) information available on the surrogate endpoint in a new trial, $i = 0$ say. Fitting either the fixed-effects model (4.1) or the mixed-effects model (4.5) to data from a meta-analysis provides estimates for the parameters and the variance components. Suppose then that a new trial $i = 0$ is considered for which data are available on the surrogate endpoint but not on the true endpoint. We then fit the following linear model to the surrogate outcomes S_{0j}:

$$S_{0j} = \mu_{s0} + \alpha_0 Z_{0j} + \varepsilon_{s0j}.$$

Estimates for m_{s0} and a_0 are

$$\begin{aligned} \widehat{m}_{s0} &= \hat{\mu}_{s0} - \hat{\mu}_s, \\ \hat{a}_0 &= \hat{\alpha}_0 - \hat{\alpha}. \end{aligned}$$

We are interested in the estimated effect of Z on T, given the effect of Z on S. To this end, observe that $(\beta + b_0 | m_{s0}, a_0)$ follows a normal distribution with mean and variance:

$$\begin{aligned} E(\beta + b_0 | m_{s0}, a_0) &= \beta + \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{s0} - \mu_s \\ \alpha_0 - \alpha \end{pmatrix}, \\ \mathrm{var}(\beta + b_0 | m_{s0}, a_0) &= d_{bb} - \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}. \end{aligned}$$

This suggests to call a surrogate 'perfect at the trial level' if the conditional variance given by the last expression is equal to zero. A measure to assess the quality of the surrogate at the trial level is therefore given by the coefficient of determination

$$R^2_{\text{trial}} = R^2_{b_i | m_{Si}, a_i} = \frac{\begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{4.6}$$

This coefficient is unitless and ranges in the unit interval if the corresponding variance-covariance matrix D is positive-definite, two desirable features for its interpretation.

Intuition can be gained by considering the special case where the prediction of b_0 can be done independently of the random intercept m_{s0}. The above expressions then

reduce to

$$\begin{aligned} E(\beta + b_0|a_0) &= \beta + \frac{d_{ab}}{d_{aa}}(\alpha_0 - \alpha), \\ \text{var}(\beta + b_0|a_0) &= d_{bb} - \frac{d_{ab}^2}{d_{aa}}, \end{aligned}$$

and we have

$$R^2_{\text{trial}} = R^2_{b_i|a_i} = \frac{d_{ab}^2}{d_{aa}d_{bb}}.$$

This implies that $R^2_{\text{trial}} = 1$ if the trial-level treatment effects are multiples of each other.

So far we have examined the surrogate endpoint validation issue from a meta-analytic standpoint, that is, it was implicitly assumed that grouping units are randomized clinical trials. This needs not always be so and grouping units can actually represent any relevant experimental unit such as center, investigator or country for example. In the sequel, however, we shall continue to refer to the corresponding R^2 surrogacy coefficient as a "trial"-level measure (R^2_{trial}).

4.2.3 Individual-Level Surrogacy

At the individual level, we consider the association between the surrogate and the final endpoints after adjustment for the treatment effect. To this end, we need the conditional distribution of T, given S and Z. From (4.1) we derive

$$\begin{aligned} T_{ij}|S_{ij} \sim\ & N\left\{\mu_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}\mu_{Si} + (\beta_i - \sigma_{TS}\sigma_{SS}^{-1}\alpha_i)Z_{ij} + \sigma_{TS}\sigma_{SS}^{-1}S_{ij}\ ;\right. \\ & \left.\sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}\right\}. \end{aligned}$$

Similarly, the random-effects model (4.5) yields

$$\begin{aligned} T_{ij}|S_{ij} \sim\ & N\left\{\mu_T + m_{Ti} - \sigma_{TS}\sigma_{SS}^{-1}(\mu_S + m_{Si}) + [\beta + b_i - \sigma_{TS}\sigma_{SS}^{-1}(\alpha + a_i)]Z_{ij}\right. \\ & \left.+\sigma_{TS}\sigma_{SS}^{-1}S_{ij};\ \sigma_{TT} - \sigma_{TS}^2\sigma_{SS}^{-1}\right\}, \end{aligned}$$

where conditioning is also on the random effects. It follows that the association between both endpoints after adjustment for the treatment effect is captured by

$$R^2_{\text{indiv}} = R^2_{\varepsilon_{Tij}|\varepsilon_{Sij}} = \frac{\sigma_{ST}^2}{\sigma_{SS}\sigma_{TT}}, \tag{4.7}$$

the squared correlation between S and T after adjustment for the trial and treatment effects.

4.2.4 Surrogate Evaluation

A surrogate endpoint will be termed 'valid' if it is both trial-level valid ($R^2_{\text{trial}} \approx 1$) and individual-level valid ($R^2_{\text{indiv}} \approx 1$). Guidelines about how close R^2_{trial} and R^2_{indiv} have to be to 1 are hard to formulate in full generality. This will be based, preferably, upon expert opinion and confidence limits for these coefficients should be examined.

To be useful in practice, a surrogate must be able to predict the effect of treatment upon the true endpoint with sufficient precision to safely distinguish between effects that are clinically worthwhile from effects that are not. This requires both that the estimate of $\beta + b_0$ be sufficiently large and that the prediction interval of this quantity be sufficiently narrow.

It should be noted that the validation criteria proposed here do not require the treatment to have a significant effect on either endpoint. In particular, it is possible to have $\alpha \equiv 0$ and yet have a perfect surrogate. Indeed, even though the treatment may not have any effect on the surrogate endpoint as a whole, the fluctuations around zero in individual trials (or other experimental units) can be very strongly predictive of the effect on the true endpoint. However such a situation is unlikely to occur since the heterogeneity between the trials is generally small compared to that between individual patients.

4.2.5 Computational Issues

In the remainder of this chapter we shall focus on the random-effects modeling strategy. Model (4.5) can be regarded as a two-level multivariate or, more precisely, as a three-level model with a four-dimensional random structure at the trial (i.e. third) level. This model can be quite challenging to fit in practice. In particular, we have repeatedly observed convergence failures of the Newton-Raphson algorithm used to obtain ML estimates in real data sets. Therefore, it is worth investigating which features of the problem may facilitate convergence of the algorithm for this kind of models.

Several factors were explored: number of trials, size of the between-trial variability (compared to residual variability), number of patients per trial, normality assumption, and strength of the correlation between random effects. Among those, number of trials and between-trial variability were found to have most impact on the convergence properties of the Newton-Raphson algorithm used to maximize the likelihood. Table 4.1 shows the number of runs for which convergence could be achieved within

Table 4.1. *Number of runs (over 500) for which convergence was achieved within 20 iterations. Percentages are given in parentheses.*

	Number of trials		
$\boldsymbol{\sigma^2}$	**50**	**20**	**10**
1	500 (100.0)	498 (99.6)	412 (82.4)
0.1	491 (98.2)	417 (83.4)	218 (43.6)

20 iterations. In each case, 500 runs were performed, assuming the following model:

$$\begin{cases} S_{ij} &= 45 + m_{Si} + (3 + a_i)Z_{ij} + \varepsilon_{Sij}, \\ T_{ij} &= 50 + m_{Ti} + (5 + b_i)Z_{ij} + \varepsilon_{Tij}, \end{cases}$$

where $(m_{Si}, m_{Ti}, a_i, b_i) \sim N(0, D)$ with

$$D = \sigma^2 \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0.8 & 1 \end{pmatrix}$$

and $(\varepsilon_{Sij}, \varepsilon_{Tij}) \sim N(0, \Sigma)$ with

$$\Sigma = 3 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

The number of trials was fixed at either 10, 20 or 50, each trial involving 10 subjects randomly assigned to treatment groups. The σ^2 parameter was set to 0·1 or 1.

From Table 4.1, we see that when between-trial variability is large ($\sigma^2 = 1$), little convergence problems occur, except when the number of trials is strongly reduced. When between-trial variability is small ($\sigma^2 = 0.1$), convergence failures are more frequent and the situation deteriorates as the number of trials decreases.

These simulation results indicate that there should be enough variability at the trial level, and a sufficient number of trials, to obtain convergence of the Newton-Raphson algorithm used to obtain parameter estimates from model (4.5). When these requirements are not fulfilled, one must rely on simpler models, such as the two-stage model (4.1) for example.

4.3 Surrogate Endpoint Validation: Two Binary Outcomes

4.3.1 The Model

In order to extend the methodology to the case of two binary endpoints, we assume a latent variable formulation as in Chapter 3. That is, we posit the existence of a pair of continuously distributed latent variables $(\tilde{S}_{ij}, \tilde{T}_{ij})$ yielding the actual binary outcomes (S_{ij}, T_{ij}) after dichotomization. These unobservable variables are assumed to have a joint normal distribution and the realized value of S_{ij} (resp. T_{ij}) equals 1 if $\tilde{S}_{ij} > 0$ (resp. $\tilde{T}_{ij} > 0$), and 0 otherwise.

We are now in a suitable position to follow the modeling strategy outlined in the previous section. Consider the random-effects model

$$\begin{cases} \tilde{S}_{ij} &= \mu_S + m_{Si} + (\alpha + a_i)Z_{ij} + \tilde{\varepsilon}_{Sij}, \\ \tilde{T}_{ij} &= \mu_T + m_{Ti} + (\beta + b_i)Z_{ij} + \tilde{\varepsilon}_{Tij}, \end{cases} \tag{4.8}$$

which is model (4.5) but on the latent variable scale. This yields the following model for the observed binary outcomes:

$$\begin{cases} \Phi^{-1}(P[S_{ij} = 1 | m_{Si}, a_i, m_{Ti}, b_i]) &= \mu_S + m_{Si} + (\alpha + a_i)Z_{ij}, \\ \Phi^{-1}(P[T_{ij} = 1 | m_{Si}, a_i, m_{Ti}, b_i]) &= \mu_T + m_{Ti} + (\beta + b_i)Z_{ij}. \end{cases} \tag{4.9}$$

As discussed in Section 3.3, not all parameters in the model are identifiable and we constrain σ_{SS} and σ_{TT} to be equal to 1. The $\mathbf{\Sigma}$ matrix defined in (4.2) can therefore be replaced by

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{ST} \\ \rho_{ST} & 1 \end{pmatrix}. \tag{4.10}$$

The above formulation is particularly attractive since the coefficients of determination defined in the previous section can readily be employed without any modification, although at the individual level the interpretation of R^2_{indiv} is bound, formally, to the postulated latent variables that give rise to the observed binary responses.

4.3.2 Model Estimation

A crucial step with the proposed methodology is to fit the above model. A direct likelihood-based approach is unlikely to be satisfactory. Indeed, the four-dimensional

random structure of the model signifies use of Monte-Carlo methods to accomplish numerical integration and these are computationally involved. Since the framework is meta-analytical in nature, which means that very large data sets can be anticipated in practical applications, the computational burden may well be unbearable.

Approximate methods such as PQL (see Section 2.4) could *a priori* be good contenders since they are computationally efficient but we will not retain this approach, for two reasons. Firstly, the two measures of surrogacy are defined in terms of the random components of the model (matrices D and Σ) but the PQL algorithm is known to provide biased estimates for these components, which is not desirable. Secondly, this algorithm tends to be numerically unreliable and the problem is aggravated with complex models such as that we wish to use.

As a consequence, we propose to fit model (4.9) using the maximum pairwise likelihood (MPL) approach studied in the preceding chapter. Since we are dealing with a three-level model, we have to consider distinct contributions of pairwise likelihoods, reflecting different types of association, as illustrated in Figure 4.1:

(i) the association between the surrogate and true endpoints measured on the same individual;

(ii) the association between the surrogate endpoints measured on two distinct individuals;

(iii) the association between the true endpoints measured on two distinct individuals;

(iv) the association between the surrogate and true endpoints measured on two distinct individuals.

More formally, the contribution of the ith trial to the log PL can be written

$$p\ell_i = \sum_{j=1}^{2n_i} \sum_{k=1}^{j-1} \ell_{jk}, \tag{4.11}$$

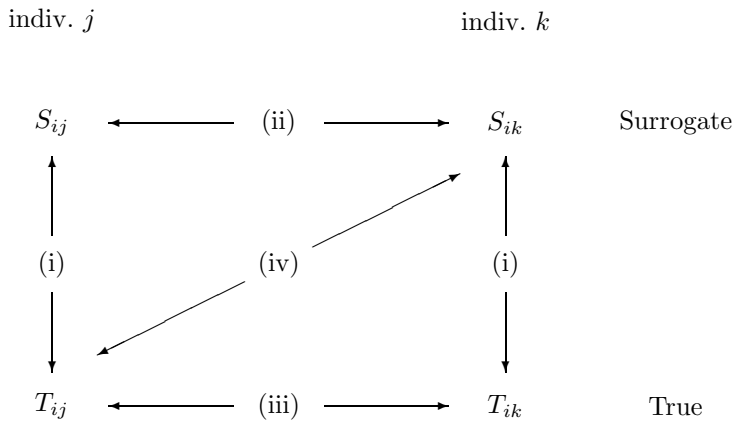
where ℓ_{jk} is the likelihood of the pair (Y_{ij}, Y_{ik}) and $\boldsymbol{Y}_i = (S_{i1}, \ldots, S_{in_i}, T_{i1}, \ldots, T_{in_i})$, that is:

$$\ell_{jk} = Y_{jk}^{(11)} \log p_{jk}^{(11)} + Y_{jk}^{(10)} \log p_{jk}^{(10)} + Y_{jk}^{(01)} \log p_{jk}^{(01)} + Y_{jk}^{(00)} \log p_{jk}^{(00)},$$

with

$$p_{jk}^{(lm)} = P[Y_{ij} = l, Y_{ik} = m]$$

Figure 4.1. *Association structure between the surrogate and true endpoints for two distinct individuals j and k in trial i.*



and

$$Y_{ij}^{(lm)} = \begin{cases} 1 & \text{if } Y_{ij} = l \text{ and } Y_{ik} = m, \\ 0 & \text{otherwise.} \end{cases}$$

As before, each term in (4.11) can be decomposed in terms of univariate and bivariate probits and inference can proceed by maximizing the log PL function (see Section 3.3 for further details). Finally, upon convergence of the algorithm, the two surrogacy measures R^2_{trial} and R^2_{indiv} are calculated using formulas (4.6) and (4.7). Approximate standard errors for these two quantities can be obtained using the delta method.

4.4 Simulations

A simulation study was conducted to further examine the behavior of the MPL estimator under different scenarios with varying trial numbers and sizes. Of particular

interest is the impact of these factors on R^2_{trial} and R^2_{indiv} and to investigate convergence issues as well.

The true underlying model in our simulation was taken to be:

$$\begin{cases} \tilde{S}_{ij} &= (0 + m_{Si}) + (-1 + a_i)Z_{ij} + \tilde{\varepsilon}_{Sij}, \\ \tilde{T}_{ij} &= (0 + m_{Ti}) + (-2 + b_i)Z_{ij} + \tilde{\varepsilon}_{Tij}, \end{cases}$$

with

$$D = \begin{pmatrix} 1 & \sqrt{0.8} & 0 & 0 \\ \sqrt{0.8} & 1 & 0 & 0 \\ 0 & 0 & 1 & \sqrt{0.8} \\ 0 & 0 & \sqrt{0.8} & 1 \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{ST} \\ \rho_{ST} & 1 \end{pmatrix}.$$

Note that additional simulations involving more complex forms of the D matrix led to similar conclusions.

Data were generated using different scenarios with fixed and variable trial sizes. We report only on a small set of simulations with fixed trial sizes (20 trials with 10 and 100 subjects) and different values of ρ_{ST}. Conclusions remained basically unchanged when trial size was allowed to vary. In each case, 250 replicates were generated. Results are presented in Table 4.2 where for each parameter, we have reported the 5%-trimmed mean, the simulation S.D. and the mean of the estimated standard errors (based on second-order numerical derivatives).

Estimates of the fixed-effects parameters seem to exhibit some bias with the smaller sample size (first two settings) but this bias is essentially eliminated when sample size increases (third setting). The estimated standard errors are relatively close to the empirical ones in all three sets of simulations. Variance parameters of the D matrix tend to be overestimated in smaller samples, owing to the higher degree of skewness in their distribution. The same comment applies, to a lesser extent, to covariance parameters, especially when their magnitude is large. These problems are not observed with the larger sample size (20 trials of size 100). It can also be noticed that standard errors of the variance parameters tend to be underestimated.

The parameter ρ_{ST} was overestimated in the first two simulation settings, especially in the first one where the bias is sizeable and the estimates exhibit more variability than in the second setting. It was, however, correctly estimated in the

Table 4.2. *Simulation results (250 replications).*

		20 trials, 10 subj. ($\rho_{ST} = \sqrt{0.5}$)			20 trials, 10 subj. ($\rho_{ST} = \sqrt{0.8}$)			20 trials, 100 subj. ($\rho_{ST} = \sqrt{0.8}$)		
Parameter	**True**	**mean**	**s.d.**	**s.e.**	**mean**	**s.d.**	**s.e.**	**mean**	**s.d.**	**s.e.**
μ_S	0	0.012	0.305	0.290	0.003	0.284	0.285	0.018	0.237	0.233
α	-1	-1.041	0.515	0.485	-1.040	0.511	0.490	-1.012	0.293	0.288
μ_T	0	0.009	0.301	0.299	0.008	0.315	0.294	0.050	0.232	0.236
β	-2	-2.048	0.820	0.731	-2.061	0.656	0.696	-1.995	0.469	0.408
d_{SS}	1	1.094	0.939	0.802	1.081	0.821	0.778	0.995	0.454	0.406
d_{ST}	$\sqrt{0.8}$	0.955	0.722	0.647	0.986	0.659	0.671	0.906	0.385	0.374
d_{TT}	1	1.291	0.956	0.860	1.157	0.819	0.812	1.031	0.404	0.415
d_{Sa}	0	-0.064	0.884	0.801	-0.072	0.765	0.750	-0.007	0.343	0.359
d_{Ta}	0	0.060	0.775	0.729	-0.018	0.736	0.711	0.002	0.328	0.351
d_{aa}	1	1.455	1.796	1.596	1.479	1.702	1.527	1.040	0.604	0.550
d_{Sb}	0	0.044	0.850	0.803	0.008	0.668	0.731	-0.026	0.463	0.427
d_{Tb}	0	-0.097	0.959	0.908	-0.105	0.741	0.805	-0.041	0.483	0.452
d_{ab}	$\sqrt{0.8}$	1.004	1.190	1.109	1.053	1.138	1.092	0.863	0.618	0.516
d_{bb}	1	1.584	2.066	1.926	1.402	1.784	1.618	1.022	1.063	0.762
ρ_{ST}	†	0.774	0.119	0.122	0.917	0.067	0.076	0.896	0.026	0.025
R^2_{trial}	0.8	0.827	0.243	0.331	0.835	0.241	0.342	0.858	0.195	0.167
R^2_{indiv}	†	0.608	0.182	0.184	0.843	0.116	0.138	0.803	0.047	0.045
Converged		203 (81%)			153 (61%)			250 (100%)		

† See column headers for true value of ρ_{ST}.

third setting, once trial size is large. Standard errors are well approximated in each case. The same comments hold for the derived parameter $R^2_{\text{indiv}} = \rho^2_{ST}$. The R^2_{trial} parameter suffers from upward bias in all three simulation settings and its distribution is strongly skewed towards small values. The amount of bias in this parameter could be attenuated by increasing replication at the trial level.

A final comment concerns convergence of the algorithm. Convergence percentages are reported in the last row of Table 4.2. It can be observed how these are affected by the magnitude of ρ_{ST} and the trial size. Note that the numbers given in the first two settings exclude cases where the solution lied close to the boundary of the parameter space (value of $\det(D)$ close to 0 or value of ρ_{ST} close to 1). As expected, this problem

was more frequent with the largest value of ρ_{ST} (second setting). No convergence problems were encountered in the last set of simulations which was characterized by a larger number of subjects in each trial.

For purposes of comparison we also used the PQL procedure, as implemented in the SAS macro GLIMMIX (Wolfinger and O'Connell, 1993), to analyze each simulated data set based on the second simulation scenario. Besides the well-known downward bias occuring in the (co)variance parameters (D and ρ_{ST}), the proportion of data sets where the algorithm converged was dramatically low. This proportion was about 44% with an unconstrained $\mathbf{\Sigma}$ matrix and dropped to about 25% when the elements on the main diagonal of $\mathbf{\Sigma}$ were constrained to equal 1. In addition, even when the algorithm did actually converge, the resulting D matrix was not always positive-definite, therefore yielding R^2_{trial} values outside the unit interval. Although the GLIMMIX macro is known to perform poorly in general, convergence difficulties are not attributable to this fact alone but rather hint upon a problem inherent to the algorithm itself. Convergence towards solutions lying on the boundary of the parameter space was also observed consistently in the analysis of real data sets using the MLwiN package, as illustrated in the next section.

4.5 Example: a Meta-Analysis of Trials in Schizophrenic Subjects

To illustrate the methodology, we use the data introduced in Section 3.6. Recall that these data come from five clinical trials comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia.

Even though we are not in a standard context for surrogate endpoint validation due to the lack of a gold standard, we consider as our primary measure (true endpoint) the CGI overall change versus baseline, dichotomized as an indicator of global improvement, i.e. a CGI score equal to 1 (='very much improved'), 2 (='much improved') or 3 (='minimally improved'). As a surrogate measure for global improvement, we take clinical response defined as a 20% or higher reduction in the PANSS score from baseline to endpoint. This corresponds to a commonly accepted criterion for defining a clinical response (Kay *et al.*, 1988). In other words, we seek to quantify the extent to which a response in PANSS, a measure of psychiatric disorder, can predict clinical improvement as observed by the physician.

Table 4.3. *Pooled data for the schizophrenia example: Surrogate endpoint (S) = response in PANSS score; True endpoint (T) = improvement in CGI overall change versus baseline.*

		T	
Z	S	0	1
Active Control	0	151 (72)†	58 (28)
	1	15 (6)	220 (94)
Risperidone	0	91 (71)	37 (29)
	1	20 (9)	213 (91)

† Frequency (row percentage)

Pooled data from the five trials are presented in Table 4.3. It can be seen that the relationship between S and T is very strong ($OR_{ST} = 31.5$, $\chi^2 = 261.4$, $P < 0.0001$), as can be expected. Note that patients were rated by the same treating physicians on both scales, thereby bringing some possible contamination bias. Table 4.3 shows parameter estimates and their standard errors for model (4.9). This model was fitted using the PQL2 procedure implemented in the MLwiN software package (Goldstein *et al.*, 1998) and using the MPL approach. Since the number of trials is too small in this example, centers were treated as grouping units. Thus, 176 units were available for analysis.

As can be seen in Table 4.4, the MPL procedure leads to an estimated D matrix that is positive-definite. With PQL2, on the other hand, some elements of D were constrained to be zero and as a result, the estimated value of the R^2_{trial} coefficient cannot even be calculated. This makes direct comparison of parameter estimates for D rather difficult between PQL2 and MPL. This put aside, fixed-effects parameter estimates are quite similar and the anticipated loss of efficiency in MPL estimates is moderate (less than 15%). Also, the parameter ρ_{ST} exhibits both a much higher point estimate and a much larger standard error.

Interestingly, the estimated value of R^2_{trial} is really low (0.006), whereas the estimated value of R^2_{indiv} is rather high (0.924). The latter confirms the strong association between S and T (at the individual level) which was seen in Table 4.3 and suggests

that they both capture overlapping components of a subject's psychotic status . The very low estimated value for R^2_{trial}, on the other hand, shows that S provides very bad predictions for treatment effects on T (at the center level), thereby making of clinical response a rather poor surrogate for clinical improvement according to our criterion. Here, we see one advantage of this approach in that individual and 'trial' (center in this example) level components of association can be completely disentangled. In such an example, this is important since both are indeed very different.

As an additional analysis, we also considered dichotomizing CGI as a score of 1 or 2 versus others. This choice gives a greater sense of confidence that a patient is effectively responding to treatment. According to guidelines for clinical trials on attention deficit hyperactivity disorder[†], which is a disease related to schizophrenia, the current consensus seems to give preference to this criterion as a response indicator. Interestingly, the resulting R^2 values were $R^2_{\text{trial}} = 0.733$ and $R^2_{\text{indiv}} = 0.824$, which seems to support this view. Note that the estimated standard error of R^2_{trial} was quite large (S.E.=2.071), which might indicate that some numerical difficulties are in play, so this result should be taken with care.

4.6 Conclusions

We have proposed an extension of the approach of Buyse *et al.* (2000) to assess the validity of a surrogate endpoint in a meta-analytic context when both the surrogate and the final endpoints are discrete in nature, the emphasis being on binary outcomes. This was done by adopting a latent variable model formulation which allows us to carry over previously proposed measures of surrogacy in a natural way, under the assumption that the latent variables are normally distributed. This, in turn, dictates the use of a joint probit model for the surrogate and the true endpoints.

The major difficulty rests in parameter estimation since, on the one hand, a direct likelihood approach would be computationally involved and, on the other hand, standard approximate methods such as PQL may not be satisfactory since interest centers directly on the random components of the model. This was our prime motivation for using the MPL approach as it provides a net balance between computational burden and bias, although at the (small) price of lower efficiency.

[†]From the consensus meeting of the European College of Neuropsychopharmacology at Nice, March 2002 (personal communication).

Table 4.4. *Results for the schizophrenia data using PQL2 and MPL. Parameter estimates and standard errors are reported.*

	PQL2		**MPL**	
Parameter	**Estimate**	**S.E.**	**Estimate**	**S.E.**
μ_S	0.227	0.056	0.233	0.062
$\alpha^\dagger$	0.166	0.046	0.161	0.049
μ_T	0.441	0.054	0.445	0.062
$\beta^\dagger$	0.100	0.050	0.109	0.057
d_{SS}	0.126	0.050	0.121	0.057
d_{ST}	0.088	0.042	0.091	0.055
d_{TT}	0.083	0.045	0.076	0.063
d_{Sa}	-	-	-0.005	0.054
d_{Ta}	-	-	-0.004	0.040
d_{aa}	-	-	0.001	0.005
d_{Sb}	-0.007	0.024	0.006	0.046
d_{Tb}	0.001	0.022	0.024	0.041
d_{ab}	-	-	-0.001	0.002
d_{bb}	0.029	0.023	0.059	0.045
ρ_{ST}	0.679	0.018	0.961	0.027
R^2_{trial}	-	-	0.006	0.082
R^2_{indiv}	0.461	0.024	0.924	0.052

† Treatment coding: -1 = active control, +1 = risperidone.

It is well-known that generalized linear mixed models are challenging to fit in general and can pose numerous estimation problems. From our personal experience, it is not so uncommon for the PQL algorithm to exhibit numerical instability and fail to converge. The problem is even worse with the second order PQL algorithm (PQL2) and with more complicated models such as (4.9). We saw in the previous chapter that MPL tends to be more robust against convergence problems, which gives an added advantage to this procedure.

Numerical problems should nevertheless be expected to occur frequently in the kind of applications sought here. In particular, such factors as the number of trials, between-trial variability and trial size can be critical for improving convergence properties of the algorithm, just as they are for normally distributed endpoints. Obviously, these problems remain topical here, if not worsened, as less information is conveyed by binary response variables than by continuous ones.

To conclude, we briefly outline how the method can be extended to ordinal endpoints. We can adopt the "threshold concept" and assume that there are unobservable latent variables that are related to the actual responses S and T through a series of cutoff points. For instance, if S has K categories, we need to define a set of $(K-1)$ threshold values $\gamma_1 < \ldots < \gamma_{K-1}$ and postulate that S and the corresponding normally distributed latent variable, U say, are connected by

$$S = k \Leftrightarrow \gamma_{k-1} < U \leq \gamma_k, \qquad k = 1, \ldots, K$$

with $\gamma_0 = -\infty$ and $\gamma_K = +\infty$ and where, for convenience, we can assume that $\gamma_1 = 0$. On the latent variable scale, we can again consider model (4.8) and the associated coefficients of determination as measures of surrogacy at the trial and individual levels. Parameter estimation can proceed as before by considering the likelihood of all possible pairs of outcomes. Threshold values used to define S and T are simply extra parameters to be estimated in the PL function. An extension to mixed situations, where one endpoint is discrete and the other is continuous, is also feasible. The same modeling strategy can be followed, with one of the components assumed to be normally distributed and the other being obtained *via* a latent, normally distributed, variable. The PL function does then involve evaluation of univariate probits only.

Part II :

Longitudinal Data Modeling

Chapter 5

Repeated-Measures Models to Evaluate a Hepatitis B Vaccination Program

5.1 Introduction

The aim of this chapter is to provide an introduction to the analysis of longitudinal data and to discuss some of the associated modeling-related issues. To illustrate the concepts, we shall utilize data from a hepatitis B vaccination program that was conducted in 1985–86 in a Belgian institution for the mentally handicapped in order to evaluate the long-term persistence of antibodies against hepatitis B surface antigen (anti-HBs) after vaccination in this population. The data, as well as the main scientific questions, will be described in Section 5.2.

To analyze the data, a standard linear mixed model will be used. An attractive feature of linear mixed models for longitudinal data is that they enable the analyst to discriminate between three specific components of variation in the data, which are between-subjects variability, serial correlation and measurement error. These different aspects will be portrayed in Section 5.3.

An important consideration with longitudinal data is the formulation of a suitable model to describe the evolution of the response variable over time, especially when prediction is a valued goal of analysis. Time trends are commonly modeled with

low-order polynomials but these are known to be poor for prediction purposes due to their asymptotic behavior. Fractional polynomials, which will be discussed in Section 5.4, constitute a highly flexible tool in this respect. They provide a wide range of functional forms and are straightforward to fit, two very appealing features.

In Section 5.5 we present the model building procedure for two specific models. The first one, saturated in time, easily deals with the nonlinearity of the profiles and is helpful for making comparisons between groups at different time points for instance. The second model describes temporal decline in antibody titers more parsimoniously and is more useful for making long-term predictions. Section 5.6 illustrates the issue of prediction one year past the study end. We conclude with a few remarks in the last section. Note that most of the contents of this chapter can be found in Renard *et al.* (2001).

5.2 Hepatitis B Vaccination Program and Scientific Questions

Mentally handicapped individuals residing in institutions are at high risk for hepatitis B virus (HBV) acquisition and subsequent carrier state. The higher risk of non-parenteral transmission in this population is due to the typical behavior of mentally retarded patients, the type of mental retardation and the closed setting of the institutions which all enhance spreading of the virus.

Hepatitis B vaccination of residents and staff is a general recommendation and has become part of today's hepatitis B prevention programs. Data on long-term persistence of antibodies against HBV are scarce, especially in this population. Data available from other high-risk populations showed that 67 to 85% of the vaccinated individuals still had antibody levels higher than 10 International Units/Liter (IU/L), 9 to 12 years after the first vaccine dose (Hadler *et al.*, 1986; Tabor *et al.*, 1993; Coursaget *et al.*, 1994; Da Villa *et al.*, 1996; Wainwright *et al.*, 1997).

In previous studies several factors have been described to cause a higher risk in the acquisition of hepatitis B virus infection. These factors include age, age at admission, duration of residency, type of mental retardation (Down's syndrome (DS) versus other types of mental retardation (OMR)), sex and use of anti-epileptic medication (Vellinga, Van Damme and Meheus, 1999). Sex, age and type of mental retardation are also of influence on the response to vaccination (Vellinga, Van Damme, Weyler *et*

al., 1999).

In 1985–86 a hepatitis B vaccination program was conducted in an Belgian institution for the mentally handicapped to evaluate the long-term persistence of anti-HBs after vaccination in this population. Blood samples were drawn from residents in that institution, who were then all vaccinated with three doses of hepatitis B vaccine (Engerix-B™, SmithKline Beecham Biologicals, Rixensart, Belgium) according to a month 0-1-6 schedule. Serum samples were taken after each vaccine dose and if residents did not meet the (arbitrary) antibody level of 100 IU/L at month 7, they received an extra vaccine dose at month 12. If the requirement of 100 IU/L was still not met at month 13, additional booster doses were administered (these residents were however not further included in the program). All residents received a booster dose after 5 years, at month 60.

Of the 196 seronegative residents originally included in the program, only 97 were included in the analysis of the follow-up after 11 years. They had blood samples taken yearly for the first 5 years and at year 11. Sixty seven of them received 4 vaccine doses (at months 0, 1, 6, plus a booster at month 60) and 30 received 5 doses (at months 0, 1, 6, 12, plus a booster at month 60). Further details can be found in Van Damme *et al.* (1989) and Vellinga, Van Damme, Weyler *et al.* (1999).

Interest focuses on describing the evolution of the mean log titer over time, while accounting for prognostic factors such as sex, body mass index, duration of residency, age at admission into the institution, type of mental retardation, use of antiepileptic drugs and number of vaccine doses received. Additional questions also involve predicting antibody level at years 11 (end of study) and 12 (one year past the end) based on the fitted model.

While the main epidemiological interest lies in the population-averaged prediction, the model enables one to perform individual-specific predictions as well. Both model building and prediction are complicated by the fact that individual and average profiles are highly nonlinear (see Figures 5.1a and 5.1b), combined with the absence of measurements between years 5 to 11.

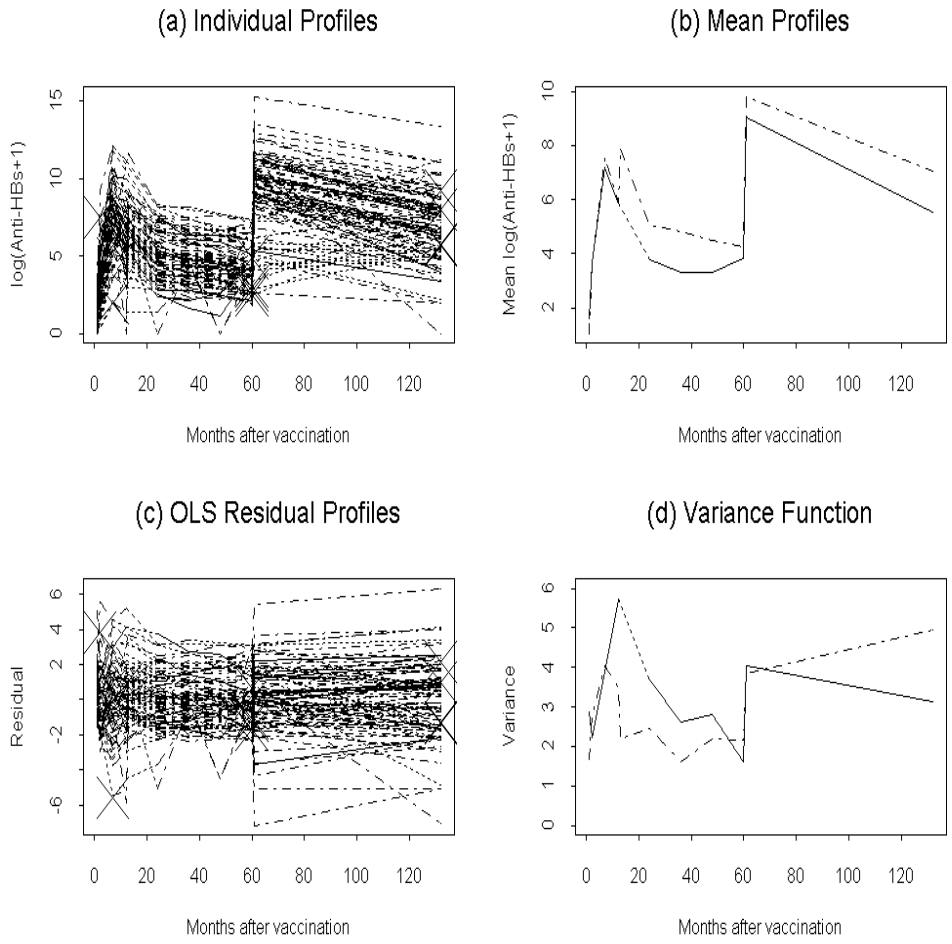


Figure 5.1. *(a) Longitudinal trends in log(anti-HBs+1) for residents with DS (solid line) and OMR (dashed line). Cross symbols indicate missing values.*
(b) Average log(anti-HBs+1) over time for residents with DS (solid line) and OMR (dashed line).
(c) Ordinary least squares (OLS) residual profiles obtained upon fitting a saturated mean structure to log(anti-HBs+1).
(d) Variance of the OLS residuals over time for residents with DS (solid line) and OMR (dashed line).

5.3 The Linear Mixed Model With Serial Correlation

In this section, we briefly review the general linear mixed-effects model, with some emphasis on components of random variability that are typically encountered in longitudinal data.

Linear mixed-effects models for longitudinal data were proposed by Laird and Ware (1982). Serial correlation was added by Diggle (1988). See Diggle, Liang and Zeger (1994) and Verbeke and Molenberghs (2000) for a general overview.

A linear mixed model for longitudinal data can be written as follows:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{5.1}$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\dagger$ is the n_i dimensional response vector for subject i, $1 \leq i \leq N$, N is the number of subjects, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ known design matrices, $\boldsymbol{\beta}$ is the p dimensional vector containing the fixed effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$ is the q dimensional vector containing the random effects, $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \Sigma_i)$ is a n_i dimensional vector of residual components, and $\boldsymbol{b_1}, \ldots, \boldsymbol{b_N}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_N}$ are assumed to be independent. Inference is based on the marginal distribution of the response vector $\boldsymbol{Y_i}$, that is:

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i). \tag{5.2}$$

Random effects in model (5.1) stem from heterogeneity between individuals. This means that various aspects of their behavior may exhibit inter-individual random variation. For example, it is conceivable that some subjects will have a high immune response, and will remain so throughout the study period.

The residual variability $\boldsymbol{\varepsilon}_i$ in (5.1) may be further refined and decomposed into the following qualitatively distinct components (Diggle, Liang and Zeger, 1994):

- **Serial correlation:** This component arises due to the fact that pairs of measurements (such as antibody titers) that are taken closer in time often show a stronger similarity than pairs taken further apart.
- **Measurement error:** Measurement errors occur when the measurement process itself introduces an element of random variability. For instance, there

†Note that we will henceforth make use of the standard index notation for longitudinal data, that is, the first index refers to individuals, while the second refers to measurement occasions.

might be substantial variation in results from bioassays of blood samples, even when two measurements are taken at the same time from the same subject, or when a sample is split into two subsamples which are then analyzed separately.

This distinction leads to the decomposition $\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_i^{(1)} + \boldsymbol{\varepsilon}_i^{(2)}$, where

$$\begin{cases} \boldsymbol{\varepsilon}_i^{(1)} \sim N(\mathbf{0}, \tau^2 H_i), \\ \\ \boldsymbol{\varepsilon}_i^{(2)} \sim N(\mathbf{0}, \sigma^2 I_{n_i}). \end{cases} \tag{5.3}$$

In this expression, the first component $\boldsymbol{\varepsilon}_i^{(1)}$ captures serial correlation, while the second represents measurement error. The covariance matrix H_i only depends on i through the number n_i of observations and through the time points t_{ij} at which measurements are taken.

The structure of the matrix H_i is determined by the autocorrelation function $\rho(t_{ij} - t_{ik})$, for some decreasing function $\rho(.)$ with $\rho(0) = 1$. A first simplifying assumption is that it depends only on the time interval between two measurements Y_{ij} and Y_{ik}, i.e. $\rho(t_{ij} - t_{ik}) = \rho(\mid t_{ij} - t_{ik} \mid)$. Two popular choices for $\rho(.)$ are the exponential and Gaussian models defined respectively as $\rho(u) = \exp(-\phi u)$ and $\rho(u) = \exp(-\phi u^2)$, with $\phi > 0$.

When model (5.1) contains solely a random intercept between subjects, a serially correlated component and a measurement error, a useful aid to the formulation of an appropriate model for the covariance structure, especially the autocorrelation function, is the variogram (Diggle, 1990). For a stochastic process $Y(t)$, the variogram is defined as $V(u) = \frac{1}{2}E[Y(t) - Y(t-u)]^2$. Under the specified model, this reduces to $V(u) = \sigma^2 + \tau^2[1 - \rho(u)]$ (Diggle, Liang and Zeger, 1994).

Decomposition (5.3) assumes that the variance of residual components $\boldsymbol{\varepsilon}_i$ is constant over time. However, individual profiles in Figure 5.1a exhibit a decrease in variability during the first half of the study, which is confirmed by the plot of the variance function displayed in Figure 5.1d. One way to accomodate variance heterogeneity is through a log-linear variance model producing exponential local effects, also called dispersion effects (Littell *et al.*, 1996). In this model, measurement errors take the form $\sigma^2 \text{diag}[\exp(U\boldsymbol{\delta})]$, where U is a design matrix and $\boldsymbol{\delta}$ a vector of dispersion parameters. This affords a way of modeling variability in terms of effects to be specified, such as time in the present study.

5.4 Fractional Polynomials with Longitudinal Data

Fractional polynomials were proposed by Royston and Altman (1994) as a flexible tool for parsimonious parametric modeling. In simple terms, a fractional polynomial $\phi(X;\boldsymbol{\beta},\boldsymbol{p})$ is a linear combination of real-valued powers of X, where X represents some (not necessarily continuous) covariate. More formally, a fractional polynomial $\phi(X;\boldsymbol{\beta},\boldsymbol{p})$ of degree m can be defined as the function

$$\beta_0 + \sum_{j=1}^{m} \beta_j X^{(p_j)},$$

where the β_j are regression parameters and $\boldsymbol{p} = (p_1, \ldots, p_m)$ is a real-valued vector of powers with $p_1 < \ldots < p_m$. The notation $X^{(p)}$ denotes the Box-Tidwell power transformation

$$X^{(p)} = \begin{cases} X^p & , \; p \neq 0 \\ \log X & , \; p = 0 \end{cases}$$

Note that the above definition can be extended to the case of equal powers (see Royston and Altman, 1994).

By definition, fractional polynomials extend the family of the Box-Tidwell power transformation and the class of conventional polynomials. A great advantage of fractional polynomials over classical polynomials is that they provide a wide range of functional forms and their behavior near the extreme values is often more reasonable. It is possible, for instance, to generate a variety of curves some of which approach a horizontal asymptote for large values of X, a feature that classical polynomials do not share.

Another advantage of fractional polynomials is that they are straightforward to fit. To determine the 'best' value of m and $\boldsymbol{p}$, Royston and Altman (1994) propose to restrict the power terms to a small predefined set of integer and non-integer values – they suggest using $\mathcal{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, \ldots, \max(3, m)\}$ – and to select the best power vector as that associated with the model with the highest likelihood (or equivalently with the lowest deviance). As with conventional polynomials, the degree m of the fractional polynomial is selected either informally on *a priori* grounds or by increasing m until no worthwile improvement in the fit can be detected. In practice, it seems that a choice of $m = 2$ or $m = 3$ is typically sufficient.

Whereas the same broad principles hold with longitudinal data, formulation of the model must be approached with greater care since in addition to modeling the mean

structure, modeling of the covariance structure must be undertaken. General guidelines for model building with longitudinal data involve the following steps (Verbeke and Molenberghs, 2000):

- **Selection of a preliminary mean structure:** it is recommended to use an overelaborated model for the mean response profile.
- **Selection of a preliminary random-effects structure:** we have to select a set of random effects to be included in the covariance model.
- **Selection of a residual covariance structure:** conditional on the previously selected set of random effects, we need to specify the residual covariance matrix Σ_i.
- **Model reduction:** based on the residual covariance structure specified in the previous step, we can investigate whether the included random effects are really needed in the model. After selection of the final covariance structure, the preliminary mean structure can be simplified.

The model building process as just described lends itself to a simple strategy to determine the best power vector $\boldsymbol{p}$: after an appropriate covariance structure has been selected, the best-fitting m-tuple of power terms from $\mathcal{P}$ can be chosen based on the likelihood criterion. This will preferably be done using several variance-covariance structures to make the procedure more robust. A drawback of the approach, however, is that time effects prespecified in the random-effects structure will typically differ from those selected in the mean structure. For example, one may have specified a random-effects structure with an intercept, t and t^2, where t denotes the time variable. On the other hand, the selection of a fractional polynomial of order 2 may result in an intercept plus the terms $\log t$ and t^{-1}. This seems rather unnatural if the model is to be interpreted as a random-coefficients model, where the random effects represent subject-specific corrections to the average response profile.

The latter remark suggests another possible strategy to choose a power vector by comparing models that include the same random and fixed time effects, in addition to possibly allow for a fixed serial correlation process. Thus, in the above example, the model including fixed and random effects for the intercept, t and t^2 could be compared to the model including fixed and random effects for the intercept, $\log t$ and t^{-1} and that with highest likelihood would be selected. After the procedure has been run, it

would be advisable to assess the appropriateness of the assumed covariance structure to check if model fit can be improved. Note that this procedure is more likely to run into convergence problems because of the diversity of random-effects structures to be fit.

As can be seen, none of the two aforedescribed procedures is perfect and one may be preferred over the other depending on the context. When there is a small number of fixed measurement occasions for example, the first method might be preferable since an unstructured matrix can be assumed for the marginal variance-covariance matrix of a response vector and emphasis put on selecting a power vector for fixed effects only. In more general situations (e.g. with unequal measurement times), random effects and/or serial correlation structures will typically be needed and either of the two proposed strategies can be used. Also, more complex procedures can presumably be conceived. For example, a selection procedure could be ideally to simplify a 'super' model incorporating all time effects with power terms in $\mathcal{P}$, both as random and fixed effects. Of course, it is unlikely that such a model can be fit in practice.

5.5 Time-evolution of Antibodies

In this section, we address the question of specifying a model that adequately describes the evolution of log antibody titer over time. Hence, we need to consider appropriate mean, variance and covariance models. Since the profiles are quite messy due to unequally spaced measurement occasions and booster effects, it is essential to conduct an exploratory data analysis.

As shown in Figures 5.1a and 5.1b, individual and mean profiles of log(anti-HBs+1) for DS and OMR are clearly nonlinear and show peaks after booster doses. Individual profiles follow approximately the same pattern with the main difference between profiles lying in the vertical shift. This suggests a strong contribution of an individual random intercept. Average profiles show a difference in anti-HBs between the two groups. Also, these profiles exhibit steep increases immediately after boosters, followed by a gradual decrease which appears to be nonlinear.

Figure 5.2 depicts an estimate of the empirical variogram for these data. It was constructed using standardized ordinary least squares residuals obtained upon fitting a saturated groups-by-times model (where group is type of mental retardation). Also shown in this figure is a smooth loess estimate of the variogram (Cleveland, 1979). The between-subject variance seems relatively large in these data, accounting for

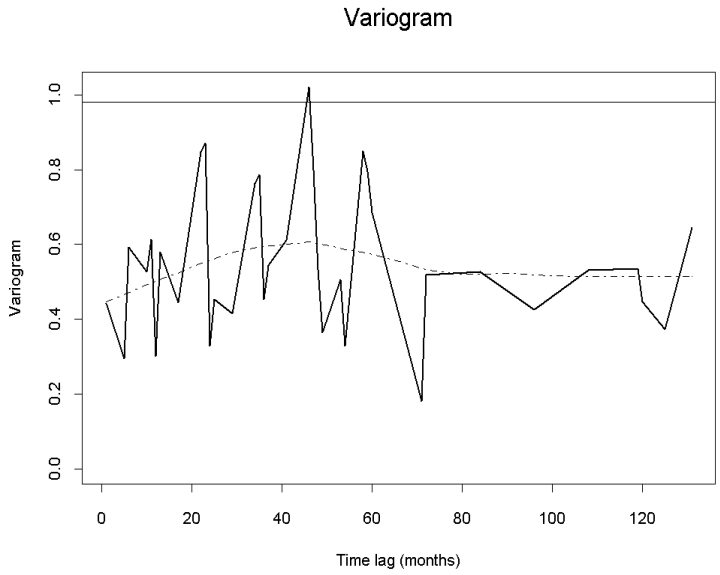


Figure 5.2. *Sample variogram of log antibody residuals (the horizontal line estimates the process variance; the dashed line represents a smooth estimate of the variogram).*

about one half of the total variability. The measurement error is also substantial, accounting approximately for the other half of the process variance. This variogram leaves little room for a serially correlated component. Note that in this context, it is essential to use standardized residuals to remove variance heterogeneity in the data, ensuring that the process variance is constant and equal to one.

Following the guidelines sketched in the previous section, we now turn to model building and outline the successive steps to retain a final model. Type of mental retardation, duration of residency and number of vaccine doses (as a group variable) were allowed to have specific effects at each sample occasion. We also included time-constant effects for sex (male versus female), use of antiepileptic drugs (yes versus no), body mass index, and age at admission in the institution, since there was no indication of a time trend for these covariates.

The preliminary variance model acknowledges the presence of serial correlation and includes the following random effects: an intercept, a linear time slope, and number of vaccine doses (0/1-coded group variable). An unstructured form is assumed for the

3×3 random-effects variance matrix D.

We first select an appropriate serial process, as shown in Table 5.1. Models with exponential (B) and Gaussian (C) serial correlation are compared to the model with no serial process (A) using the likelihood ratio test statistic (denoted G^2 in the table). These tests strongly reject the null hypothesis of no serial process. At this stage, we decided to keep the exponential model for comparison purposes with a first analysis of the data. As a rule of thumb, however, one should generally go for the model with the largest value of the deviance (or another criterion, such as Akaike's or Schwarz's Bayesian information criteria). Note that substantially no differences were seen between the two models.

Next, the random-effects structure may be simplified. Three hierarchically ordered models are presented in the second part of Table 5.1. One must be very careful in interpreting the significance of random effects using the likelihood ratio test statistic G^2: the associated testing problem is indeed non-standard as the null hypothesis lies on the boundary of the parameter space of the alternative hypothesis (Verbeke and Molenberhs, 2000). The reference distribution for the B–D comparison is a 50:50 mixture of χ^2_2 and χ^2_3. Similarly, for the comparison of Models D and E we obtain a 50:50 mixture of χ^2_1 and χ^2_2 variables. These distributions have been utilized to calculate the corresponding p-values. Thus, at this stage we select model D, comprising a random intercept and a random time slope.

Finally, retaining the covariance structure we have just selected, the mean model can be reduced. Effects kept in the final model were time, type of mental retardation and number of vaccine doses (with unstructured time effects), duration of residency (with a linear time trend), use of antiepileptic medication and sex (time-constant effects). Although not significant, sex was kept in the model for reasons of external comparison.

Parameter estimates for this model are shown in Table 5.2. Note that the random intercept and random time slope are assumed to be independent. In fact, a comparison between model-based and empirical (or robust) standard errors revealed large discrepancies (relative increases more than three-fold for most of the estimates) in the final model. Empirical standard errors correct for potential misspecification of the covariance structure (Liang and Zeger, 1986; Diggle, Liang and Zeger, 1994) and disagreement between both types of standard errors might point to an inadequately specified covariance structure. Arguably, we had little reason to believe that the selected covariance structure is substantially incorrect. Therefore, it is wise to attain

Table 5.1. *Selection of a serial correlation process and a random-effects structure.*

		Nr. of parameters			Comp.			
Model	**Description**	**Random**	**Serial**	**Deviance**	**Model**	G^2	**d.f.**	***p*-value**
Selection of a serial correlation process:								
A	Without serial process	6	0	2623.72				
B	Exponential process	5†	2	2575.23	A	48.49	1	< 0.0001
C	Gaussian process	6	2	2574.31	A	49.41	2	< 0.0001
Selection of a random-effects structure:								
B	Int., time and nr. of doses	5†	2	2575.23				
D	Intercept and time	3	2	2578.54	B	3.31		0.269††
E	Intercept	1	2	2606.22	D	23.68		< 0.0001†††

† One parameter could not be estimated due to parameter constraints.

†† From a 50-50 mixture of χ^2_2 and χ^2_3 distributions.

††† From a 50-50 mixture of χ^2_1 and χ^2_2 distributions.

a trade-off between model fit as reported by likelihood ratios and differences occuring between model-based and empirical standard errors. In particular, assuming a diagonal instead of an unstructured covariance matrix for the random effects yields a much better model in this respect and was therefore retained as our final model. Most of the estimated empirical standard errors in Table 5.2 do not exhibit changes of more than 25% compared to model-based standard errors.

It is worth noting that the effect of Down's syndrome on antibody titer was significant at months 24, 36 and 48, indicating a faster decline in anti-HBs in this population than in other mentally retarded. There did not seem to be a difference in immediate response to vaccination between these two groups. Also, we see that the extra dose given at month 12 in G2[†] had sufficiently elevated antibody titer so as to render it almost indistinguishable from antibody titer in G1 until year 5. Yet, administration of a booster dose at that time again led to better responses in G1 and this was still visible at year 11.

A similar modeling exercice can be performed on post-vaccination data, that is, data available after the last vaccination, at month 6 (G1) or 12 (G2). We simply need to specify specific models for both groups since post-vaccination times are different. We can set up a model for pre-booster data (until month 60) and then transpose this model to post-booster data, using an indicator variable for the time of booster administration. For instance, a simple model ignoring potential covariates could be written as follows:

$$E[Y_{ij}] = \begin{cases} \beta_0^{(1)} + \beta_1^{(1)} I(t_j \geq 55) + \phi(t_j - 55.I(t_j \geq 55)) & \text{in group G1,} \\ \beta_0^{(2)} + \beta_1^{(2)} I(t_j \geq 49) + \phi(t_j - 49.I(t_j \geq 49)) & \text{in group G2,} \end{cases}$$

where $\phi(t)$ is a fractional polynomial.

The process of selecting a covariance structure resulted in similar conclusions as before, with the difference that no spatial process was found necessary. For the selection of the mean structure, the antibody titer measurement obtained at the last vaccination could possibly be included as a baseline value in addition to the other covariates, but this was not possible because no measurements were taken at month 6. Instead, the first log antibody titer measurement was used as baseline. Selected covariates were type of mental retardation (with a different effect depending on the

[†]G1 (resp. G2) refers to the group of residents vaccinated according to a month 0-1-6 (resp. 0-1-6-12) schedule.

Table 5.2. *Parameter estimates and standard errors (model-based; empirically corrected) for the final model (original data).*

Effect	Time	Estimate (S.E.)	Effect	Time	Estimate (S.E.)
Mean Structure:			***Mean Structure (continued):***		
Intercept		9.361 (0.408; 0.378)	Number of doses	36	-0.250 (0.272; 0.303)
Time	1	-6.801 (0.207; 0.238)	Number of doses	48	0.003 (0.331; 0.346)
Time	2	-4.296 (0.210; 0.198)	Number of doses	60	0.243 (0.324; 0.335)
Time	7	- †	Number of doses	61	-2.009 (0.410; 0.445)
Time	12	-1.664 (0.181; 0.126)	Number of doses	132	-1.997 (0.410; 0.412)
Time	13	-0.539 (0.358; 0.366)	Residency		-0.040 (0.015; 0.012)
Time	24	-3.361 (0.196; 0.188)	Residency⋆Time		0.005 (0.002; 0.002)
Time	36	-3.682 (0.213; 0.187)	Sex		-0.013 (0.225; 0.230)
Time	48	-4.214 (0.275; 0.268)	Antiepileptic drugs		-0.629 (0.239; 0.232)
Time	60	-4.687 (0.308; 0.246)			
Time	61	1.621 (0.338; 0.253)			
Time	132	-1.924 (0.585; 0.549)	***Random-Effects:***		
DS/OMR	1	0.598 (0.638; 0.535)	Intercept		0.659
DS/OMR	2	0.374 (0.676; 0.623)	Time		0.015
DS/OMR	7	-0.023 (0.564; 0.562)	***Serial Structure:***		
DS/OMR	12	-0.303 (0.605; 0.821)	Variance		0.581
DS/OMR	13	- ††	Rate of exponential decrease ($1/\rho$)		2.319
DS/OMR	24	-1.562 (0.527; 0.739)	***Measurement Error:***		
DS/OMR	36	-1.749 (0.459; 0.596)	Time	1	1.323
DS/OMR	48	-1.498 (0.559; 0.621)	Time	2	1.367
DS/OMR	60	-0.613 (0.543; 0.436)	Time	7	0.758
DS/OMR	61	-0.833 (0.692; 0.541)	Time	12	0.701
DS/OMR	132	-1.180 (0.691; 0.386)	Time	13	0.778
Number of doses	1	-1.337 (0.367; 0.238)	Time	24	0.479
Number of doses	2	-1.781 (0.402; 0.358)	Time	36	0.000
Number of doses	7	-2.447 (0.330; 0.357)	Time	48	0.455
Number of doses	12	-2.708 (0.348; 0.418)	Time	60	0.474
Number of doses	13	-†††	Time	61	1.307
Number of doses	24	-0.106 (0.305; 0.361)	Time	132	0.310

DS/OMR: 1 = DS, 0 = OMR; Number of doses: 1 = 5 doses, 0 = 4 doses.
Sex: 1 = male, 0 = female; Antiepileptic drugs: 1 = use, 0 = no use.
† Month 7 taken as reference point because the decision to give an extra booster dose was taken at that time.
†† No measurements were available at month 13 in DS patients.
††† No measurements were available at month 13 in the group that was administered 4 vaccine doses.

number of vaccine doses administered), use of antiepileptic medication and sex. Table 5.3 presents the parameter estimates of the final model, for the fixed effects only.

In order to visually assess the fit of these two models, Figure 5.3 shows observed and predicted average profiles for combinations of number of vaccine doses and type

Table 5.3. *Parameter estimates and standard errors (model-based; empirically corrected) for the fixed effects of the final model (post-vaccination data).*

Effect	Estimate (S.E.)	Effect	Estimate (S.E.)
Intercept G1	8.808 (0.299; 0.284)	**Intercept G2**	6.389 (0.503; 0.582)
$\boldsymbol{I(t \geq 55)}$	3.501 (0.151; 0.163)	$\boldsymbol{I(t \geq 49)}$	1.358 (0.272; 0.346)
$\boldsymbol{(t-55I(t \geq 55))^{-3}}$	-0.384 (0.202; 0.178)	$\boldsymbol{(t-49I(t \geq 49))^{-3}}$	1.601 (0.455; 0.446)
$\boldsymbol{\log(t-55I(t \geq 55))}$	-1.120 (0.052; 0.057)	$\boldsymbol{\log(t-49I(t \geq 49))}$	-0.435 (0.131; 0.154)
DS/OMR G1	-0.494 (0.673; 0.702)	**DS/OMR G2**	-2.706 (0.688; 0.533)
Sex	0.037 (0.278; 0.276)	**Use of antiepileptic drugs**	-0.632 (0.284; 0.281)

of mental retardation. For predicted average profiles, all other covariate effects were set equal to their mean values.

5.6 Prediction at Year 12

This section addresses the issue of predicting antibody titer at year 12, i.e. one year after last follow-up contact. This extrapolation problem was complicated by the design feature that no measurements were available between months 61 and 132, whereas apparently we have to cope with nonlinear profiles. While the use of a time-saturated model for the mean structure is viable to tackle these features, it is less useful when it comes to prediction purposes, in particular when interest centers on future prediction. We nevertheless propose two simple methods to perform such a prediction and compare the results to the straightforward approach provided by the fractional-polynomial model on post-vaccination data.

The first approach merely uses a linear extrapolation based on an individual's last two measurements (at months 61 and 132). The resulting extrapolations are then averaged out to obtain prediction at month 144. Obviously, this approach can be criticized as being overly simple since the profiles are clearly nonlinear over the first five years. A refinement of this method might consist of overlaying profiles for the month 61-132 period with profiles from the first part of the study and then extrapolating until month 144. This raises some technical difficulties though, since the starting point of the first period (time of the last vaccine dose) depends upon the

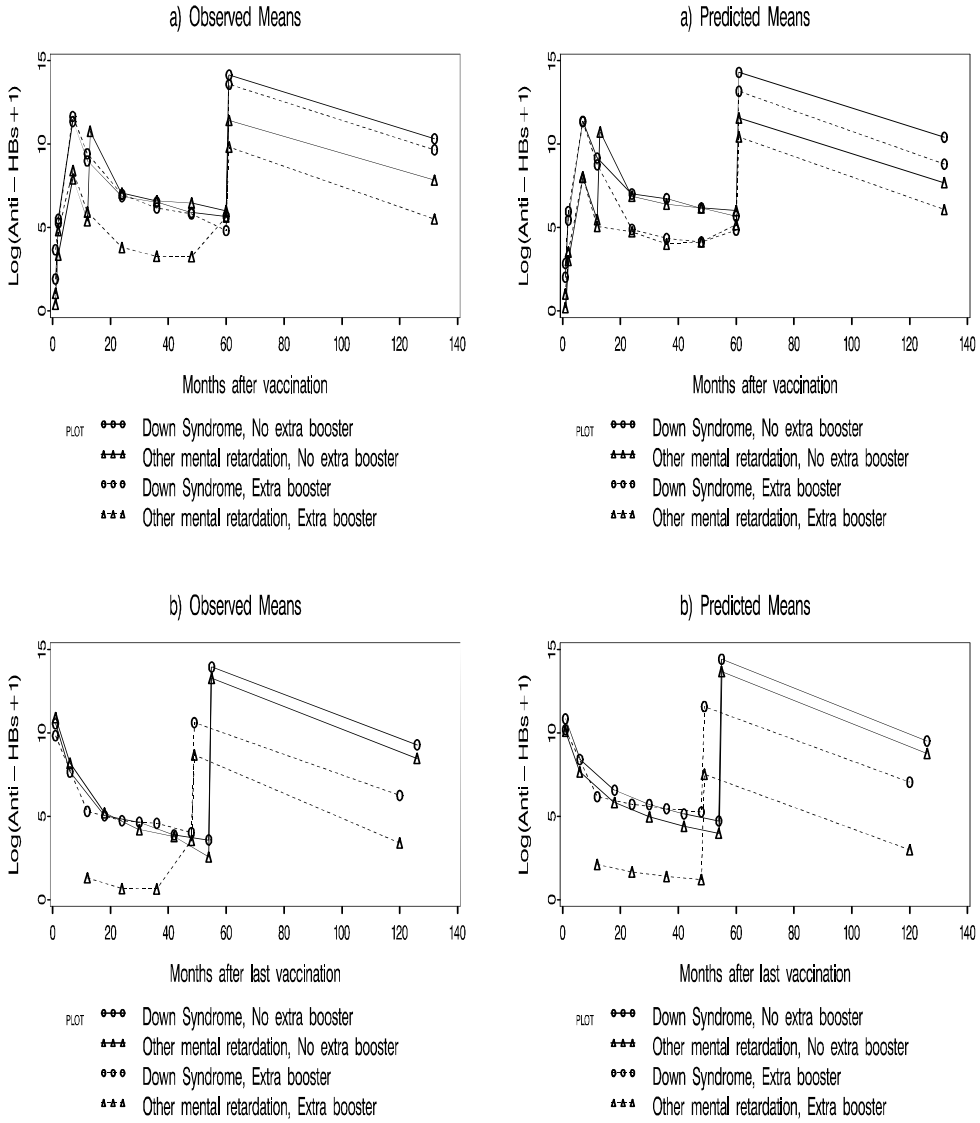


Figure 5.3. *Observed and predicted mean profiles for combinations of number of vaccine doses and type of mental retardation: a) original data; b) post-vaccination data*

Table 5.4. *Predicting log antibody titer (IU/L) at year 12: a) Approach 1: linear interpolation (original data); b) Approach 2: refined linear interpolation (original data); c) Approach 3: fractional-polynomials model (post-vaccination data)*

Group	Approach 1	Approach 2	Approach 3
4 vaccine doses (G1)	7.05	7.38	7.13 (0.23)†
5 vaccine doses (G2)	5.01	5.42	5.41 (0.35)†

† Standard errors are reported in parentheses.

group being considered: month 7 for group G1 and month 13 for G2. Using month 24 as a cut-off point to split the first time period into two pieces, we can linearly approximate the profiles in these two time windows, translate them to the month 61-132 period, extrapolate until month 144 and eventually average the results out accross the two groups.

Predictions at year 12 based on these two extrapolation schemes are displayed in Table 5.4, together with the prediction inferred from the model on post-vaccination data. As can be seen, all three approaches yield quite similar results if we look at the variability of the predictions in the third model.

We conclude this section with two remarks. Firstly, in this study prediction takes place only one year after completion of the study, which is not too distant in time compared to the duration of the study. Had prediction to be done several years later, we would presumably observe larger discrepancies. Secondly, in studies where more emphasis is to be put on prediction, it is a good idea to plan some intermittent assessment occasions to aid in modeling long-term temporal evolution. A simple model, e.g. using fractional polynomials, might then be used straightforwardly for making long-term inference with more confidence.

5.7 Conclusions

We have illustrated linear mixed models methodology for longitudinal data in the evaluation of a hepatitis B vaccination program. This approach was based upon recognition that response to vaccination may be attributed to a combination of individual-

specific characteristics but also to different other sources of (unknown) variability. Use of random effects in this setting was already proposed by Coursaget *et al.* (1994) and Gilks *et al.* (1993) who considered between-individuals variability in a Bayesian random-effects model.

On accomodating individual-specific effects, the model enables a much more precise assessment of important explanatory variables, such as number of vaccine doses received, whether or not a person has Down's syndrome, and of course time effects. In particular, the strong contributions of random intercepts and serial correlation show the importance of the initial response as well as the individual trajectory for the further evolution of anti-HBs profiles. Models that restrict attention to geometric mean titers (GMT) calculation are not able to include such individual-specific effects, typically resulting in less precise inference, also for the fixed effects.

In this study, no difference could be detected between DS and OMR patients in their immediate response, but we found that DS induces an accelerated decrease in antibody titers, implying that the rate of decline in antibody titers might be different in these two populations. This might explain why some other studies attempting to demonstrate a difference between DS and OMR patients in their anti-HBs response after vaccination have failed (see Vellinga, Van Damme, Bruckers *et al.* (1999) for a further discussion). Another point concerns whether antiepileptic medication has an influence on the immune system. However, it is hard to decide whether this is due to the medication itself, or rather an indication for the influence of epilepsy, or both (De Ponti *et al.* (1993) further discuss this point).

While there is some interest in modeling the complete set of data from a descriptive viewpoint, this could only be achieved with the help of a time-saturated model to account for the high nonlinearity in the profiles. If one is interested in a more parsimonious, parametric description of the temporal decline in antibody titer, to address such questions as long-term inference, one needs to resort to an alternative solution. Focusing the analysis on post-vaccination data was a suitable alternative, permitting simple parametric modeling of the anti-HBs evolution over time. Obviously, the absence of intermittent measurements between year 5 and 11 may somehow weaken the long-term prediction process, and we need to make certain assumptions such that the rate of decline after booster administration at month 60 is similar to the rate of decline after time of last vaccination. It was nevertheless reasurring to see that predicted values at year 12 were all in good agreement, independently of the model or method chosen. All in all, we see that each of these two models may bring

their own insight into the data and their combined use may better serve the purpose of a sensitivity analysis.

In conclusion, we suggest that linear mixed models are considered as a viable alternative to analyze data from vaccination evaluation studies.

Chapter 6

Estimating Reliability Using Non-Linear Mixed Models With Repeated Binary Data

6.1 Introduction

Measurement in psychiatric health sciences seldomly relies on objective criteria. The subjective nature of the information to be gathered renders the development of scales in this area far from easy. One difficulty is that external conditions can influence the response that is given on such a scale like, for example, the person who administers the test or the time of measurement. Therefore, whenever a mental health measurement scale is developed, its psychometric properties are typically checked. An important property in this respect is *reliability*, which reflects the amount of error inherent in any measurement and hence, in a general sense, how replication of the administration would give a different result (Streiner and Norman, 1995).

In classical test theory, the outcome of a test is modeled as

$$X = \tau + \varepsilon, \tag{6.1}$$

where X represents an observation or measurement, τ is the true score and ε the corresponding measurement error. It is assumed that the measurement errors are mutually uncorrelated as well as with the true scores.

The reliability of a measuring instrument is defined as the ratio of the true score variance to the observed score variance, i.e.,

$$R = \frac{\text{var}(\tau)}{\text{var}(X)} = \frac{\text{var}(\tau)}{\text{var}(\tau) + \text{var}(\varepsilon)}. \tag{6.2}$$

In practice, repeated measures are needed to disentangle these two components of variability. Thus, in the case of two parallel measurements, we have $X_1 = \tau + \varepsilon_1$ and $X_2 = \tau + \varepsilon_2$ with ε_1 and ε_2 independent; hence

$$\text{corr}(X_1, X_2) = \frac{\text{var}(\tau)}{\text{var}(\tau) + \text{var}(\varepsilon)} = R. \tag{6.3}$$

In other words, the reliability of an instrument is merely the correlation between independent replications of measuring the same subject. The outcomes X_1 and X_2 can, for example, be two subscores of a test, in which case we are also referring to *split-halve reliability*. If the scores are two measurements of the same instrument, measured at different moments in time, then we are dealing with *test-retest reliability*. When the scores are obtained by two different raters, at one moment in time, then the measure is called *inter-rater reliability*.

There are also situations (e.g. a clinical trial) where repeated or longitudinal measures have been taken, rather than a single measure or pair of measures. This was exemplified in Section 3.6 using data from five clinical trials where repeated measures on the PANSS and CGI scales were planned over the course of the study period. When the rating scale is continuous or quasi-continuous (e.g. PANSS), reliability can be assessed by taking full advantage of the modeling power of linear mixed models. This includes correcting for important covariate effects and allowing for a possibly complicated covariance structure between measurements. Unlike in the classical approach, reliability is no longer summarized by a single figure but rather as a time-dependent function (Laenen *et al.*, 2002).

The objective in this chapter is to utilize related methodology when the rating scale is binary (e.g. dichotomized CGI) rather than (quasi-)continuous. This raises a number of issues. Firstly, estimation of reliability through formula (6.3) is not straightforward and we shall see in Section 6.2 how this can be accomplished in the context of generalized linear mixed models (GLMMs). Secondly, formulation of a random-effects model that acknowledges for the presence of autocorrelation is not trivial with repeated binary data. Standard GLMMs assume that random effects are the unique source to account for residual correlation among responses and we are not

aware of a likelihood-based model formulation that can further handle autocorrelation. In Section 6.3, we propose an extension of the latent variable approach described in Chapter 3, where the (latent) residual terms are assumed to be realizations of a Gaussian process. This artifice allows to introduce some form of autocorrelation in the model, albeit on the latent variable scale, and to assess reliability in a simple manner.

6.2 Estimating Reliability in Generalized Linear Mixed Models

If we adopt the view that reliability is nothing else but a measure of correlation between pairs of measurements, then we must derive the marginal variance-covariance matrix of an outcome vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$. We shall now examine how this can be done in GLMMs and then turn to the specific case of the probit model where simplifications occur.

6.2.1 General Model

The model can be specified as follows:

$$g(E[y_{ij}|\boldsymbol{u}_i]) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_i, \tag{6.4}$$

where y_{ij} is the binary outcome at the jth occasion for subject i, $\boldsymbol{x}_{ij}$ is a set of explanatory variables, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \Sigma_u)$ and $g(.)$ is a preset link function. Furthermore, conditionally on the random effects $\boldsymbol{u}_i$, the y_{ij}'s are assumed to be independent Bernoulli variables.

The marginal covariance matrix of $\boldsymbol{y}_i$ is

$$V_i = \text{cov}[E(\boldsymbol{y}_i|\boldsymbol{u}_i)] + E[\text{cov}(\boldsymbol{y}_i|\boldsymbol{u}_i)] \tag{6.5}$$

with j, k element

$$\begin{aligned}[V_i]_{jk} = &\int (f_{ij} - \pi_{ij})(f_{ik} - \pi_{ik})\phi(\boldsymbol{u}_i; \Sigma_u)d\boldsymbol{u}_i \\ &+ I(j = k)\int f_{ij}(1 - f_{ij})\phi(\boldsymbol{u}_i; \Sigma_u)d\boldsymbol{u}_i,\end{aligned} \tag{6.6}$$

where $f_{ij} = g^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_i)$, $\pi_{ij} = E[y_{ij}] = \int f_{ij}\phi(\boldsymbol{u}_i; \Sigma_u)d\boldsymbol{u}_i$, and $\phi(\boldsymbol{u}; \Sigma_u)$ denotes the density function of a $N(\boldsymbol{0}, \Sigma_u)$ random variable.

Except in some special cases (identity link), numerical integration must be undertaken to evaluate (6.6). It is easy to check that

$$[V_i]_{jj} = \pi_{ij}(1 - \pi_{ij}), \qquad j = 1, \ldots, n_i$$

and

$$[V_i]_{jk} = \int f_{ij} f_{ik} \phi(\boldsymbol{u}_i) d\boldsymbol{u}_i - \pi_{ij}\pi_{ik}, \qquad j \neq k,$$

so the evaluation of $n_i(n_i + 1)/2$ integrals is required.

The problem of calculating $\text{corr}(y_{ij}, y_{ik})$ is, in fact, similar to that of estimating the intraclass correlation coefficient in GLMMs. Goldstein, Browne and Rasbash (2002) discuss four different methods that provide at least approximate estimates of an intraclass correlation measure in generalized linear multilevel models. These are: 1. model linearization by using a first order Taylor expansion; 2. simulation from the fitted model; 3. fitting a binary linear model where the binary response is treated as continuous; and 4. a latent variable approach. We will not further consider the first and third methods which are based on approximations. Although the fourth approach could be followed, the resulting intraclass correlation measure formally stands at the postulated latent variable level and thus is less relevant. We will therefore focus on the simulation method in the rest of this section.

The following procedure can be used to compute $\text{corr}(y_{ij}, y_{ik})$ with the simulation method:

1. From the fitted model (6.4), generate a large number m (say 10000) of values for the random effects $\boldsymbol{u}_i$ from the normal distribution $N(\mathbf{0}, \widehat{\Sigma}_u)$. Let us denote these simulated values by $\boldsymbol{u}_i^{t,\star}$ $(t = 1, \ldots, m)$.

2. Compute the m corresponding values of f_{ij} and f_{ik}, denoted $f_{ij}^{t,\star}$ and $f_{ik}^{t,\star}$ respectively $(t = 1, \ldots, m)$. Let $v_{ij}^{t,\star} = f_{ij}^{t,\star}(1 - f_{ij}^{t,\star})$.

3. Estimate $\text{cov}(y_{ij}, y_{ik})$, $\text{var}(y_{ij})$ and $\text{var}(y_{ik})$ from (6.6) after noting that:

   - the first integral in (6.6) can be approximated by the sample covariance $\widehat{\text{cov}}(f_{ij}^{t,\star}, f_{ik}^{t,\star})$;
   - the second integral in (6.6) can be approximated by the sample mean $\overline{v_{ij}^{t,\star}}$.

Whereas in linear mixed models reliability is a function of the random-effects covariates $\boldsymbol{z}_{ij}$ only, a further difficulty with GLMMs is that it depends on all of

the covariates included in the model, making it complicated to display this measure concisely.

We now illustrate the computations on the schizophrenia data using a logit link specification. A saturated treatment by time model with random intercept and slope is assumed. The model was fitted with the SAS procedure NLMIXED using adaptive Gaussian quadrature with 20 nodes. Table 6.1 presents the values of reliability as a function of treatment group and measurement occasions. It shows that reliability decreases with increasing time lag and is larger at later measurement occasions. Also, estimates are somewhat higher in the risperidone group.

Table 6.1. *CGI: Reliability as a function of treatment group and measurement occasions (logit link). The model includes a random intercept and a random slope.*

	Active Control					***Risperidone***				
Time	**1**	**2**	**4**	**6**	**8**	**1**	**2**	**4**	**6**	**8**
1		0.529	0.426	0.333	0.276		0.593	0.504	0.416	0.348
2			0.567	0.479	0.418			0.592	0.520	0.459
4				0.654	0.608				0.675	0.637
6					0.741					0.745
8										

6.2.2 Probit Model

Under a probit link specification, computation of $\text{corr}(y_{ij}, y_{ik})$ greatly simplifies, after noting that

$$\pi_{ij} = P(y_{ij} = 1) = \Phi\left(\frac{\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1 + \boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ij}}}\right), \tag{6.7}$$

and

$$\begin{aligned}&\int f_{ij}f_{ik}\phi(\boldsymbol{u}_i)d\boldsymbol{u}_i = P(y_{ij} = 1, y_{ik} = 1)\\ &= \Phi_2\left(\frac{\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{1 + \boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ij}}}, \frac{\boldsymbol{x}_{ik}^T\boldsymbol{\beta}}{\sqrt{1 + \boldsymbol{z}_{ik}^T\Sigma_u\boldsymbol{z}_{ik}}}; \frac{\boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ik}}{\sqrt{1 + \boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ij}}\sqrt{1 + \boldsymbol{z}_{ik}^T\Sigma_u\boldsymbol{z}_{ik}}}\right).\end{aligned} \tag{6.8}$$

These equalities are easy to verify in the latent variable model that can (at least conceptually) be associated with model (6.4). See Section 3.3 for further details.

As an illustration, we fit the same model as above, using the probit link instead of the logit. Values of the log likelihood are almost identical in both cases (logit: -2ℓ=2839.7; probit: -2ℓ=2839.4) so the two models provide comparable fits based on this criterion. Table 6.2 presents values of reliability as a function of treatment group and measurement occasions. As can be seen, they are close to those presented in Table 6.1.

Table 6.2. *CGI: Reliability as a function of treatment group and measurement occasions (probit link). The model includes a random intercept and a random slope.*

	Active Control					***Risperidone***				
Time	**1**	**2**	**4**	**6**	**8**	**1**	**2**	**4**	**6**	**8**
1		0.527	0.423	0.328	0.270		0.587	0.496	0.407	0.339
2			0.563	0.476	0.415			0.588	0.516	0.456
4				0.656	0.611				0.674	0.638
6					0.744					0.748
8										

6.3 Estimating Reliability in the Probit Model with Autocorrelation

Models discussed in the previous section attempt to explain the source of residual correlation, after correcting for covariate effects, exclusively by specifying random effects in the model. Clearly, just as it might be in linear models, the conditional independence assumption may be untenable and therefore, we would like to have a model which explicitly recognizes the possibility of an autocorrelation structure. This might also afford an alternative way of parameterizing our model, perhaps more parsimonously.

While this can be accomodated without great difficulty in linear mixed models by assuming a certain autocorrelation structure on the residuals, it is not obvious how

to proceed in the context of non-linear models with binary responses. When there is a small number of fixed occasions, a simple solution to introduce dependence among the binary responses after conditioning on the covariates and random effects is to consider a multilevel multivariate model in which each individual's response sequence is treated as a multivariate outcome vector of fixed length. Yang, Heath and Goldstein (2000), for instance, use this artifice to analyze longitudinal binary responses from a panel study of voting intentions in Great Britain. The advantage of this approach is that the dependence between the responses is modeled by the covariance structure at the individual level (i.e. level 2) rather than at the measurement occasion level (i.e. level 1).

When measurement occasions are not fixed, Barbosa and Goldstein (2000) propose to extend the standard multilevel model for binary outcomes by allowing the (level 1) residuals to be correlated. More precisely, they write the covariance between residuals for individual i at occasions j and k

$$\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})}f(|t_{ij}-t_{ik}|),$$

where the conditional mean $\pi_{ij} = E[y_{ij}|\boldsymbol{u}_i]$ is modeled by (6.4) and $f(s) = \alpha + \exp(-h(s))$. The PQL algorithm can be employed to fit this model and its implementation follows that for continuously distributed responses as explained in Goldstein, Healy and Rasbash (1994).

A drawback with the proposal of Barbosa and Goldstein (2000) is that it corrects for autocorrelation in an *ad hoc* way and thus is outside the likelihood framework. We are actually not aware of a general likelihood-based model formulation that addresses this problem. The model described hereafter is such an attempt, albeit restricted to probit link specification.

6.3.1 The Model

The model that we propose for repeated binary data has many similarities with the model discussed by Heagerty and Lele (1998), which deals with binary spatial data. A random-effects structure can be incorporated as in standard GLMMs and the model further acknowledges for a form of autocorrelation, albeit not directly on the observed responses.

As in Chapter 3, the model is introduced from a latent variable perspective:

$$\tilde{y}_{ij} = \boldsymbol{x}_{ij}^{T}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{T}\boldsymbol{u}_j + \tilde{\varepsilon}_{ij}. \tag{6.9}$$

The residual error terms $\tilde{\varepsilon}_{ij}$ are no longer assumed to be mutually independent, however. Instead, we suppose that they are realizations from a Gaussian stationary process $\tilde{\varepsilon}(t)$ with autocorrelation function

$$\text{corr}(\tilde{\varepsilon}(s), \tilde{\varepsilon}(t)) = \rho(|t-s|).$$

Following Goldstein, Healy and Rasbash (1994), we shall assume that $\rho(u) = \exp(-g(u))$, where $g(u)$ is a positive increasing function, not necessarily linear. For example, one can have $g(u) = \alpha u$ (exponential model), $g(u) = \alpha u^2$ (Gaussian model) or, more generally, $g(u) = \sum_k \alpha_k u^k$ for any polynomial constrained to take on positive values on $[0, +\infty[$. As pointed out by Goldstein, Healy and Rasbash (1994), a difficulty when $g(u)$ is a polynomial is that successive powers tend to be highly correlated and this may cause estimation difficulties. Another possible choice is then to add an inverse polynomial term such as in $g(u) = \alpha_1 u + \alpha_2 u^{-1}$, which avoids the high correlations associated with the ordinary polynomial. Another useful extension is to make the parameters α_k explicitly dependent on some explanatory variables as, for example, with $g(u) = (\alpha_0 + \alpha_1 z_{ij})u$. As to the choice of the function g, Goldstein, Healy and Rasbash (1994) state that it should "contain as few parameters as necessary to be flexible enough to describe real data. [...] There seems to be little substantive guidance on choice, and it is likely that different functional forms will be appropriate for different kinds of data."

6.3.2 Model Estimation

The log likelihood for the observed (binary) data can be written

$$\sum_{i=1}^{N} \sum_{a_{i1},\ldots,a_{in_i}=0}^{1} \delta_{a_{i1},\ldots,a_{in_i}} \log \int P[y_{i1} = a_{i1}, \ldots, y_{in_i} = a_{in_i} | \boldsymbol{u}_i] \phi(\boldsymbol{u}_i) d\boldsymbol{u}_i, \tag{6.10}$$

with

$$\delta_{a_{i1},\ldots,a_{in_i}} = \begin{cases} = 1 & \text{if } y_{i1} = a_{i1}, \ldots, y_{in_i} = a_{in_i}; \\ = 0 & \text{otherwise.} \end{cases}$$

This expression entails the evaluation of multivariate normal probabilities. For instance, we have

$$\begin{aligned} P[y_{i1} = 1, \ldots, y_{in_i} = 1 | \boldsymbol{u}_i] &= P[\tilde{y}_{i1} > 0, \ldots, \tilde{y}_{in_i} > 0 | \boldsymbol{u}_i] \\ &= \int_{-\infty}^{\xi_{i1}} \ldots \int_{-\infty}^{\xi_{in_i}} \phi\big(x_1, \ldots, x_{n_i}; R(\boldsymbol{t}_i)\big) dx_1 \ldots dx_{n_i}, \end{aligned}$$

where $\xi_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + z_{ij}^T\boldsymbol{u}_i$ and the matrix $R(\boldsymbol{t}_i)$ has jkth element equal to $\rho(|t_{ij} - t_{ik}|)$.

Unless the number of measurements is very small ($n_i = 2$), the log likelihood is cumbersome to evaluate and therefore, an alternative estimation method would be desirable. We propose, again, to use maximum pairwise likelihood (MPL). As a matter of fact, the MPL estimation procedure described in Chapter 3 requires only minor modifications in the computation of bivariate marginal probabilities to be applicable in the present context. For instance, previously we had

$$P[y_{ij} = 0, y_{ik} = 0] = \Phi_2\left(\xi_{ij}, \xi_{ik}; \rho_{ijk}\right)$$

with

$$\xi_{ij} = \frac{-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}{\sqrt{\text{var}[\tilde{y}_{ij}]}}$$

and

$$\rho_{ijk} = \frac{\boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ik}}{\sqrt{1 + \boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ij}}\sqrt{1 + \boldsymbol{z}_{ik}^T\Sigma_u\boldsymbol{z}_{ik}}},$$

whereas for the new model, this needs to be updated as

$$\rho_{ijk} = \frac{\boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ik} + \rho(|t_{ij} - t_{ik}|)}{\sqrt{1 + \boldsymbol{z}_{ij}^T\Sigma_u\boldsymbol{z}_{ij}}\sqrt{1 + \boldsymbol{z}_{ik}^T\Sigma_u\boldsymbol{z}_{ik}}}.$$

Apart from such minor modifications, the whole estimation procedure remains unchanged.

6.3.3 Simulations

To assess the finite sample properties of MPL with the proposed model, we conducted a small simulation study using a longitudinal model with random intercept $N(0, \sigma_u^2)$ and time as sole covariate. Different numbers of measurement occasions ($n_i = 5$, 10, 20) and individuals ($N = 100$, 500, 1000) were considered. In each case, measurement times were taken to be equally spaced on the interval $[0, 4]$. An exponential model for the autocorrelation structure was assumed. Note that the autocorrelation parameter α was log transformed to avoid constrained optimization.

Results are presented in Tables 6.3–6.5 based on 100 simulated data sets under each scenario. For each parameter the mean, simulation S.D. and mean estimated S.D. are reported, as well as the relative bias compared to the true value assumed by

Table 6.3. *MPL Estimates based on 100 simulations ($N = 100$ subjects).*

Parameter	Model	Estimate (mean)	Relative bias (%)	Simulation S.D.	Estimated S.D. (mean)
$\boldsymbol{n_i = 5}$ (Converged = 70)					
β_0	-2	-2.655	-32.8	0.745	1.032
β_1	0.5	0.665	33.0	0.186	0.256
σ_u	1	1.551	55.1	0.622	1.161
α	0.25	0.729	191.6	0.599	0.870
$\boldsymbol{n_i = 10}$ (Converged = 70)					
β_0	-2	-2.584	-29.2	0.870	0.824
β_1	0.5	0.648	29.6	0.217	0.199
σ_u	1	1.467	46.7	0.699	0.856
α	0.25	0.525	110.0	0.394	0.355
$\boldsymbol{n_i = 20}$ (Converged = 74)					
β_0	-2	-2.466	-23.3	0.732	0.830
β_1	0.5	0.619	23.8	0.182	0.202
σ_u	1	1.371	37.1	0.625	0.865
α	0.25	0.448	79.2	0.271	0.295

the model. Standard error calculations for the mean and (transformed) dependence parameters were based on the estimator (3.16) in Chapter 3. Precision estimates for α were then obtained using the delta method.

These simulations indicate that the mean and dependence parameters are strongly biased with a small number of subjects ($N = 100$), the parameter α being mostly affected. Increasing the number of measurements somewhat reduces the extent of bias. With a medium number of subjects ($N = 500$), parameters are still largely biased when the number of measurement occasions is small ($n_i = 5$) but the bias falls within more acceptable limits with an increased number of measurement occasions. The parameter α is still noticeably biased, though. With a large number of subjects ($N = 1000$), the bias for the mean parameters and the parameter σ_u becomes small but for the autocorrelation parameter α it is still sizeable with data sets containing as many as 20,000 observations. Restricting attention to precision estimates, we see that they tend to slightly overestimate the sampling variability, especially for the variance parameter σ_u.

Also reported in each table are the numbers of simulated data sets where conver-

Table 6.4. *MPL Estimates based on 100 simulations ($N = 500$ subjects).*

Parameter	Model	Estimate (mean)	Relative bias (%)	Simulation S.D.	Estimated S.D. (mean)
$\boldsymbol{n_i = 5}$ (Converged = 86)					
β_0	-2	-2.266	-13.3	0.470	0.542
β_1	0.5	0.571	14.2	0.122	0.136
σ_u	1	1.224	22.4	0.436	0.582
α	0.25	0.389	55.6	0.202	0.219
$\boldsymbol{n_i = 10}$ (Converged = 88)					
β_0	-2	-2.177	-8.9	0.391	0.418
β_1	0.5	0.547	9.4	0.098	0.105
σ_u	1	1.159	15.9	0.372	0.436
α	0.25	0.322	28.8	0.123	0.132
$\boldsymbol{n_i = 20}$ (Converged = 86)					
β_0	-2	-2.135	-6.8	0.405	0.461
β_1	0.5	0.535	7.0	0.101	0.115
σ_u	1	1.073	7.3	0.410	0.637
α	0.25	0.297	18.8	0.118	0.132

Table 6.5. *MPL Estimates based on 100 simulations ($N = 1000$ subjects).*

Parameter	Model	Estimate (mean)	Relative bias (%)	Simulation S.D.	Estimated S.D. (mean)
$\boldsymbol{n_i = 5}$ (Converged = 89)					
β_0	-2	-2.068	-3.4	0.328	0.424
β_1	0.5	0.516	3.2	0.081	0.106
σ_u	1	1.041	4.1	0.327	0.486
α	0.25	0.288	15.2	0.107	0.137
$\boldsymbol{n_i = 10}$ (Converged = 90)					
β_0	-2	-2.087	-4.4	0.311	0.354
β_1	0.5	0.518	3.6	0.077	0.088
σ_u	1	1.055	5.5	0.308	0.375
α	0.25	0.281	12.4	0.099	0.104
$\boldsymbol{n_i = 20}$ (Converged = 86)					
β_0	-2	-2.037	-1.9	0.293	0.321
β_1	0.5	0.507	1.4	0.073	0.080
σ_u	1	1.016	1.6	0.288	0.352
α	0.25	0.271	8.4	0.086	0.091

gence could be achieved. Several types of convergence failure can be detected. As usual, this might be due to a failure of the optimization procedure to converge. Also, some problems arise when the parameter σ_u converges to the boundary value 0 or when the parameter α grows large (which means no autocorrelation according to the exponential model), and we call these convergence failures too. In practice, they can easily be traced since the hessian matrix is near-singular and yields very large values of variance estimates. All cases of convergence failures were ignored in the computation of summary statistics shown in Tables 6.3–6.5. From the simulations reported here, number of subjects appears to be the most determining factor to reduce convergence difficulties.

6.3.4 Application to the Schizophrenia Data

We take a saturated treatment by time model for the mean structure and a random intercept. For the autocorrelation structure, we assumed that $g(u) = \alpha u^{\gamma}$ and tried several values of γ (=-1, 0.5, 1, 2). The exponential model ($\gamma = 1$) provided the best fit. We also fitted a model with $g(u) = \alpha_1 u + \alpha_2 u^{-1}$ but $\widehat{\sigma}_u$ converged to the boundary value 0.

In Table 6.6, parameter estimates and standard errors are reported for the random-intercept model and the model with random intercept and exponential autocorrelation structure. As can be seen, parameter estimates for the model with autocorrelation are all attenuated by an amount of around 30%. This is largely due to the fact that the residual error terms in (6.9) are allowed to be autocorrelated. The $\log PL$ value for this model shows an improvement in the fit of the model. In comparison, the $\log PL$ value for the model with random intercept and random slope is -1714.8, which shows yet a bigger improvement over the random-intercept model, although at the cost of 2 extra parameters. We also tried to add an exponential autocorrelation structure to the random-intercept-and-slope model but the algorithm converged to a solution with non positive-definite matrix $\widehat{\Sigma}_u$.

Finally, reliability can be easily estimated using formulas (6.7) and (6.8), by adjusting the latter as indicated in Section 6.3.2. Table 6.7 presents the values of reliability as a function of treatment group and measurement occasions for the random-intercept model with exponential autocorrelation structure. Compared to Table 6.2, these estimates tend to be higher at early measurement occasions and smaller at later measurement occasions.

Table 6.6. *Schizophrenia data: MPL parameter estimates and standard errors for the random-intercept model with and without autocorrelation. The exponential model was taken for the autocorrelation structure.*

	Random Intercept		**Random Intercept + Autocorrelation**	
Parameter	**Estimate**	**S.E.**	**Estimate**	**S.E.**
Intercept	-0.269	0.158	-0.179	0.120
Week1	-1.881	0.175	-1.343	0.197
Week2	-1.165	0.166	-0.878	0.158
Week4	-0.702	0.157	-0.519	0.132
Week6	-0.209	0.144	-0.155	0.109
Treat×Week1	0.288	0.211	0.190	0.152
Treat×Week2	0.581	0.207	0.425	0.155
Treat×Week4	0.542	0.208	0.388	0.156
Treat×Week6	0.333	0.315	0.236	0.160
Treat×Week8	0.198	0.222	0.139	0.168
σ	1.830	0.108	1.117	0.231
$\log \alpha$			-1.338	0.332
$\log PL$	-1727.0		-1722.2	

Coding for 'Treat': 0 = active control, 1 = risperidone.

6.4 Conclusions

In this chapter we have focused on situations where clinical trial data with repeated measures have been collected on a particular measuring instrument and interest centers on assessing its reliability. If the rating scale is continuous or quasi-continuous, reliability can be estimated by taking full advantage of the modeling power of linear mixed models. Unlike in the classical approach, it is no longer summarized by a single figure but rather as a time-dependent function.

When repeated binary data have been gathered, we may want to employ related methodology for examining reliability of the instrument. Our first goal was therefore to clarify how reliability can be estimated when a standard generalized linear mixed model has been fitted to the data. A difficulty is that reliability is then a function of the covariates incorporated in the model and hence is much more awkward to

Table 6.7. *CGI: Reliability as a function of treatment group and measurement occasions for the random-intercept model with exponential autocorrelation structure.*

	Active Control					***Risperidone***				
Time	**1**	**2**	**4**	**6**	**8**	**1**	**2**	**4**	**6**	**8**
1		0.642	0.467	0.365	0.339		0.699	0.544	0.471	0.430
2			0.580	0.447	0.402			0.609	0.500	0.444
4				0.577	0.481				0.607	0.499
6					0.609					0.609
8										

summarize. We have also shown that reliability is appreciably easier to estimate when a probit link specification is assumed since only simple probit functions must be evaluated in this case.

As a complementary manner of accounting for the dependence among the binary responses, it might be useful to introduce an autocorrelation structure in a generalized linear model with random effects. To that end, we adopted a latent variable perspective where the residual error terms are assumed to be realizations of a Gaussian process, which constraints the link function to be probit. This afforded us a way to introduce a source of autocorrelation in the model, albeit on the latent variable scale.

Since the likelihood function is cumbersome to evaluate, we proposed MPL for estimation purposes as it involves only minor modifications compared to the procedure discussed in Chapter 3. The series of simulations presented in Section 6.3.3 suggest that in practice, a very large sample size might be necessary for the asymptotic properties of the MPL estimator to hold and that autocorrelation parameters may be subject to substantial bias. This warrants further research on alternative estimation methods or on computationally efficient approaches to evaluate the likelihood.

Finally, convergence difficulties should be anticipated to occur quite frequently in practical applications. Even in linear mixed models, convergence failures are relatively common when modeling of the covariance structure involves joint specification of random effects and a term of serial correlation, simply because these two variability components cannot easily be disentangled. Obviously, this phenomenon should be amplified with binary data, which convey less information.

Chapter 7

Validation of a Longitudinally Measured Surrogate Marker for a Time-to-Event Endpoint

7.1 Introduction

In recent years, interest in modeling the relationship between a time-to-event endpoint and longitudinally measured data has developed considerably. This problem occurs naturally in many biomedical or public health studies where participants are followed over time. In such studies, measurements on a number of outcomes can be obtained at different occasions and times to certain clinical events observed as well.

In randomized clinical trials the main question is often whether a new treatment has some beneficial effect on the time to a specific clinical event, the endpoint of primary interest. The time elapsed between randomization and this event, however, can be very long and it may therefore be desirable to find a surrogate for the clinical outcome of interest that is less distant in time, thereby permitting a trial to be completed sooner and making a potentially useful treatment available earlier to a wider range of patients. A well-known example is in AIDS research where an early proposal of surrogate marker for clinical outcomes such as disease progression or survival was the number of CD4 T-lymphocytes (see e.g. De Gruttola *et al.*, 1993; Choi *et al.*, 1993; Tsiatis, De Gruttola and Wulfsohn, 1995).

The objective here is to extend the methodology of Buyse *et al.* (2000), which was introduced in Section 4.2, to the case of a biomarker measured repeatedly over time and a time-to-event endpoint. Technically, a joint model for longitudinal measurements and event time data is required. Research on this topic has received substantial attention over recent years and some useful references include Pawitan and Self (1993), De Gruttola and Tu (1994), Taylor, Cumberland and Sy (1994), Faucett and Thomas (1996), Lavalley and De Gruttola (1996), Hogan and Laird (1997), Wulfsohn and Tsiatis (1997), Henderson, Diggle and Dobson (2000) and Xu and Zeger (2001). The model of Henderson, Diggle and Dobson (2000) will be adopted here. Their approach assumes standard models for the longitudinal and event time data and postulates a latent bivariate Gaussian process inducing stochastic dependence between the measurement and event processes.

The chapter is organized as follows. Section 7.2 introduces the motivating example which involves a set of two randomized clinical trials in advanced prostate cancer and where we seek to evaluate the usefulness of prostate-specific antigen (PSA) level as a surrogate for survival. Section 7.3 shows how the methodology of Buyse *et al.* (2000) can be adapted to the case of a longitunally measured marker and a time-to-event endpoint. The methodology is then applied to the prostate cancer data in Section 7.4. Note that the results presented in this chapter can be found in Renard *et al.* (2002b).

7.2 Motivating Study

We consider a set of two open-label multicenter clinical trials in which patients with advanced prostate cancer were randomized either to oral liarozole, an experimental retinoic acid metabolism-blocking agent developed by Janssen Research Foundation, or to an antiandrogenic drug: cyproterone acetate (CPA) in the first trial (Debruyne et al., 1998) and flutamide in the second. The two trials accrued 312 and 284 patients in centers spread over 9 and 10 countries, respectively. All patients were in relapse after first-line endocrine therapy.

The primary endpoint in each trial was survival time after randomization. Assessments were undertaken before the start of treatment and repeated at 2 weeks, monthly for six months and every three months thereafter, until patients show clinical progression or develop a serious adverse event. All patients were then followed up until death. The assessments included measurement of prostate-specific antigen (PSA) level. PSA is a glycoprotein that is found almost exclusively in normal and

neoplastic prostate cells. Serum PSA usually rise in men who have prostate cancer, but also with some infections of the prostate or non-malignant diseases such as benign prostatic hyperplasia. As a consequence, changes in PSA often antedate changes in bone scan, and they have been used as a response indicator in patients with androgen-independent prostate cancer (Kelly et al., 1993; Sridhara et al., 1995; Smith et al., 1998). It is therefore of interest to study more formally to which extent a sequence of PSA measurements can be a valuable surrogate for a patient's survival.

Figure 7.1 shows plots of the individual log-transformed PSA profiles. To avoid overly cluttered plots, profiles were shadowed and 30 randomly chosen subjects are depicted using darker lines. As can be seen, the length of the individual sequences of PSA measurements is highly variable accross patients, with only a few individuals having very long sequences. Figure 7.2 displays PSA and survival summaries for each trial. The (log-transformed) PSA data were smoothed using the LOESS technique (Cleveland, 1979); the survival curves were obtained using the Kaplan-Meier estimator (Kaplan and Meier, 1958). Notice the scatter of points in the left-hand plots: most of the subjects had their PSA measurements taken within the first few months after treatment randomization.

To further investigate the effect of "dropout" induced by patients being taken off study upon clinical progression, we plotted the mean profiles per dropout pattern according to visits as they were planned in the protocol (thus not using the exact date of PSA measurement). This is shown in Figure 7.3 for the combined data from the two trials, with the label "control" referring to CPA/flutamide and "experimental" to liarozole. Late-dropout patterns were not represented because of the scarcity of data after 1.5 year. Noticeable in the plot is that patients who progressed early tend to have a higher initial PSA value and do not exhibit an early decline in their PSA level. Also, the mean PSA evolution among subjects who progressed belatedly can be contrasted with the relatively flat curves displayed in Figure 7.2.

7.3 Modeling Approach

As explained in Section 4.2, Buyse *et al.* (2000) examine the surrogacy issue in a meta-analytic context at two distinct levels. At the trial level, the surrogacy measure can be approached from two different modeling perspectives: one is to fit the random-effects model (4.5) from which R^2_{trial} can be immediately obtained *via* expression (4.6); the other is to fit the fixed-effects model (4.1) and, in a second stage, to estimate R^2_{trial}

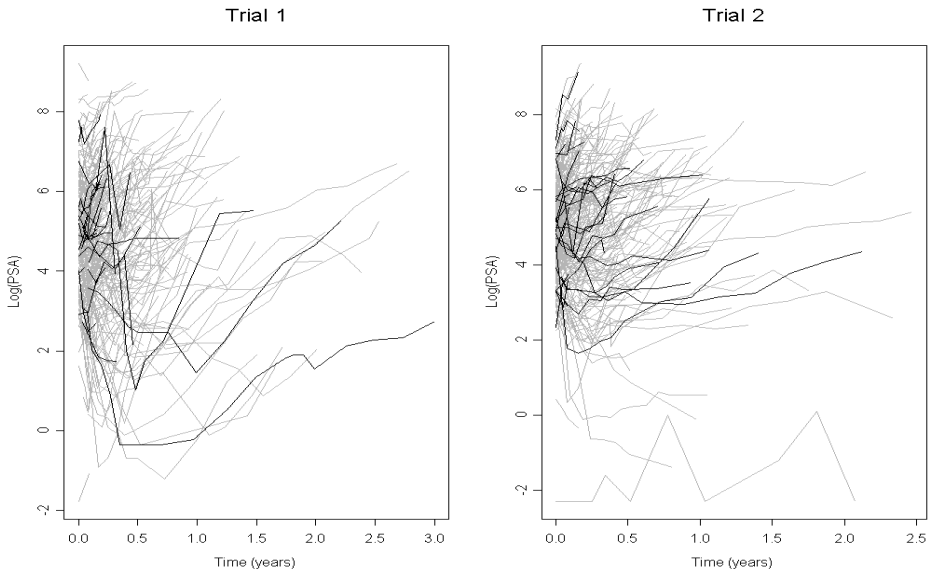


Figure 7.1. *Individual log-transformed PSA profiles for the liarozole trials (30 randomly chosen subjects are plotted using darker lines).*

as the coefficient of determination from the linear regression model

$$\hat{\beta}_i = \lambda_0 + \lambda_1 \hat{\mu}_{S_i} + \lambda_2 \hat{\alpha}_i + \varepsilon_i.$$

In either case, the association at the individual level (R^2_{indiv}) is a by-product of the joint model. Of course, inference will preferably be based on (4.5) but this model is difficult to fit in practice and does not go without numerous convergence problems (see Section 4.2.5). Whereas emphasis in Chapter 4 was on random-effects modeling, we will focus on the two-stage approach here.

To extend the methodology, a joint model for longitudinal measurements and event time data is required. To that end, we consider the model proposed by Henderson, Diggle and Dobson (2000). We will follow their notation and thus, we consider a set of N grouping units (trial, center, etc.) with subjects within the ith unit being followed for some time τ_i. The jth subject in unit i provides a set of measurements $\{y_{ijk} : k = 1, \ldots, n_{ij}\}$ at times $\{t_{ijk} : k = 1, \ldots, n_{ij}\}$, together with the realization of a counting process $\{N_{ij}(u) : 0 \leq u \leq \tau_i\}$ for the time-to-event endpoint and a zero-one process $\{H_{ij}(u) : 0 \leq u \leq \tau_i\}$ indicating whether a subject is at risk of

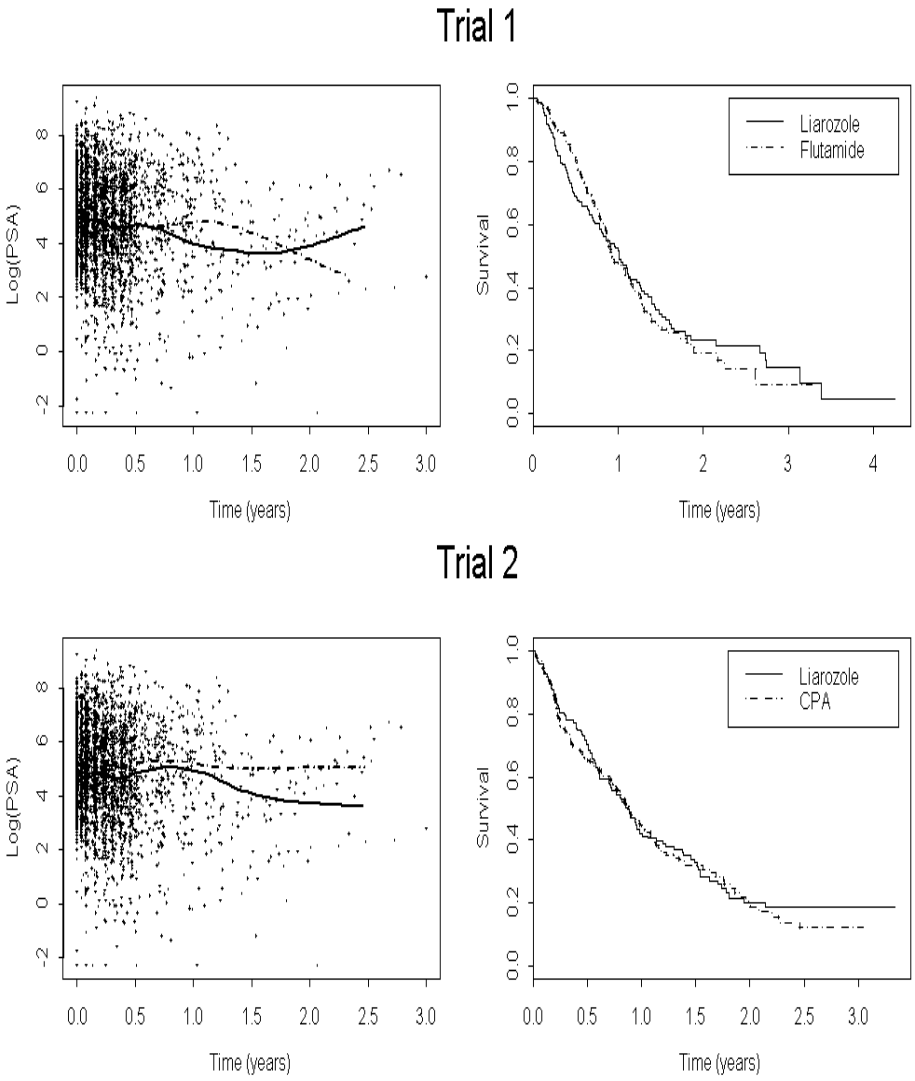


Figure 7.2. *Longitidunal and event time summaries for the liarozole trials (left: smoothed PSA profiles; right: survival curves).*

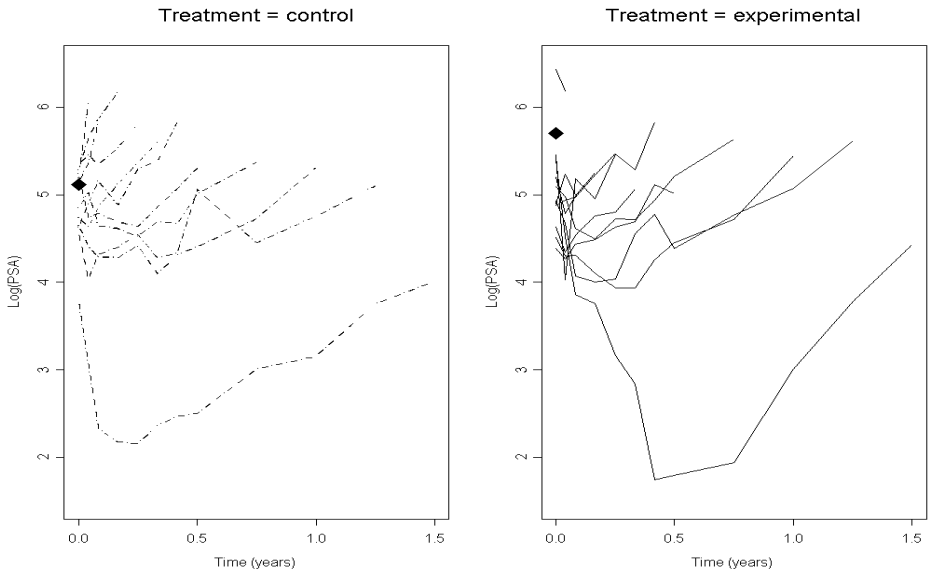


Figure 7.3. *Mean PSA profiles per "dropout" patterns (the black diamonds represent the mean PSA level of those patients who only have a baseline measurement).*

experiencing an event at time u.

A central feature of the model is to postulate an unobserved (latent) zero-mean bivariate Gaussian process, $W_{ij}(t) = \{W_{1ij}(t), W_{2ij}(t)\}$, to describe the association between the longitudinal measurement and event processes. The measurement and intensity models are linked as follows:

(1) The sequence of measurements $\{y_{ijk} : k = 1, \ldots, n_{ij}\}$ of a subject is modeled using a standard linear mixed model, possibly allowing for a serially correlated component:

$$Y_{ijk} = \mu_{ij}(t_{ijk}) + W_{1ij}(t_{ijk}) + \varepsilon_{ijk}, \tag{7.1a}$$

where $\mu_{ij}(t_{ijk})$ describes the mean response profile and $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$ is a sequence of mutually independent measurement errors. We will let $\boldsymbol{\alpha}_i$ denote the vector of parameters for the trial-specific treatment effects used in modeling the mean response profile. Examples will be given in what follows.

(2) The event intensity process is modeled using a semi-parametric model

$$\lambda_{ij}(t) = H_{ij}(t)\lambda_0(t)\exp\{\beta_i Z_{ij} + W_{2ij}(t)\}, \tag{7.1b}$$

where the form of $\lambda_0(t)$ is left unspecified. The parameters β_i represent trial-specific treatment effects on the hazard function.

The specification of W_{1ij} and W_{2ij} can take a variety of forms. As a basic example, suppressing indices for notational simplicity, one could consider $W_1(t) = U_1 + U_2 t$, with (U_1, U_2) being normally distributed with mean zero and covariance matrix G, to specify a model with random intercept and random slope for the longitudinal marker. The $W_2(t)$ process could then include different effects for the initial value (U_1), the slope (U_2) or the current value ($U_1 + U_2 t$) of the marker according to the assumed model, yielding $W_2(t) = \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3(U_1 + U_2 t)$. Inclusion of a frailty component, orthogonal to the measurement process, is also possible if necessary.

Following Henderson, Diggle and Dobson (2000), the expectation maximization (EM) algorithm can be employed to fit the model. Upon convergence of the algorithm, the coefficients of determination R^2_{trial} and R^2_{indiv} can practically be obtained as follows. The inclusion of (fixed) trial-specific coefficients in both the longitudinal measurement and intensity models permits estimation of R^2_{trial}. Unlike in the simpler normal setting which involves solely trial-specific intercepts and treatment effects, the longitudinal measurement model will require, in general, extra terms to model time evolution of the marker. For practical purposes, we will therefore suppose that the mean response profile within each treatment group can be specified parsimoniously, as a low-order (conventional or fractional) polynomial or as a continuous piecewise linear function of time. To illustrate the calculation of R^2_{trial}, suppose that the trajectory of the marker is quadratic over time within each treatment group. Then $\mu_{ij}(t_{ijk})$ can be written

$$\mu_{ij}(t_{ijk}) = \mu_{0i} + \mu_{1i} t_{ijk} + \mu_{2i} t_{ijk}^2 + \alpha_{0i} Z_{ijk} + \alpha_{1i} Z_{ijk} t_{ijk} + \alpha_{2i} Z_{ijk} t_{ijk}^2$$

and R^2_{trial} defined as the coefficient of determination in the regression model

$$\hat{\beta}_i = \lambda_0 + \lambda_1 \hat{\alpha}_{0i} + \lambda_2 \hat{\alpha}_{1i} + \lambda_3 \hat{\alpha}_{2i} + \varepsilon_i.$$

At the individual level it is natural to consider the association between $W_1(t)$ and $W_2(t)$ in the above model. In consequence, R^2_{indiv} will not refer directly to the association between the two endpoints but rather, to the association between the two components of the bivariate latent process which governs the longitudinal and

event processes. This association can no longer be summarized by a single number, however. It will now be a time-dependent measure since the association between the marker and the event process can be defined relative to any time over the course of measurement of the marker. In fact, this could even be extended to the association between the marker, as measured at some time t_1, and the event process defined at a later time $t_2 \geq t_1$, thereby yielding a surface to describe the association between the longitudinal and event processes. This feature can be important in selecting an optimal time at which the marker should be evaluated, either to enhance clinical judgment or even further, to predict the event time of interest.

To illustrate the derivation of $R^2_{\text{indiv}}(t)$, we consider the aforementioned example with $W_1(t) = U_1 + U_2 t$ and $W_2(t) = \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3(U_1 + U_2 t)$. The correlation between $W_1(t)$ and $W_2(t)$, for any fixed time t, can be easily calculated since $W_1(t)$ and $W_2(t)$ have a joint normal distribution. Thus, if $(U_1, U_2) \sim N(0, G)$, we have:

$$\begin{aligned}
\text{var}[W_1(t)] &= G_{11} + 2G_{12}t + G_{22}t^2, \\
\text{var}[W_2(t)] &= (\gamma_1^2 + 2\gamma_1\gamma_3)G_{11} + 2(\gamma_1\gamma_2 + \gamma_1\gamma_3 t + \gamma_2\gamma_3)G_{12} \\
&\quad + (\gamma_2^2 + 2\gamma_2\gamma_3 t)G_{22} + \gamma_3^2 \text{var}[W_1(t)], \\
\text{covar}[W_1(t), W_2(t)] &= \gamma_1 G_{11} + (\gamma_2 + \gamma_1 t)G_{12} + \gamma_2 G_{22} t + \gamma_3 \text{var}[W_1(t)],
\end{aligned}$$

from which the (squared) correlation between $W_1(t)$ and $W_2(t)$ can be easily derived by plugging in estimates for γ_1, γ_2, G_{11}, G_{12} and G_{22}. This function, that will be termed "model-based", is entirely based on the assumptions made in our model. A more heuristic estimate, which we will refer to as "empirical", could be derived along the same lines of development, except that sample estimators based on the expected U values obtained at the final step of the EM algorithm are substituted for the elements of G. Thus, G_{11} is replaced by $\widehat{\text{var}}\{\hat{U}_{1i}\}$, G_{22} by $\widehat{\text{var}}\{\hat{U}_{2i}\}$ and G_{12} by $\widehat{\text{covar}}\{\hat{U}_{1i}, \hat{U}_{2i}\}$.

It should be stressed that the resulting curve is still strongly dependent on some aspects of the model. For example, should we assume that $W_2(t) = \gamma W_1(t)$, then $R^2_{\text{indiv}}(t) \equiv 1$. As one departs from this simple model and further terms are added, a finer characterization of the curve is allowed in its admissible forms. Because of this, we recommend to include a sufficiently large number of association parameters $\{\gamma_k\}$ in the model.

7.4 Application to the Advanced Prostate Cancer Data

In this section, we apply the proposed methodology to the liarozole data introduced in Section 7.2. We will utilize pooled data from the two trials and will refer to control and experimental arms as in Figure 7.3. Since our methodology requires the estimation of the treatment effects in multiple trials or other meaningful groups of patients, we will use country as a grouping unit within each trial in order to have a sufficient number of patients in each unit. This enables us to define 19 groups containing between 3 and 69 patients per group. For the analysis, however, two of these groups had to be excluded: in one of them ($n = 3$), subjects were accrued in only one treatment arm and no events were observed in the second ($n = 8$).

Figure 7.4 depicts summaries of the data in terms of the basic entities connected through model (7.1a)-(7.1b), that is, the sequences of longitudinal measurements on the marker and the hazard function for survival. The bottom figure was obtained by smoothing the Nelson-Aalen estimator of the hazard rate with an Epanechnikov kernel function (Ramlau-Hansen, 1983).

A first step in the analysis is to specify a parsimonious model that captures time evolution of the marker within each treatment group. A simplistic attempt could involve second-order polynomials. While this choice may, at first, seem odd after inspection of the average profiles (top panel in Figure 7.4), this is more in agreement with what Figure 7.3 suggested. This was also confirmed by a likelihood ratio test as the introduction of a quadratic term (with random coefficient) in the model yields a large drop in deviance.

As a possible refinement, we can employ fractional polynomials to better characterize time evolution of the marker. We used the second strategy described in Section 5.4 to select the power vector, while allowing for treatment-specific curves. Starting from the set of powers ranging from -2 to 2 by step of 0.5 and restricting the search over fractional polynomials of degree 2, the values 0.5 and 1 were selected. Note that using the other selection strategy, with given random effects for the intercept, t and t^2, the pair (0.5, 1) also yields one of the best fits. Our final model for the longitudinal PSA measurements therefore includes fixed (treatment-specific) and random effects for the intercept, t and $\sqrt{t}$ [Note: comparison of this final model with the original one also yields a large drop in deviance].

Now that we have chosen a parsimonious description of the temporal evolution of

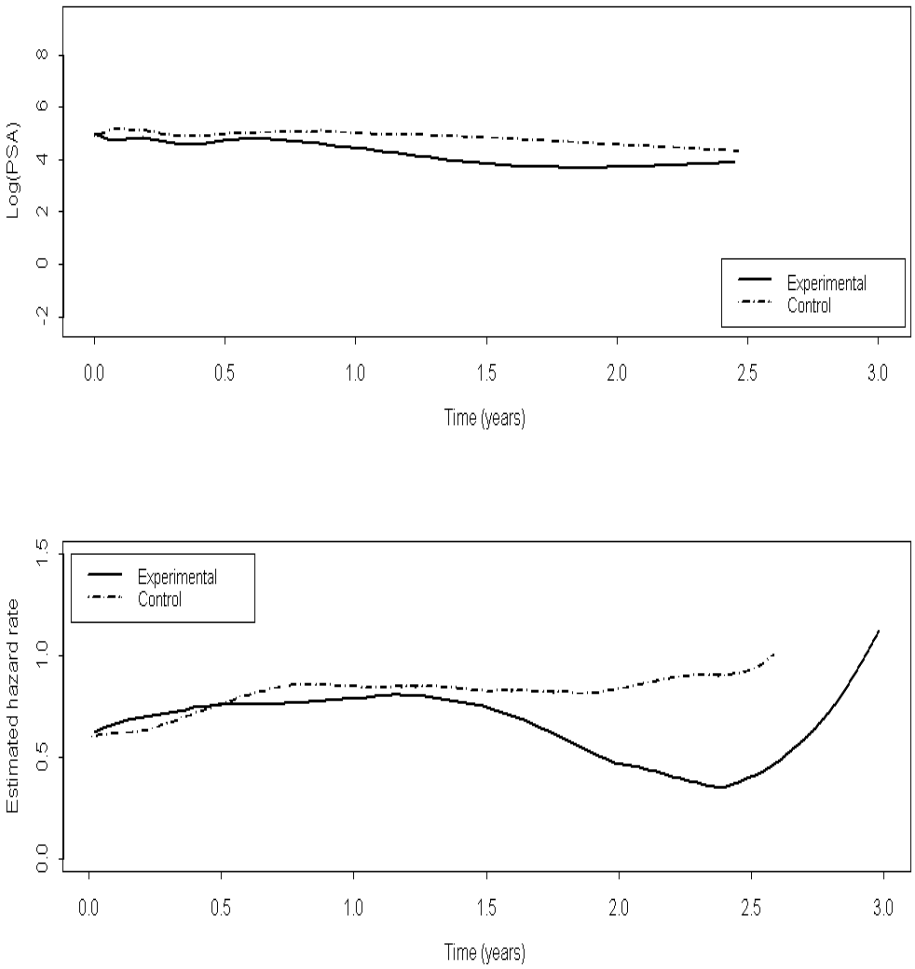


Figure 7.4. *Longitudinal and event time summaries for the combined liarozole trials. Top panel: smoothed log PSA profiles; Bottom panel: smoothed estimates of the hazard rate.*

the marker, the next step is to formulate the joint model for the PSA measurements and survival time data, with trial-specific effects for each term. The model can be written:

$$\begin{aligned} Y_{ijk} &= \mu_{0i} + \mu_{1i} t_{ijk} + \mu_{2i}\sqrt{t_{ijk}} + \alpha_{0i} Z_{ij} + \alpha_{1i} Z_{ij} t_{ijk} + \alpha_{2i} Z_{ij}\sqrt{t_{ijk}} \\ &\quad + U_{0j} + U_{1j} t_{ijk} + U_{2j}\sqrt{t_{ijk}} + \varepsilon_{ijk}, \end{aligned} \tag{7.2a}$$

and

$$\lambda_{ij}(t) = \lambda_0(t)\exp\{\beta_i Z_{ij} + \gamma_0 U_{0j} + \gamma_1 U_{1j} + \gamma_2 U_{2j} + \gamma_3(U_{0j} + U_{1j} t + U_{2j}\sqrt{t})\}, \tag{7.2b}$$

with i denoting country (within trial), j referring to individual patients and k to measurement occasions.

As explained in Section 7.3, R^2_{trial} can be calculated as the coefficient of determination in the regression of $\{\hat{\beta}_i\}$ on $\{\hat{\alpha}_{0i}, \hat{\alpha}_{1i}, \hat{\alpha}_{2i}\}$, which yields a value of 0.517. This mid-range value is presumably too low to permit reliable prediction of treatment effects on survival, having observed the effect of treatment on the marker. Confidence limits on R^2_{trial} can be obtained based on the assumption that $\boldsymbol{\alpha}_i$ and β_i are normally distributed (as in (4.3)), from which the distribution of the coefficient of determination can be derived (Algina, 1999; Ding, 1996). More specifically, a 95% confidence interval can be obtained by finding values of R^2_{trial} for which the corresponding estimates are approximately equal to the 2.5% and 97.5% quantiles of the cumulative distribution function of R^2. In our example, the resulting confidence limits for R^2_{trial} are $[0.013, 0.748]$, thus showing that the trial-level association is estimated rather imprecisely due to the relatively small number of units available to estimate treatment effects.

Remark that dependence between the marker and survival endpoint is a complicating assumption with our methodology. If interest centers on trial-level surrogacy alone, a naive approach might be to assume independence between the two outcomes, which greatly simplifies computations since the two models can then be fitted separately. Tibaldi *et al.* (2002) explore this issue in the case of normally distributed endpoints (Section 4.2) and conclude that simplified computational methods perform quite well. Obviously, as one departs from the multivariate Gaussian framework, it is not at all clear whether such a simplistic approach works effectively well. For comparitive purposes, we calculated R^2_{trial} by fitting separately models (7.2a) and (7.2b) with $\gamma_0 = \gamma_1 = \gamma_2 = \gamma_3 = 0$. This results in a value of $R^2_{\text{trial}} = 0.291$ which is much

lower than the one found above (although confidence limits should not be overlooked). Thus, ignoring dependence between the marker and the survival endpoint might give misleading inference on R^2_{trial} in this setting, although this issue should be further explored.

Figure 7.5a shows the model-based and empirical curves $R^2_{\text{indiv}}(t)$ for model (7.2a)-(7.2b). Both curves agree fairly well over the time range considered. They start from a relatively low level (~ 0.3), then raise sharply until a value of about 0.9 at year 1 and stabilize at that level thereafter. Although the interpretation of such a plot holds, strictly speaking, at the level of the latent processes $W_1(t)$ and $W_2(t)$, this would suggest that initially, PSA level bears relatively little information on a patient's future survival but as information on the marker is gathered over time (mostly within the first year of treatment), it achieves a better capability to predict survival, with no further gain in subsequent years. For comparison purposes, the plot in the right panel (Figure 7.5b) shows the same curves under the model with quadratic time evolution and individual-level random effects for the intercept, t and t^2. The two curves show a similar behavior within the first year after randomization, but then a dip can be observed. Also, it can be noticed that the two curves do not coincide so well. It is not clear whether this is caused by the inferior fit of the model, or by constraints imposed by the model itself, but this calls for caution when interpreting such curves. We do believe that they might shed some light on the basic intricacies between the marker and the survival endpoint under study, but they should not be over-interpreted as they may be strongly model-dependent.

7.5 Conclusions

An extension of the surrogate endpoint validation methodology of Buyse et al. (2000) was proposed for the case where a longitudinally measured biomarker is a potential surrogate for a survival endpoint. To that end, the formulation of Henderson, Diggle and Dobson (2000) was adopted for the joint model relating the marker and the survival time.

A limiting feature arises from the inherent complexity of joint modeling longitudinal measurements and event time data, which is most noticeable in the computational aspect of this approach. As a result, intensive computing times can be expected in the type of applications covered here because of the typically large size of the meta-analytic data sets required for our validation exercise.

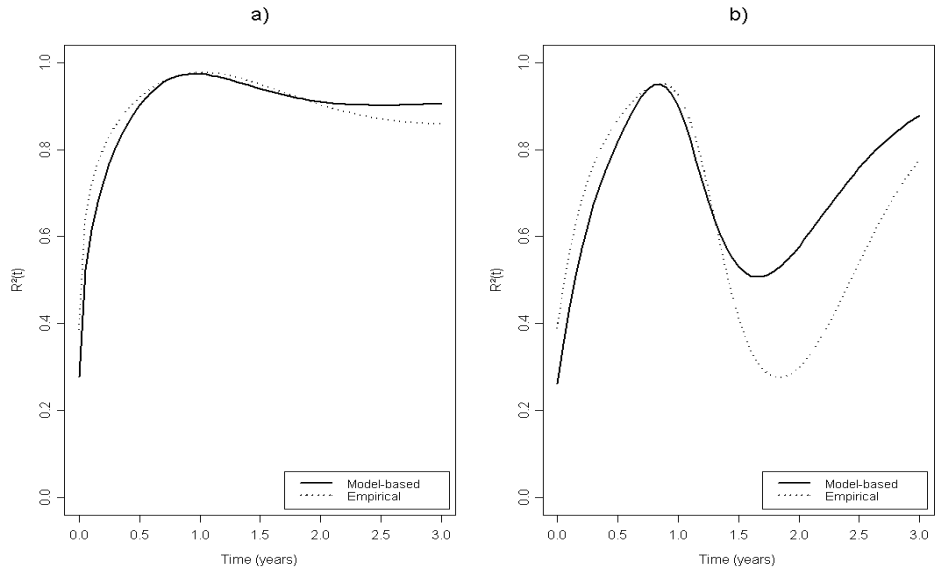


Figure 7.5. *Plots of the model-based and empirical $R^2_{indiv}(t)$ curves. Left panel: final model (int., t, $\sqrt{t}$). Right panel: original model (int., t, t^2).*

Another limiting feature comes from the use of the EM fitting algorithm, which fails to deliver precision estimates of parameters. In their paper, Henderson, Diggle and Dobson (2000) obtained standard errors by a Monte-Carlo method, refitting the model to simulated data sets generated using parameter values taken from the original analysis. Clearly, such a procedure may be exceedingly time-consuming here, unless one has a powerful computer at one's disposal.

In view of these comments, there is a strong need to further investigate the performance of simplified computational methods (such as based on the independence assumption) compared to our joint modeling approach. In particular, an extensive simulation study would be mostly welcome but would necessitate access to powerful computational resources, which we did not have.

At the trial level, which is of most interest in practice, the R^2_{trial} surrogacy measure can be easily derived by considering extra terms to characterize the longitudinal evolution of the marker, and our method provides point estimates and uncertainty measures for this parameter. In addition, the individual-level surrogacy can be ex-

plored through the function $R^2_{\text{indiv}}(t)$ which captures the association induced by the two underlying Gaussian processes, $W_1(t)$ and $W_2(t)$, used in the joint model. Since the latter quantity is primarily of interest for exploratory purposes and since computation of the precision estimates within the joint model is cumbersome, we do not attach uncertainty measures to $R^2_{\text{indiv}}(t)$ here. Note that if precision estimates were available for the model parameters, it could also help to incorporate measurement error introduced by the fact that estimates of the $\boldsymbol{\alpha}_i$'s and the β_i's are effectively employed when estimating R^2_{trial}.

Finally, it would be desirable to further investigate model adequacy (with an application to $R^2_{\text{indiv}}(t)$ in mind, for example) or the diagnostic assessment of fitted models. Unfortunately, such tools are currently lacking and this is an area for further research, as pointed out by Henderson, Diggle and Dobson (2000).

As to the clinical interpretation of this work, we saw that PSA level and survival seem, as expected, to be strongly related, at least when a sufficiently large amount of information has been gathered on the marker. While bearing in mind that the R^2_{trial} coefficient was estimated with large uncertainty, the value that was found stands in the mid-range of the unit interval and would prevent us from formulating any firm conclusion, had it been estimated more precisely. This points to an issue, not of the methodology, but rather of the biological nature of the marker. Thus, we may tentatively state that PSA level has some value as a surrogate marker for survival (for the class of treatments considered in the two trials at least) but presumably is not a very good one. Obviously, these results should be taken with caution since this study involved only a couple of clinical trials with a relatively limited number of subjects. The issue of validiting a surrogate marker will, ideally, be based on a much more extended set of randomized trials and will cover different classes of therapies commonly used for the treatment of patients with the disease in question.

Chapter 8

Concluding Remarks and Further Research

This work has focused on methods aimed at modeling data that are correlated. More specifically, we have been concerned with two types of data frequently encountered in applied statistics: multilevel and longitudinal data. Even though formally, multilevel data structures can be regarded as encompassing nearly all other types of correlated data, including longitudinal data, the distinction between these two data types is maintained since they each have their own pecularities and field of research.

8.1 Pairwise Likelihood Estimation

A significant part of our work (Chapters 3, 4 and 6) was devoted to models for binary responses. The difficulty in evaluating the likelihood for models with discrete correlated data has led to alternative methods of estimation and to an extensive body of research in the literature. The aim of our work was to study in more depth the use of maximum pairwise likelihood (MPL) as an estimation tool in multilevel models with binary responses. Pairwise likelihood is a special example of pseudo-likelihood, a notion first introduced by Besag (1975) and which amounts to construct a product (of pieces) of likelihoods.

Although several authors have already used MPL to model clustered binary data, it was not investigated, to our knowledge, in the multilevel modeling framework, that

is, using a 'subject-specific' rather than 'population-averaged' approach. In Chapter 3 we examined the merits of MPL for estimation purposes in multilevel models with binary responses. To summarize, MPL enjoys appealing asymptotic properties (such as consistency and asymptotic normality) and is computationally simple. In compromise, MPL estimators are subject to some loss of efficiency.

Our work has concentrated on a very limited aspect of modeling however, and there are yet a number of issues to examine if we wish to use MPL estimation routinely. These are briefly discussed below.

8.1.1 Model Checking and Diagnostics

An important step in the process of data modeling is to check various features of the fitted model. This usually involves checking goodness-of-fit of the model, checking model assumptions, and detecting possibly influential observations.

Surprisingly, little work has been done on model checking and model diagnostics in generalized linear mixed models and this topic alone would definitely deserve further research. In relation to our work, we think that the following points would be of particular interest.

Normality of Random Effects

An important assumption in (generalized) linear mixed models is that the random effects are normally distributed. A first issue is therefore to obtain individual estimates for the random effects. In linear mixed models, Empirical-Bayes (EB) estimates are often used for diagnostic purposes (Verbeke and Molenberghs, 2000). EB estimates are also employed in generalized linear models with random effetcs, although they have no closed-form expression and thus are typically approximated by the mode of the posterior distribution using estimated values of the parameters (see, e.g., Stiratelli, Laird and Ware, 1994).

Basic checks of the normality assumption for random effects are usually limited to examination of the distribution of EB estimates, using histograms or normal probability plots for example. As shown by Verbeke and Lesaffre (1996) however, such histograms may be misleading and these authors suggest to extend the standard linear mixed model to the so-called "heterogeneity model" which involves a mixture of normal distributions for the random effects. The two models can then be compared and the assumption of normality for random effects formally tested. Obviously, this

work would need to be pursued within the realm of generalized linear mixed models, whatever method is used for estimation of model parameters.

Influence Measures

In general, influence measures aim at determining whether some observations have undue influence on the estimates of the model parameters and hence how sensitive is the fitting of the model to such observations. With longitudinal (or, more generally, clustered) data we actually need to distinguish between influential subjects (influence of the observations from a particular subject) and influential observations (a particular observation from a particular subject).

One approach to detecting influential observations is local influence (Cook, 1986). Using a case-weight pertubation scheme where it is investigated how much the parameter estimates are affected by changes in the weights of the log-likelihood contributions of specific subjects, Lesaffre and Verbeke (1998) derive local influence measures in linear mixed models. More recently, Ouwens, Tan and Berger (2001) extended these local influence measures in generalized linear models with random effects.

Since these local influence measures are based on perturbations of the likelihood function, it would be worthy to investigate whether similar measures can be derived for pseudo-likelihood functions and, if this path proves to be unsuccessful, to propose other influence measures for use with pseudo-likelihood estimation.

8.1.2 Missing Data

Another area where pseudo-likelihood methods would benefit from further research is when data are incomplete (or missing). A problem with incomplete data is that ignoring the missingness mechanism can result in misleading inference. Based on the well-known terminology of Rubin (1976) and Little and Rubin (1987), missing data mechanisms can be classified as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). A remarkable fact about likelihood-based estimation is that MCAR and MAR mechanisms are ignorable (provided an unrestrictive separability condition on the parameter spaces is satisfied). In practical terms this means that MCAR and MAR missingness mechanisms can simply be ignored in the analysis of the data within the likelihood framework.

Unfortunately, the MAR mechanism will *a priori* not be ignorable if pseudo-likelihood estimation is to be pursued. This can be seen if we consider the following

factorization, which is the basis of the so-called selection models,

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i) = f(\boldsymbol{y}_i) f(\boldsymbol{r}_i | \boldsymbol{y}_i),$$

where $\boldsymbol{r}_i$ denotes the missing data indicator. In previous chapters we have been only concerned with modeling of the first factor, that is, of the joint distribution of the response vector. The above formula shows, however, that an additional factor must be taken into consideration with missing data, which allows the data analyst to formulate a model for the missingness mechanism.

While pseudo-likelihood estimation may result in biased estimators under MAR, it would still be useful to evaluate the robustness of the method to missing data. Since a pseudo-likelihood is essentially composed of likelihoods, an hypothesis would be that it is more robust than other methods purely based on estimating equations (such as GEE). Also, it would be worth investigating the performance of a "weighted" approach, in the sense of Robins, Ronitzky and Zhao (1995) and Fitzmaurice, Molenberghs and Lipsitz (1995) who discuss weighted GEE, an extension of GEE which yields asymptotically unbiased estimates under MAR provided the models for the mean structure and missingness mechanism are correctly specified.

8.1.3 Crossed Random-Effects Models

Throughout we have exclusively considered models where population units are hierarchically structured. There are cases, however, where units at the same level of a hierarchy are simultaneously classified by more than one factor. For example, school pupils may be classified by the school they attend as well as the neighborhood they live in. Since schools usually attract pupils from several neighborhoods, these two factors are crossed.

Random effects can be specified (as usual) for each of the crossed factors; hence crossed random-effects models extend the standard multilevel models. Interestingly, it can be shown that this kind of models can be analyzed using procedures designed for purely hierarchical or multilevel structures, although they are technically and computationally more demanding.

We now briefly discuss some work in progress about crossed random-effects models for binary responses, which are much more challenging to fit than the standard generalized linear multilevel models.

Consider the following model:

$$g(\pi_{i(j_1j_2)}) = \boldsymbol{x}_{ij}^{T}\boldsymbol{\beta} + u_{1j_1} + u_{2j_2}, \tag{8.1}$$

where i is the level 1 index, and j_1 and j_2 are level 2 indices for two classification factors (the parentheses mean grouped classifications at the same level). Thus, the model assumes two random intercepts to represent heterogeneity between units from both classifications.

The major difficulty with likelihood estimation stems from the fact that the local independence assumption no longer holds and hence the integral over the random effects does not simplify. More specifically, the marginal likelihood takes the form

$$L = \int \prod_{j_1,j_2,i} \pi_{i(j_1j_2)}^{y_{i(j_1j_2)}} (1-\pi_{i(j_1j_2)})^{1-y_{i(j_1j_2)}} \phi(u_{1j_1}, u_{2j_2}) d\boldsymbol{u}, \tag{8.2}$$

where $\boldsymbol{u}$ contains all u_{1j_1}'s and u_{2j_2}'s. In consequence, the dimension of integration is now dramatically higher and the only practical way to carry out likelihood estimation is to resort to Monte Carlo techniques, which are computationally intensive.

Thus, we are clearly in a situation where pairwise likelihood might be of great help when fitting of a model such as (8.1) is needed. This entails considering all pairwise probabilities of the kind $P(y_{i(j_1j_2)}, y_{i'(j_1j_2)})$, $P(y_{i(j_1j_2)}, y_{i'(j_1j_2')})$, and $P(y_{i(j_1j_2)}, y_{i'(j_1'j_2)})$. Of course, this may represent a considerable amount of contributions depending on the size of the data set.

A problem yet to solve with this approach is to estimate standard errors of the parameters. Indeed, we saw in Section 3.2 that the asymptotic variance-covariance matrix of the pseudo-likelihood estimator takes the form

$$J(\boldsymbol{\theta}_0)^{-1} K(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}, \tag{8.3}$$

with

$$J(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left[-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log L \right]$$

and

$$K(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log L \frac{\partial}{\partial \boldsymbol{\theta}^T} \log L \right] = \mathrm{var}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log L \right].$$

In this expression, $J(\boldsymbol{\theta}_0)$ can be easily estimated, using second-order derivatives as in (3.15) or crossed products of first-order derivatives as in (3.16). For $K(\boldsymbol{\theta}_0)$, the estimator

$$K_N(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log L(\boldsymbol{\theta}_0)$$

would be consistent, but $\boldsymbol{\theta}_0$ is not available. Of course, the plug-in estimator

$$K_N(\tilde{\boldsymbol{\theta}}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\tilde{\boldsymbol{\theta}}) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log L(\tilde{\boldsymbol{\theta}})$$

cannot be used since it is identically zero by definition of $\tilde{\boldsymbol{\theta}}$. In fact, what we are lacking here is independent replication over subjects (or clusters), which allowed us to use the following estimator in Chapter 3:

$$K_N(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}} \log p\ell_i(\tilde{\boldsymbol{\theta}}) \frac{\partial}{\partial \boldsymbol{\theta}^T} \log p\ell_i(\tilde{\boldsymbol{\theta}}).$$

In marginal generalized linear models, some empirical variance estimators have been proposed to solve this problem (see Lumley and Heagerty (1999) for example), albeit mostly in the context of spatial data. The work of Lumley and Mayer-Hamblett (2002), however, is directed towards sparse correlation structures, including those with crossed design, and contains interesting results (e.g. central limit theorem) that might ultimately prove useful to derive imprecision measures for MPL estimators.

8.2 Evaluation of Surrogate Endpoints

In Chapter 4 we have presented an approach to surrogate endpoint validation using data from multiple randomized clinical trials, as proposed by Buyse *et al.* (2000). A first critical step was to extend the method initially developed for two normally distributed endpoints. In this work the approach was extended to deal with discrete outcomes and to the case where a longitudinally measured biomarker is used as a surrogate for a failure-time variable.

In addition to the references already cited in Section 4.1, we can mention the work of Burzykowski (2001) who studied in detail the proposed methodology when the true endpoint is a failure-time variable and introduced a useful concept, the surrogate threshold effect, which is defined as the minimum value of treatment effect on the surrogate endpoint for which the predicted effect on the true endpoint would be significantly different from 0. In addition to providing information relevant to the practical use of a surrogate endpoint, the surrogate threshold effect also has a natural interpretation from a clinical point of view, as it can be expressed in terms of treatment effect necessary to be observed to predict a significant treatment effect on the true endpoint. Therefore, its use might facilitate communication between statisticians and clinicians regarding results of a validation of a surrogate endpoint.

Also, in the context of repeated measurements on both the surrogate and the true endpoint, we refer to the work of Alonso, Geys, Molenberghs *et al.* (2002) who generalize the R^2 surrogacy measures using the so-called variance reduction factor (VRF), a measure which summarizes the variability of the repeated measurements on the true endpoint over all trials. As shown by Alonso, Geys, Kenward *et al.* (2002), the VRF can even be regarded as a special member of a large class of canonical correlation functions that can be used to study surrogacy at the trial and individual levels.

On related topics worth of further research, we think that more work should be done to evaluate computationally simpler modeling strategies and compare them to the more elaborate modeling techniques currently proposed. As shown by Tibaldi *et al.* (2002) in the case of two normally distributed endpoints, such strategies produce results reasonably close to those obtained from the full random-effects model (4.5). It would be interesting to investigate whether this is true in other settings, such as the longitudinal-survival situation of Chapter 7, and if not, what is the amount of bias and/or loss of precision attributed to such strategies.

References

Agresti, A., Booth, J.B., Hobert, J.P., and Caffo, B., (2000). Random-effects modeling of categorical response data. *Sociological Methodology*, **30**, 27–80.

Algina, J. (1999). A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient. *Multivariate Behavioral Research*, **34**, 494–504.

Alonso, A., Geys, H., Kenward, M., Molenberghs, G., and Vangeneugden, T. (2002). Validation of Surrogate Markers in Multiple Randomized Clinical Trials with Repeated Measurements. *Submitted for publication.*

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M. (2002). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: a canonical correlation approach. *Submitted for publication.*

Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203–210.

Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya, Series B*, **53**, 233–243.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, **81**, 767–775.

Bahadur, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, H. Solomon (ed.), Stanford Mathematical Studies in the Social Sciences VI. Stanford: Stanford University Press, pp. 158–168.

Barbosa, M.F. and Goldstein, H. (2000). Discrete response multilevel models for repeated measures: an application to voting intentions data. *Quality and Quantity*, **34**, 323–330.

Besag, J.E. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**, 179–195.

Boissel, J.P., Collet, J.P., Moleur, P., and Haugh, M. (1992). Surrogate endpoints: a basis for a rational approach. *European Journal of Clinical Pharmacology*, **43**, 235–244.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park: Sage Publications.

Burzykowski, T. (2001). *Validation of Surrogate Endpoints from Multiple Randomized Clinical Trials with a Failure-Time True Endpoint*. Unpublished Ph.D. dissertation, Limburgs Universitair Centrum.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics*, **50**, 405–422.

Burzykowski, T., Molenberghs, G., Tafforeau, J., Van Oyen, H., Demarest, S., and Bellamammer, L. (1999). Missing data in the Health Interview Survey 1997 in Belgium. *Archives of Public Health*, **57**, 107–130.

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**, 49–67.

Carey, V.C., Zeger, S.L., and Diggle, P.J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517–526.

Choi, S., Lagakos, S., Schooley, R.T., and Volberding, P.A. (1993). CD4+ Lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking Zidovudine. *Annals of Internal Medicine*, **118**, 674–680.

Chuang-Stein, C. and DeMasi, R. (1998). Surrogate endpoints in AIDS drug development: current status. *Drug Information Journal*, **32**, 439–459.

Cleveland, W.S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Coursaget, P., Leboulleux, D., Soumare, M., le Cann P., Yvonnet, B., Chiron, J.P., and Collseck A.M. (1994). Twelve-year follow-up study of hepatitis immunization of Senegalese infants. *Journal of Hepatology*, **21**, 250–254.

Cox, D.R. (1972). The analysis of multivariate binary data. *Applied Statistics*, **21**, 113–120.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*, 2nd edition. London: Chapman and Hall.

Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Crouch, E.A.C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association*, **85**, 464–469.

Dale, J.R. (1986). Global cross-ratio models for bivariate discrete ordered responses. *Biometrics*, **42**, 909–917.

Da Villa, G., Peluso, F., Picciotto, L., Bencivenga, M., Elia, S., and Pelliccia, M.G. (1996). Persistence of anti-HBs in children vaccinated against viral hepatitis B in the first year of life: follow-up at 5 and 10 years. *Vaccine*, **14**, 1503–1505.

Debruyne, F.J.M., Murray, R., Fradet, Y., Johansson, J.E., Tyrrell, C., Boccardo, F., *et al.* (1998). Liarozole - a novel treatment approach for advanced prostate cancer: results of a large randomized trial versus cyproterone acetate. *Urology*, **52**, 72–81.

De Gruttola, V., Fleming, T.R., Lin, D.Y., and Coombs, R. (1997). Validating surrogate markers – Are we being naive? *Journal of Infectious Diseases*, **175**, 237–246.

De Gruttola, V. and Tu, X.M. (1994). Modelling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.

De Gruttola, V., Wulfsohn, M., Fischl, M.A., and Tsiatis, A. (1993). Modelling the relationship between survival and CD4 lymphocytes in patients with AIDS and AIDS-related complex. *Journal of Acquired Immune Deficiency Syndromes*, **6**, 359–365.

De Ponti, F., Lecchini, S., Cosentino, M., Castelletti, C.M., Malesci, A., and Frigo, G.M. (1993). Immunological adverse effects of anticonvulsants. What is their clinical relevance? *Drug Safety*, **8**, 235–250.

Diggle, P.J. (1988). An approach to the analysis of repeated measures. *Biometrics*, **44**, 959–971.

Diggle, P.J. (1990). *Time series: a biostatistical introduction*. Oxford: Oxford University Press.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Ding, C.G. (1996). On the computation of the distribution of the square of the sample multiple correlation coefficient. *Computational Statistics and Data Analysis*, **22**, 345–350.

Ellenberg, S.S. and Hamilton, J.M. (1989). Surrogate endpoints in clinical trials: Cancer" *Statistics in Medicine*, **8**, 405–413.

Faucett, C.L. and Thomas, D.C. (1994). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1685.

Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood-based method for analyzing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Fitzmaurice, G.M., Molenberghs, G., and Lipsitz, S.R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society, Series B*, **57**, 691–704.

Flandre, P. and Saidi, Y. (1999). Letters to the Editor: Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, **18**, 107–115.

Fleming, T.R. and DeMets, D.L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine*, **125**, 605–613.

Fleming, T.R., Prentice, R.L., Pepe, M.S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine*, **13**, 955–968.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.

Geys, H. (1999). *Pseudo-likelihood Methods and Generalized Estimating Equations: Efficient Estimation Techniques for the Analysis of Correlated Multivariate Data*. Unpublished Ph.D. dissertation, Limburgs Universitair Centrum.

Geys, H., Molenberghs, G., and Lipsitz, S. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginally specified odds ratio models with exchangeable association structure. *Journal of Computational Statistics and Simulations*, **62**, 45–71.

Geys, H., Molenberghs, G., and Ryan, L. (1997). Pseudo-likelihood inference for clustered binary data. *Communications in Statistics: Theory and Methods*, **26**, 2743–2767.

Gilks, W.R., Wang, C.C., Yvonnet, B., and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*, **49**, 441–453.

Gilmour, A.R., Anderson, R.D., and Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72**, 593–599.

Glonek, G.F.V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, **57**, 533–546.

Godambe, V.P. (1991). *Estimating Functions*. Oxford: Oxford University Press.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, **78**, 45–51.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edition. London: Edward Arnold.

Goldstein, H., Browne, W., and Rasbash, J. (2002). Partitioning variation in generalised linear multilevel models. *Understanding Statistics*, **1**, 000–000.

Goldstein, H., Healy, M.J.R., and Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, **13**, 1643–1655.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **159**, 505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., and Healy, M. (1998). *A user's guide to MLwiN*. London: Multilevel Level Models Project, Institute of Education, University of London.

Graubard, B.I. and Korn, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, **5**, 263-281.

Hadler, S.C., Francis, D.P., Maynard, J.E., Thompson, S.E., Judson, F.N., Echenberg, D.F., *et al.* (1986). Long-term immunogenicity and efficacy of hepatitis B vaccine in homosexual men. *New England Journal of Medicine*, **315**, 209–214.

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.

Heagerty, P.J. and Lele, S.R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, **93**, 1099–1111.

Heagerty, P.J. and Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, **15**, 1–26.

Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933-944.

Hedeker, D. and Gibbons, R.D. (1996). MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, **49**, 157-176.

Henderson R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.

Herson, J. (1975). Fieller's theorem vs. the delta method for significance intervals for ratios. *Journal of Statistical Computations and Simulations*, **3**, 265–274.

Hjort, N.L. (1993). A quasi-likelihood method for estimating parameters in spatial covariance functions. Technical Report SAND/93, Norwegian Computing Centre, Oslo.

Hogan, J.W. and Laird, N.H. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239–257.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, 3rd edition. Englewood Cliffs, NJ: Prentice-Hall.

Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, **13**, 261–276.

Kelly, W.K., Scher, H.I., Mazumdar, M., Vlamis, V., Schwartz, M., and Fossa, S.D. (1993). Prostate-specific antigen as a measure of disease outcome in metastatic hormone-refractory prostate cancer. *Journal of Clinical Oncology*, **11**, 607–615.

Korn, E.L. and Graubard, B.I. (1995). Analysis of large health surveys. *Journal of the Royal Statistical Society, Series A*, **158**, 263–295.

Kreft, I. and de Leeuw, J. (1998). *Introducing Multilevel Modeling.* London: Sage Publications.

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society, Series B*, **57**, 395–407.

Kuk, A.Y.C. and Nott D.J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters*, **47**, 329–335.

Laenen, A., Geys, H., Vangeneugden, T., and Molenberghs, G. (2002). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Submitted for publication.*

Lagakos, S.W. and Hoth, D.F. (1992). Surrogate markers in AIDS: Where are we? Where are we going? *Annals of Internal Medicine*, **116**, 599–601.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lang, J.B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**, 625–632.

Lange, K. (1998). *Numerical Analysis for Statisticians.* New York: Springer.

Lavalley, M.P. and De Gruttola, V. (1996). Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine*, **15**, 2289–2305.

Le Cessie, S. and Van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, **43**, 95–108.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lehmann, E.L. (1983). *Theory of Point Estimation.* New York: John Wiley & Sons.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics*, **50**, 325–335.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Liang, K.-Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007–1016.

Lindsay, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

Lipsitz, S.R., Laird, N.M., and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153–160.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Longford, N.T. (1993). *Random Coefficient Models*. London: Oxford University Press.

Longford, N.T. (1994). Logistic regression with random coefficients. *Computational Statistics and Data Analysis*, **4**, 12–35.

Lumley, T. and Heagerty, P. (1999). Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society, Series B*, **61**, 459–477.

Lumley, T. and Mayer-Hamblett, N. (2002). Asymptotics for marginal generalized linear models with sparse correlations. *Submitted for publication*.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C.E. (1994). Maximum likelihood variance components estimation 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.

McCulloch, C.E. and Searle, R.E. (2000). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons.

McGilchrist, C.A. (1994). Estimation in generalized linear models. *Journal of the Royal Statistical Society, Series B*, **56**, 61–69.

Molenberghs, G., Buyse, M., Burzykowski, T., Renard, D., and Geys, H. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Submitted for publication.*

Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20**, 3023–3038.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

Molenberghs, G. and Ryan, L. (1999). An exponential family model for clustered multivariate binary data. *Environmetrics*, **10**, 279–300.

Monaham, J.F. (2001). *Numerical Methods of Statistics.* Cambridge: Cambridge University Press.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Ouwens, M., Tan, F., and Berger, M. (2001). Local influence to detect influential data structures for generalized linear mixed models. *Biometrics*, **57**, 1166–1172.

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, **88**, 719–726.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, 23–40.

Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

Quataert, P., Van Oyen, H., Tafforeau, J., Schiettecatte, L., Lebrun, L., Bellamammer, L., and Molenberghs, G. (1997) *Health Interview Survey, 1997. Protocol for the Selection of the Households and the Respondents.* Brussels, S.P.H/EPISERIE N12.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics*, **11**, 453–466.

Renard, D., Bruckers, L., Molenberghs, G., Vellinga, A., and Van Damme, P. (2001). Repeated-measures models to evaluate a hepatitis B vaccination programme. *Statistics in Medicine*, **20**, 951–963.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002a). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal*, **44**, 1–15.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Bijnens, L., and Vangeneugden, T. (2002b). Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics*, **29**, 000–000.

Renard, D. and Molenberghs, G. (2002). Multilevel modeling of complex survey data. In *Topics in Modelling of Clustered Data*, M. Aerts, H. Geys, G. Molenberghs and L. Ryan (Eds.). London: Chapman and Hall, pp. 235–243.

Renard, D., Molenberghs, G., and Geys, H. (2002). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, **38**, 000-000.

Renard, D., Molenberghs, G., Van Oyen, H., and Tafforeau, J. (1998). Investigation of the clustering effect in the Belgian Health Interview Survey 1997. *Archives of Public Health*, **56**, 345–361.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rodríguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **158**, 73–89.

Rotnitzky, A. and Jewell, P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered correlated data. *Biometrika*, **77**, 485–497.

Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, **54**, 221–226.

Royston, P. and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, **43**, 429–467.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

SAS Institute Inc. (1995). *SAS/IML Software: Changes and Enhancements Through Release 6.11.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2000). *SAS/STAT User's Guide, Version 8.* Cary, NC: SAS Institute Inc.

Shall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–727.

Skinner, C.J., Holt, D., and Smith, D.M.F. (1989). *Analysis of Complex Surveys.* Chichester: John Wiley & Sons.

Smith, D.C., Dunn, R.L., Stawderman, M.S., and Pienta, K.J. (1998). Change in serum prostate-specific antigen as a marker of response to cytotoxic therapy for hormone-refractory prostate cancer. *Journal of Clinical Oncology*, **16**, 1835–1843.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to the Basic and Advanced Multilevel Modeling.* London: Sage Publications.

Spiegelhalter D.J., Thomas, A., Best, N.G., and Gilks, W.R. (1995). *BUGS Manual and Examples: Version 0.50.* Cambridge: MRC Biostatistics Unit, Institute of Public Health, University of Cambridge.

Sridhara, R., Eisenberger, M.A., Sinibaldi, V.J., Reyno, L.M., and Egorin, M.J. (1995). Evaluation of prostate-specific antigen as a surrogate marker for response of hormone-refractory prostate cancer to suramin therapy. *Journal of Clinical Oncology*, **13**, 2944–2953.

StataCorp. (2001). *Stata Statistical Software: Release 7.0.* College Station, TX: Stata Corporation.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, **40**, 961–971.

Streiner, D.L. and Norman, G.R. (1995). *Health Measurement Scales.* Oxford: Oxford University Press.

Tabor, E., Cairns, J., Gerety, R.J., and Bayley, A.C. (1993). Nine-year follow-up study of a plasma-derived hepatitis B vaccine in a rural Africal setting. *Journal of Medical Virology*, **40**, 204–209.

Taylor, J.M.G., Cumberland, W.G. and Sy, J.P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, **89**, 727–736.

Tibaldi, F., Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2002). Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Submitted for publication.*

Tsiatis, A.A., De Gruttola, V., and Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, **90**, 27–37.

Van Damme, P., Vranckx, R., Safary, A., Andre F.E., and Meheus, A. (1989). Protective efficacy of a recombinant desoxyribonucleic acid hepatitis B vaccine in institutionalized mentally handicapped clients. *American Journal of Medicine*, **87**, 26S–29S.

Van Oyen, H., Tafforeau, J., Hermans, H., Quataert, P., Schiettecatte, E., Lebrun, L., and Bellamammer, L. (1997). The Belgian Health Interview Survey. *Archives of Public Health*, **55**, 1–13.

Vellinga, A., Van Damme, P., Bruckers, L., Weyler, J.J., Molenberghs, G., and Meheus, A. (1999). Modelling long term persistence of hepatitis B antibodies after vacination. *Journal of Medical Virology*, **57**, 100–103.

Vellinga, A., Van Damme, P., and Meheus, A. (1999). Hepatitis B and C in institutions for individuals with mental retardation: a review. *Journal of Intellectual Disability Research*: (1999), **43**, 445–453.

Vellinga, A., Van Damme, P., Weyler, J.J., Vranckx, R., and Meheus, A. (1999). Hepatitis B vaccination in mentally retarded: effectiveness after 11 years. *Vaccine*, **17**, 602–606.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects distribution. *Journal of the American Statistical Association*, **91**, 217–221.

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*. New York: Springer.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wainwright R.B., Bulkow, L.R., Parkinson, A.J., Zanis, C., and McMahon, B.J. (1997). Protection provided by hepatitis B vaccine in a Yupik Eskimo population – results of a 10-year study. *Journal of Infectious Disease*, **175**, 674–677.

Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31**, 949–952.

Wolfinger, R.D. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood appoach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Xu, J. and Zeger, S.L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, **50**, 375–387.

Yang, M., Heath, A., and Goldstein, H. (2000). Multilevel models for repeated binary responses: attitudes and vote over the electoral cycle. *Journal of the Royal Statistical Society, Series A*, **163**, 49–62.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Samenvatting

Dit werkt richt zich op het formuleren van modellen en het schatten van parameters en hun bijhorende precisiematen voor hiërarchische (*multilevel*) en longitudinale gegevens. Dergelijke gegevens komen zeer frequent voor. We noemen eerst enkele belangrijke voorbeelden.

In de Belgische Volksgezondheidsenquête (1997 en 2001; Hoofdstuk 2) worden respondenten bevraagd. De selectie van respondenten gebeurt in verscheidene stappen. Eerst wordt een deel van de steekproef toegekend aan de drie regio's van het land, waarna een proportionele verdeling over de provincies volgt. Binnen elke provincie worden gemeenten geselecteerd, met selectiekans proportioneel aan de grootte van de provincie. Wanneer een gemeente geselecteerd wordt betekent dit dat er 50 interviews zullen afgenomen worden. Bij grotere gemeenten kunnen verscheidene groepen geselecteerd worden. Binnen gemeenten worden huishoudens geselecteerd en binnen huishoudens respondenten. Het zal duidelijk wezen dat er in een zorgvuldige analyse rekening dient gehouden met de stratificatie over regio's en provincies, en met de hiërarchische selectie van respondenten binnen huishoudens en huishoudens binnen gemeente. Bovendien zijn de selectiekansen verschillend van individu tot individu, waardoor met wegingsfactoren dient rekening gehouden te worden. Eén van de gevolgen is dat respondenten niet noodzakelijk van elkaar onafhankelijk zijn.

Een ander belangrijk voorbeeld is surrogaatrespons in klinische studies (Hoofdstuk 4). Met surrogaatrespons bedoelen we een eindpunt in een klinische studie (bijv. prostaat specifiek antigen) dat gebruikt wordt ter vervanging van het werkelijke eindpunt (bijv. overlijden ten gevolge van kanker). De rationale voor de vervanging van een werkelijk eindpunt door een surrogaat is besparing in tijd en/of steekproefgrootte. Een dergelijke vervanging is echter slechts mogelijk na een zorgvuldige studie van de kwaliteit van een surrogaat. Waar Prentice (1989) gebruikt maakte van één enkele

studie om een surrogaat te valideren, hebben Buyse *et al* (2000) een meta-analytisch kader voorgesteld. Hierdoor ontstaat opnieuw een hiërarchische gegevensstructuur. Ten eerste worden patiënten binnen studies geselecteerd en ten tweede worden twee eindpunten (surrogaat en werkelijk) voor elke patiënt opgetekend. In dit werk wordt uitgegaan van een meta-analyse in schizofrenie en van studies in prostaatkanker.

Een derde gebied waar hiërarchische gegevens ontstaan is longitudinale studies. Met deze term duiden we studies aan waarbij patiënten niet één enkele keer doch herhaald in de tijd gemeten worden. Ook metingen aan dezelfde patiënt kunnen a priori niet als onafhankelijk beschouwd worden.

Het beschouwen van hiërarchische gegevens heeft belangrijke methodologische implicaties. Vooreerst dient men zorgvuldig na te denken, meer dan bij onafhankelijke gegevens, over de formulering van de modellen. Dit geldt reeds wanneer de responsvariabelen continu (normaal verdeeld) zijn (Verbeke en Molenberghs 2000), maar a fortiori ook wanneer gegevens discreet (binair, categorisch) zijn (Fahrmeir en Tutz 1994, Diggle, Liang en Zeger 1994). Immers, de keuze van model heeft gevolgen voor de aard en de interpretatie van de modelparameters. Het beantwoorden van vragen op populatieniveau (bijv. het verschil tussen twee armen in een klinische studie) vereist een andere aanpak dan het beantwoorden van vragen op individueel niveau (bijv. de predictie van een later niveau van antilichamen bij een patiënt op basis van eerdere niveaus, in vaccinatiestudies; zie Hoofdstuk 5). In het eerste geval formuleert men best marginale modellen, in het tweede geval zijn modellen met random effecten meer aangewezen.

Een tweede implicatie van hiërarchische modellen is dat er zeer zorgvuldig dient nagedacht over schattingsmethoden. Waar dit probleem beheersbaar blijft bij continue hiërarchische gegevens (lineair gemengde modellen of multilevel modellen, zoals geïmplementeerd in de SAS procedure MIXED of in MLwiN), zijn er meer problemen bij niet-continue gegevens. Een veelbelovende techniek in dit verband is pseudo-likelihood. Hierbij wordt een ingewikkelde likelihood vervangen door een andere, beter beheersbare functie. Een prominente vorm van pseudo-likelihood is paarsgewijze likelihood. Deze methode wordt voorgesteld in de context van een multilevel probit model (Hoofdstuk 3). Het uitgangspunt is dat interesse uitgaat naar de regressieparameters en eventueel ook naar de associatie tussen twee responsen binnen eenzelfde hiërarchisch niveau, maar niet naar hogere orde associaties. In plaats van een likelihood te moeten neerschrijven voor een cluster van hoge dimensie, wordt een dergelijk cluster vervangen door alle mogelijk paren. Er wordt aangetoond dat deze

procedure consistente en asymptotisch normaal verdeelde schatters oplevert, en dat de efficiëntie behoorlijk is, daar waar tegelijk belangrijke winsten in termen van computertijd worden opgeleverd. Het is ook mogelijk deze procedure met wegingsfactoren te combineren.

Paarsgewijze likelihood kan ook gebruikt worden in de context van surrogaatvalidatie met categorische respons (Hoofdstuk 4). Het is belangrijk op te merken dat deze aanpak toelaat het meta-analytische kader van Buyse *et al* (2000), ontwikkeld voor continue gegevens, ook toe te passen op binaire gegevens, en dit met behoud van interpretatie van parameters, de mogelijkheid om met random effecten te werken, en de mogelijkheid om de kwaliteit van surrogaten uit te drukken met behulp van intuïtief aantrekkelijke R^2 maten.

In Hoofdstuk 5 worden herhaalde metingen, afkomstig van vaccinatiestudies, bekeken. Belangrijke problemen zijn het voorkomen van booster doses en de noodzaak aan predictie te doen op jaar 12, gegeven vroegere waarnemingen.

In Hoofdstuk 6 wordt een uitbreiding van een hiërarchisch model voor binaire gegevens voorgesteld, waarbij naast random effecten ook seriële correlatie wordt ingebouwd. Een flexibel model van dit type werd niet eerder voorgesteld. Ook hier wordt gebruik gemaakt van paarsgewijze likelihood. Het model wordt toegepast op het schatten van psychometrische betrouwbaarheid in de context van klinische studies in psychiatrie. Deze aanpak heeft als voordeel dat zeer flexibele hiërarchische datasets kunnen gebruik worden, wat de klassieke vormen van betrouwbaarheid, zoals test-hertest betrouwbaarheid, uitbreidt.

Een bijzondere belangrijke vorm van surrogaatrespons is een longitudinaal surrogaat (zoals prostaat-specifiek antigen) voor een overlevingstijd als werkelijk eindpunt. Een dergelijke setting vereist het gemeenschappelijk modeleren van beide responsen. Zeker met het voorkomen van censurering is dit een situatie die de nodige zorgvuldige reflectie verdiend. Een mogelijk model, met toepassing in de validering van surrogaatrespons, wordt gegeven in Hoofdstuk 7.