METHODS FOR THE ANALYSIS OF INCOMPLETE LONGITUDINAL DATA

A THESIS SUBMITTED TO
LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE
IN THE SUBJECT OF MEDICAL STATISTICS
FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

By
Claudio John Verzilli
2003

Abstract

Unplanned missing data commonly arise in longitudinal trials. When the mechanism driving the missing data process is related to the outcome under investigation, traditional methods of analysis may yield seriously biased parameter estimates. Motivated by data from two clinical trials, this thesis explores various approaches to dealing with data incompleteness.

In the first part, a Monte Carlo EM algorithm is developed and used to fit so called random-coefficient-based dropout models; these models relate the probability of a patient's dropout in follow-up studies to some subject-specific characteristics such as their deviation from the average rate of progression of the disease over time. The approach is used to model incomplete data from a 5-year study of patients with Parkinson's disease.

The validity of the results obtained using these methods however, depends in general on distributional and modelling assumptions about the missing data that are inherently untestable as no data were collected. For this reason, many have advocated the need for a sensitivity analysis aimed at assessing the robustness of the conclusions from an analysis that ignores the missing data mechanism. In the second part of the thesis we address these issues. In particular, we present results from sensitivity analyses based on local influence and sampling-based methods used in conjunction with the random-coefficient-based dropout model described

in the first part.

Recently, a more formal approach to sensitivity analysis for missing data problems has been proposed whereby traditional point estimates are replaced by intervals encoding our lack of knowledge due to incompleteness of the data. In the third part of the thesis, we extend these methods to longitudinal ordinal data. Also, for cross-sectional discrete data having distribution belonging to the exponential family, we propose using the proportion of possible estimates of a parameter of interest, over all solutions corresponding to all sample completions, as a measure of ignorance. We develop a computationally efficient algorithm to calculate this proportion and illustrate our methods using data from a dental pain trial.

Acknowledgments

I would like to thank my supervisor, Dr James Carpenter, for his guidance and encouragement throughout my thesis project and for patiently reading draft versions of each Chapter.

This research was funded by a scholarship awarded by GlaxoSmithKline Pharmaceuticals to whom I am grateful. In particular I would like to thank Dr Sandy Macrae, Richard Davies and Mick Ireson from GlaxoSmithKline for proving the data used in this project.

I am also grateful to my family and my partner Monica for their support during my academic studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Longitudinal data arise naturally in many areas of statistics, not least medical applications. Whenever interest lies in the temporal evolution of a certain phenomena, there are distinct advantages in collecting repeated measurements on the same units as opposed to a single one in a cross-sectional approach (Diggle et al., 1994). For instance, in drug development studies, assessment of the efficacy and safety of competing treatments often requires the follow-up of patients over time. Such study designs are inevitably expensive both in terms of time required to reach any conclusions and effort on the part of the investigator to make sure that data are as complete as possible. Despite careful planning however, data collected longitudinally are likely to be incomplete for many reasons. For example, in a clinical trial a patient may move out of the area and be lost to follow-up or may miss one or more of the scheduled visits because of illness. The first consequence of this is that, an intended balanced data set (with measurements taken at the same time on every patient), becomes unbalanced, and the efficiency of the parameter estimators is reduced. However, from a data analysis perspective, this does not in general constitute a serious problem as many techniques are available that can handle unbalanced data sets. A second problem is related to the validity of the results obtained from an incomplete data set. In the case mentioned

above, a patient's moving away from the area where the study is being conducted is likely to be unrelated to the study itself and valid conclusions can be obtained from analysing the incomplete data (as long as maximum likelihood methods are used). In the second example however, the missing data mechanism could well be related to the outcome that would have been observed had the patient attended the schedule visits and ignoring this could lead to substantially biased estimates.

The issue of how to deal with missing data in longitudinal studies has received the attention of many researchers in the last ten years or so. Early approaches focused on obtaining adjusted point estimates by joint modelling the response and missing data mechanism whenever the latter is though to be related to the outcome of interest. However, it was soon realized that such methods rely on inherently untestable assumptions and their scope can only be as part of a sensitivity analysis of the results obtained from an analysis of available cases. In most cases, the mechanism driving the missing data process is not known and any adjustment will depend ultimately on investigators' subjective beliefs. Thus, there cannot be a single solution to the problem valid across different study settings and for different data sets. Rather, every single study represents a unique challenge and input from the investigators, far from being detrimental, will serve the purpose of narrowing down possible missing data mechanisms in order to obtain sensible conclusions. Contradictory as it may sound, missing data represent and important and integral part of the information collected and just like any other data collected, their modelling will require careful thought.

The work presented in this thesis was motivated by two incomplete data sets from phase III clinical trials. The first data set consists of scores of Activity of Daily Living (ADL) from a randomised double-blind study comparing the safety and efficacy of Ropinirole and Levodopa in the treatment of patients in early stages of Parkinson's disease (Rascol et al., 2000). Many patients did not com-

plete the intended follow-up period of five years. To account for this, a random-coefficient-based model for the missing data mechanism is jointly fitted with a linear mixed model for the response variable using a Monte Carlo Expectation-Maximization (MCEM) algorithm. This choice for the missing data model and the description of the MCEM algorithm constitute the focus of Chapter 3.
The results from this joint modelling approach are supplemented with sensitivity analyses in Chapter 4. In particular two methods are considered: Local Influence adapted to the chosen model for the missing data mechanism and a sampling-based approach using finite-element methods. The latter method is also used to account for the effect of rescue medication, as in this trial patients whose symptoms were not adequately controlled by the randomised treatment could receive supplemental open-label Levodopa.

The second data set motivates the work presented in Chapters 5 and 6. Data consists of pain scores form a double-blind placebo-controlled trial conducted to assess the efficacy of two drugs in the treatment of post-surgical pain following third molar extraction. There is severe attrition with almost 80% of patients in the placebo arm leaving the study prematurely. To account for the large number of missing data, the response variable is modelled using the adaptation of the GEE method to longitudinal ordinal data and pessimistic-optimistic bounds on parameter estimates are obtained using a modified Fisher scoring algorithm under different scenarios for the missing data. This approach is the focus of Chapter 5. Bounds on parameter estimates are also obtained for continuous data using a modified IGLS algorithm and results from the application to the Parkinson's disease trial are presented.

The estimation of bounds for parameter estimates enables a degree of quantification of the extra uncertainty induced by the unplanned missing data. However, with discrete data, a more interesting question from a clinician's perspective

might be the proportion of pseudo-complete data sets that would result in estimates that are greater or smaller than an user-specified threshold. A method is described in Chapter 6 which enables calculation of these proportions based on a Fisher scoring algorithm with nested saddlepoint approximations. We also extend methods based on exact conditional inference for generalized linear models and calculate 'expected' exact levels of significance for coefficients of interest, where expectations are taken over the distribution of the missing data. Finally, Chapter 7 contains a discussion and outlines further areas of research.

Both data sets are described in more detail in Chapter 2 where Rubin's classification of missing data mechanisms and the notation used throughout are also introduced.

Chapter 2

Background

This chapter is organized in four Sections. A detailed description of the motivating data sets from a Parkinson's disease and dental pain trial is given in the first part. In the second, the taxonomy of missing data mechanisms due to Rubin (1976) and Little and Rubin (1987) is introduced. Various approaches proposed in the literature for dealing with incomplete data are set within this general framework. In Section 3 we review EM-type algorithms available for dealing with missing data problems; these underlie some of the methods presented in later Chapters. Section 4 contains some concluding remarks.

2.1 Data

2.1.1 Parkinson's Disease trial

Parkinson's disease is a common progressive neurological disorder caused by degeneration of nerve cells (neurons) in a region of the brain that controls movement. This degeneration results in a shortage of the brain-signaling chemical (neurotransmitter) known as dopamine, causing impaired movement.

There is no cure for Parkinson's disease. Many patients are only mildly affected and need no treatment for several years after diagnosis. When symptoms grow severe, doctors usually prescribe Levodopa (L-dopa), which helps replace the

brain's dopamine but has long-term complications such as dyskinesia – i.e. a set of abnormal, involuntary movements of the mouth or facial area. As a result of this, alternative therapies have been sought. One such drug is Ropinirole, a dopamine agonist which works by mimicking dopamine.

The data considered here come from a study conducted to assess the safety and efficacy of Levodopa and Ropinirole in the treatment of patients in early stages of the disease. A total of 268 patients were randomised to receive either Ropinirole (179) or Levodopa (89). Patients whose symptoms were not adequately controlled, received supplemental open-label doses of Levodopa. The main outcome was time to incidence of dyskinesia. Other symptoms recorded included scores of Activities of Daily Living (ADL) which is the outcome analyzed below. Values range from 0 to 52 with 0 indicating no disability and 52 maximum impairment.

The analysis conducted by the investigators compared mean changes from baseline in ADL score among subjects completing the 5-year follow-up and found no statistically significant difference between the two treatment groups (p=0.08). The aim in the first part of this thesis will be to assess how robust these conclusions are to assumptions about the dropout mechanism while taking into account the longitudinal nature of the data collected.

Figure 2.1 plots the mean profiles of ADL scores using all available data at each time point over the 5 years in the two treatment arms. The number of patients remaining in the study is shown above the x-axis. In all subsequent analyses, data from the first two visits have been ignored since, on average, treatment seems to attain its full effect by the third visit (Figure 2.1). Thus, we refer to measurements made 12, 24, 48, 72, 96, 120, 144, 168, 192, 216 and 240 weeks after randomisation.

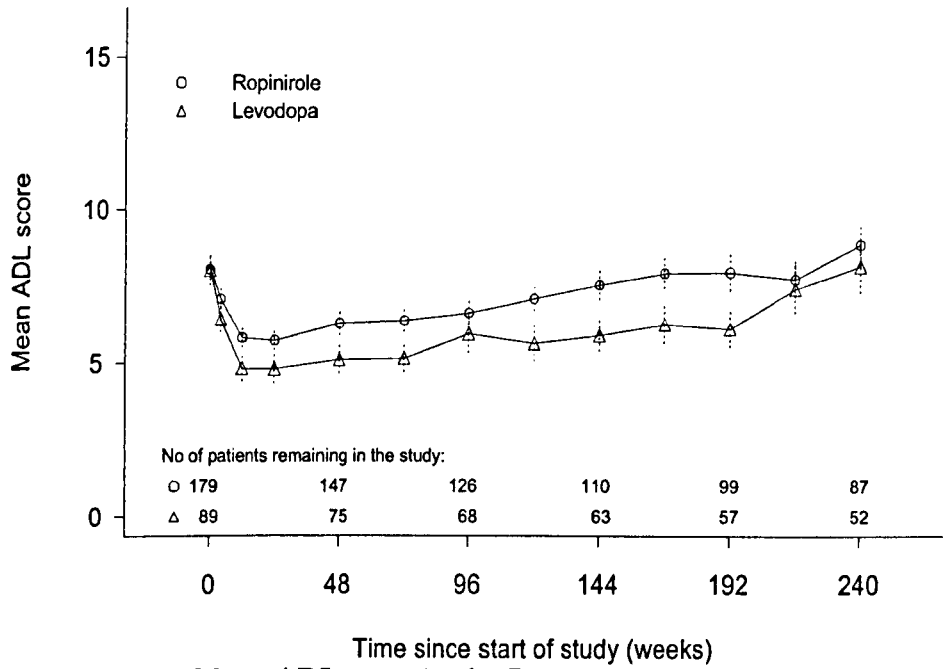

Figure 2.1: Mean ADL score in the Ropinirole and Levodopa arms.

2.1.2 *Dental pain trial*

The data come from a multicentre, double-blind placebo-controlled study which randomised patients with moderate to severe postsurgical pain following third molar extraction to receive a single oral dose of either a test drug at five different increasing doses (referred to as Test Doses 1 to 5), a positive control or a placebo. A total of 366 patients entered the study and were randomised to one of seven groups of roughly 50 each. Among other measures of efficacy, pain relief from initial onset of pain was recorded using the following self-rating scale: 1=none, 2=a little, 3=some, 4=a lot and 5=complete. Following administration of study medication, patients provided pain evaluations at 15, 30 and 45 minutes, 1 hour, 1.5 hours, every hour from 2 to 12 hours, 18 and 24 hours.

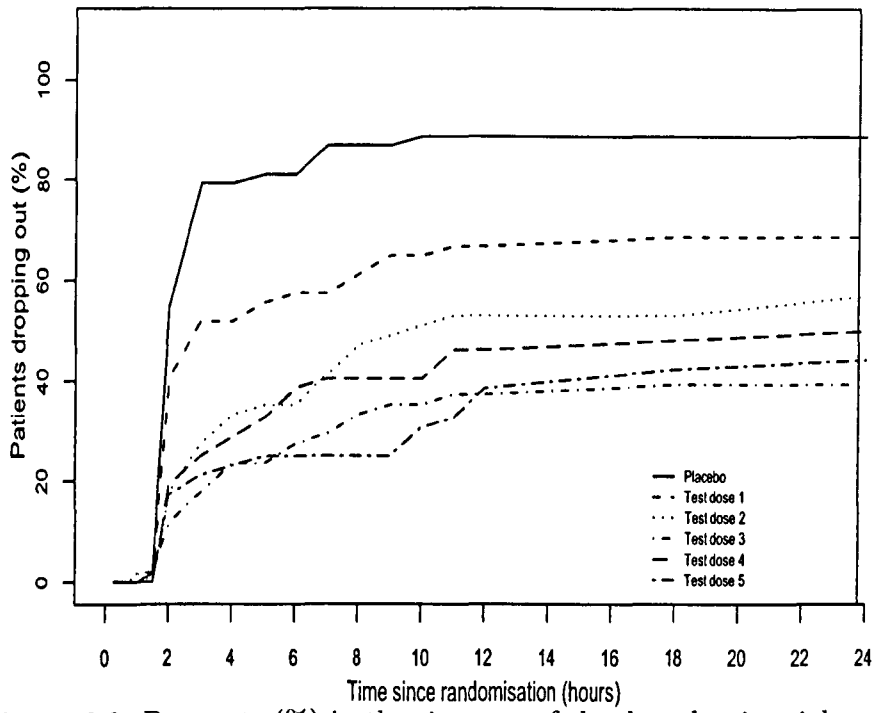

Figure 2.2: Dropouts (%) in the six arms of the dental pain trial.

In our analyses, we focus on the contrasts between placebo and the test dose groups and omit patients in the positive control group leaving us with 313 subjects. The primary endpoint defined in the protocol was total pain relief 8 hours after randomisation, obtained by summing the 12 time specific measurements for each patient. In the analysis conducted by the investigators using this endpoint, treated patients showed a statistically significant improvement in total pain relief compared to the placebo ($p< 0.001$), with dose levels two or above apparently not giving increased relief. As a result, Test Dose 2 was considered to be the lowest dose clinically effective. However, as with the Parkinson's disease trial, in this trial it would be interesting to assess how this conclusion is affected by the large number of missing scores and use methods for repeated measurements.

Figure 2.2 shows the percentage of patients dropping out of the study. By 24 hours after randomisation, around 90% and 40-70% of the data are missing in the placebo and active groups respectively. Both interim missing data and

dropouts occur. In the original analysis performed by the drug company, linear interpolation was used to impute interim missing values; in case of dropouts, the patient's last available measurement was carried forward to obtain a pseudo-complete data set.

2.2 Missing data mechanisms

In a broad sense, all studies that involve randomisation of units to treatments are affected by missing data since a unit's response for the treatment(s) it was not allocated is unobserved. As long as the intended sample is fully observed, so that uncertainty about parameter estimates falls within theory of sampling inference, the missing mechanism is under the control of the statistician (Little and Rubin, 1987). Often, however, some of the intended observations will be missing from the sample and the mechanism that caused this will be outside the control of the statistician. In a longitudinal study for instance, some patients may withdraw prematurely or miss some of their intended visits. Failure to adjust the analysis to take account of this can lead to bias in parameter estimates and hence misleading inferences being drawn.

We now introduce the general notation used in this thesis. Additional notation will be described as and when needed.

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ indicate a vector valued random variable of intended measurements on subject i and $\mathbf{R}_i = (R_{i1}, \ldots, R_{in_i})^T$ the corresponding vector of binary indicators flagging whether the intended measurement at time t_j was observed ($R_{ij} = 0$) or not ($R_{ij} = 1$), $j = 1, \ldots, n_i$. Suppose further that the distribution of $\mathbf{R}$ is parametrised by a q-dimensional vector ψ. In what follows, we suppose $f(\mathbf{Y}_i, \mathbf{b}_i \mid \boldsymbol{\theta}, \mathbf{X}_i)$ specifies a mixed model for $\mathbf{Y}_i$ and let $\mathbf{b}_i$ be a vector of normally distributed subject-specific random coefficients with mean zero as in Laird and Ware (1982). The vector $\boldsymbol{\theta}$ comprises both the variance components

and the parameters relating $\mathbf{Y}_i$ to $\mathbf{X}_i$, the $n_i \times p$ design matrix of covariates assumed to be completely observed.

Continuing to focus on an individual, but dropping the subscript i for simplicity, two approaches can be distinguished depending on how the joint distribution $f(\mathbf{Y}, \mathbf{R}, \mathbf{b} \mid \mathbf{X}, \theta, \psi)$ of the outcome variable and response indicators is factorized (Little, 1995; Kenward and Molenberghs, 1999).

In *Selection Models* it is assumed that

$$f(\mathbf{Y}, \mathbf{R}, \mathbf{b} \mid \mathbf{X}, \theta, \psi) = f(\mathbf{Y}, \mathbf{b} \mid \mathbf{X}, \theta) f(\mathbf{R} \mid \mathbf{Y}, \mathbf{b}, \mathbf{X}, \psi) \tag{2.1}$$

where $f(\mathbf{R} \mid \mathbf{Y}, \mathbf{b}, \mathbf{X}, \psi)$ is the conditional distribution of the missing-data indicators conditioning on the $\mathbf{Y}$ and $\mathbf{b}$. Thus the incomplete data can be thought of being the result of a process of selection based on the values of $\mathbf{Y}$ or $\mathbf{b}$.

There are circumstances where describing the behaviour of $\mathbf{Y}$ across different patterns of missingness or dropout might be of interest and seems a more sensible approach to the problem. Consider for example a follow-up study with four intended measurement times common to all subjects. Loss of follow up caused by attrition may result in a monotone pattern of missingness i.e. for subject i, $Y_{i,j}$ is observed only if $Y_{i,j-1}$ is observed $j = 2, \ldots, 4$ (Little and Rubin, 1987). Then the distribution of the response variable within each of the four patterns of dropout might be of interest in its own right. This leads to *Pattern-Mixture* models. Different parameters characterise the response variable across the different patterns or dropout times; the marginal distribution of the response variable is then a mixture over these patterns (Little, 1993).

This uses the factorization

$$f(\mathbf{Y}, \mathbf{R}, \mathbf{b} \mid \mathbf{X}, \theta, \psi) = f(\mathbf{Y}, \mathbf{b} \mid \mathbf{R}, \mathbf{X}, \theta) f(\mathbf{R} \mid \mathbf{X}, \psi). \tag{2.2}$$

Within these two approaches, different missing-data mechanisms can be distinguished following Rubin (1976), Little and Rubin (1987) and Little (1995). In the remainder of this thesis however, we will be concerned only with selection models, the main reason being that, as discussed later in Chapter 3, for the Parkinson's disease data this seems a more sensible choice.

Also, explicit restrictions about the distribution of the missing data conditional on the observed data have to be made in order to identify parameters in the pattern–mixture framework; for some classes of identifying restrictions the selection and pattern–mixture approaches have been show to be equivalent (Molenberghs et al., 1998).

2.2.1 *Missing completely at random (MCAR) mechanism*

Under this mechanism, the missing-data or dropout mechanism does not depend on the observed data or on the missing data (i.e. on the values that would have been observed had they not been missing for whatever reason).

Therefore, partitioning the response vector into the observed and missing components $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{miss}$ gives $f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{b}, \mathbf{X}, \psi) = f(\mathbf{R} \mid \mathbf{X}, \psi)$. As long as the parameters θ and ψ of the response and missing-data model are distinct (separability condition), the missing-data mechanism can be ignored when calculating ML estimates of θ, which can be obtained by maximising the log-likelihood of the observed data $\mathbf{Y}^{obs}$ alone.

To see this note that the likelihood of the observed data is formally obtained by integrating expression (2.1) over the distribution of $\mathbf{Y}^{miss}$ and $\mathbf{b}$. This can be written as

$$\begin{aligned} \mathrm{L}(\theta, \psi \mid \mathbf{Y}^{obs}, \mathrm{R}, \mathrm{X}) &= \int f(\mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathrm{b} \mid \mathrm{X}, \theta) f(\mathrm{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathrm{X}, \psi)\, d\,\mathbf{Y}^{miss} d\,\mathrm{b} \\ &= \int f(\mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathrm{b} \mid \mathrm{X}, \theta) f(\mathrm{R} \mid \mathrm{X}, \psi)\, d\,\mathbf{Y}^{miss} d\,\mathrm{b} \\ &= f(\mathbf{Y}^{obs} \mid \mathrm{X}, \theta) f(\mathrm{R} \mid \mathrm{X}, \psi) \end{aligned}$$

so that the corresponding log-likelihood is given by

$$\ell_{obs} = \log f(\mathbf{Y}^{obs} \mid \mathbf{X}, \boldsymbol{\theta}) + \log f(\mathbf{R} \mid \mathbf{X}, \boldsymbol{\psi}).$$

Covariates that are predictive of missingness should however be included in the model for the response variable (Carpenter et al., 2002).

An intuitive interpretation is that under an MCAR mechanism the observed values of $\mathbf{Y}$ are assumed to be a random subsample of the intended complete data set conditional on observed covariates (Little and Rubin, 1987). Examples include all design experiments where the investigator has complete control over the values of $\mathbf{R}$. In most cases however, the investigator is faced with data that are not MCAR by design and adopting such a mechanism a posteriori makes assumptions that are unlikely to hold.

2.2.2 *Missing at Random (MAR) mechanism*

Here we assume that the missing-data mechanism depends on the observed component $\mathbf{Y}^{obs}$ of $\mathbf{Y}$ but not on $\mathbf{Y}^{miss}$ or $\mathbf{b}$ conditional on $\mathbf{Y}^{obs}$. Algebraically, $f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{b}, \mathbf{X}, \boldsymbol{\psi}) = f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{X}, \boldsymbol{\psi})$.

Likelihood based inferences are still valid under the separability condition. This is because we can write

$$\begin{aligned}
\mathrm{L}(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{Y}^{obs}, \mathbf{R}, \mathbf{X}) &= \int f(\mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{b} \mid \mathbf{X}, \boldsymbol{\theta}) f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{X}, \boldsymbol{\psi})\, d\mathbf{Y}^{miss} d\mathbf{b} \\
&= \int f(\mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{b} \mid \mathbf{X}, \boldsymbol{\theta}) f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{X}, \boldsymbol{\psi})\, d\mathbf{Y}^{miss} d\mathbf{b} \\
&= f(\mathbf{Y}^{obs} \mid \mathbf{X}, \boldsymbol{\theta}) f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{X}, \boldsymbol{\psi})
\end{aligned}$$

and therefore the observed log-likelihood is

$$\ell_{obs} = \log f(\mathbf{Y}^{obs} \mid \mathbf{X}, \boldsymbol{\theta}) + \log f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{X}, \boldsymbol{\psi}).$$

In longitudinal studies with dropouts, a simple graphical inspection of mean profiles for completers and subjects lost to follow-up could rule out an MCAR assumption and show more evidence of an MAR mechanism (Carpenter et al., 2002). In a more formal approach, logistic regression could be used to investigate if "post-randomisation" $\mathbf{Y}$ predict dropout after accounting for baseline covariates (Carpenter et al., 2002; Jacqmin-Gadda et al., 1997).

A more formal test for the null hypothesis of MCAR against the alternative of MAR in repeated measurement studies was proposed by Diggle (1989). He considers at each time point, possibly within each treatment group, whether those subjects that are about to drop out of the study are characterized by values of the response variable that are significantly different from the average level, the latter obtained by considering all available measurements up to that time point. A set of p-values is obtained, one for each time point within each treatment group, which are tested for departure from a uniform distribution on $(0, 1)$.
In some circumstances data are missing at random by design. For instance, in a longitudinal study involving serial measurements of blood pressure, the investigator may decide to withdraw a patient from the study for ethical reasons if his blood pressure crosses a threshold (Murray and Findlay, 1988).

Estimation

For continuous outcomes in longitudinal studies, any of the proposed methods that handle incomplete data can be used to obtain ML parameter estimates if the missing-data mechanism is MCAR or MAR (Laird and Ware, 1982; Murray and Findlay, 1988; Goldstein, 1986; Schluchter, 1988). There are however some subtle implications for the expected information matrix. The main point is that, under an MAR mechanism, the expected information matrix should be calculated over the marginal unconditional joint distribution of $\mathbf{Y}$ and $\mathbf{R}$ and not the

"naive" sampling distribution of $\mathbf{Y}^{obs}$ (Verbeke and Molenberghs, 1997). In linear mixed models for instance, the former is no longer block diagonal with respect to fixed effect parameters and variance components as would be the case under a MCAR mechanism. Consequently, estimation procedures like the scoring algorithm and estimation of standard errors based on the expected Hessian matrix should use inversion of the full matrix and not just of the fixed-effect block. A more practical alternative is to obtain the variance-covariance matrix of the fixed effect parameter estimates from the observed information.

For incomplete longitudinal discrete data on the other hand, a standard fitting procedure like GEE requires the more stringent assumption of data missing completely at random. This is because only when the probability of missingness does not depend on the outcome (either observed or unobserved), do the estimating equations have mean zero so that an unbiased parameter estimate is given by the root of the score function (Liang and Zeger, 1986; Kenward et al., 1994; Paik, 1997; Laird, 1988).

Fitzmaurice et al. (1995), use the EM algorithm to find ML estimates of the marginal parameters which are valid under the more relaxed MAR assumption. More recently, EM-type algorithms have been proposed for ML estimates for the broader class of models known as generalized linear mixed models. These methods approximate the E-step of the algorithm, which can be intractable in this context, with Monte Carlo or quasi-Monte Carlo integration (McCulloch, 1997; Quintana et al., 1999; Shi and Lee, 2000; Pan and Thompson, 1998).

2.2.3 *Nonignorable outcome-based missing-data mechanism*

In this case the model for the missing-data mechanism is

$$f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{b}, \mathbf{X}, \psi) = f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{X}, \psi)$$

and to obtain ML estimates, we need to simultaneously maximise the joint model for the response and missing data as the separability condition no longer holds. The likelihood of the observed data no longer simplifies as in this case we have

$$\mathrm{L}(\theta, \psi \mid \mathbf{Y}^{obs}, \mathbf{R}, \mathbf{X}) = \int f(\mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{b} \mid \mathbf{X}, \theta) \times f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{X}, \psi)\, d\mathbf{Y}^{miss} d\mathbf{b}. \quad (2.3)$$

Thus, in this setting, an approach which ignores the missing-data mechanism will give biased estimates and, in practice, this bias can often be substantial.

Testing whether a missing-data process is at random or nonignorable requires testing hypotheses about parameters whose standard errors depend crucially on the distribution chosen for the missing data. However, this distribution has to be assumed, because no data were observed. Thus, in practice, we cannot differentiate between the two scenarios. Indeed it is now generally agreed that the role of nonignorable models is to check the sensitivity of the results from a MAR approach (Verbeke and Molenberghs, 1997; Kenward, 1998; Carpenter et al., 2002).

Diggle and Kenward (1994) propose a modelling framework to allow for informative dropouts in longitudinal studies. A logistic model relates the probability of dropping out at each time point to the history of the measurement process up to and including dropout. The likelihood is maximised using the simplex algorithm.

The performance of the approach is tested using simulation. Data are simulated under different assumptions about the dropout mechanism (completely at random, at random and informative). The results show that as long as the dropout mechanism is correctly specified, the method yields unbiased parameter estimates. In the discussion of the paper, Little, Rubin and Laird point out that the method

might not be robust to misspecification of the dropout model and/or the distribution of the response, and that the use of likelihood ratio tests to distinguish between different models can be misleading. Thus, as Verbeke and Molenberghs (1997) propose, the main use of these models is to assess the robustness of inferences from MAR analyses to scientifically plausible nonignorable mechanisms. This point is clearly illustrated by Kenward (1998). He explicitly considers the effect of distributional assumptions and in particular the assumption of multivariate normality for the response variable and how outliers under this assumption cause the models to appear informative. This is because the model tends to impute atypically low (for the particular data set considered) values for the missing observations in order to counterbalance the presence of the "outliers". The informative mechanism is no longer supported if the outlying observations are deleted or if t-distributions are used instead of the multivariate normal.

Within the same modelling framework, the role of distributional assumptions has been further investigated by Attay (1999) who used the Stochastic EM algorithm (Dielbolt and Celeux, 1993) to fit the models. The results reinforce those of Kenward (1998); in a variety of settings the choice of response distribution, with its implicit characterization of outliers, influences whether ML methods detect non-ignorability.

Within a Bayesian framework, Carpenter et al. (2002) fit an outcome-based model to investigate the impact of missing data from a longitudinal asthma study. They assess the sensitivity of the results from a MAR analysis by allowing the coefficient which relates the probability of missing a visit to the unseen measurements in the dropout model to vary within a plausible range of values, with zero corresponding to MAR. The attraction of this approach is that a flexible class of models can be fitted easily using available software and can be used to explore the effect of various missing-data mechanisms in a fairly routine way. The authors

use vague priors, so that their estimates are close to ML estimates. Again, the inferences are potentially sensitive to the distributional assumptions.

For binomial responses Ibrahim and Lipsitz (1996) propose a method for estimating parameters in binomial regression where some measurements are missing and the missing mechanism is thought to be nonignorable. The latter is modelled using logistic regression and the approach is fitted using the EM algorithm by the method of weights. In the E-step, the expectation of the complete-data log-likelihood is obtained as a weighted sum with weights given by conditional binomial probabilities conditioning on the missing indicator; these are readily calculated given current estimates of all relevant parameters. The maximization step involves standard fitting routines for generalized linear models that allow for weights.

Molenberghs et al. (1997) extend the Dale model for longitudinal ordinal data to allow for a nonignorable dropout mechanism. Maximum likelihood parameter estimates are obtained using the EM algorithm. The response and dropout models are assumed independent conditioning on the complete data and this results in a straightforward maximization step.

2.2.4 *Nonignorable random-coefficient-based missing-data mechanism*

Although outcome-based selection models have a direct and intuitive interpretation, there are many examples in the literature of so-called random-coefficient-based models.

Here it is assumed that $f(\mathbf{R} \mid \mathbf{Y}^{obs}, \mathbf{Y}^{miss}, \mathbf{b}, \mathbf{X}, \psi) = f(\mathbf{R} \mid \mathbf{b}, \mathbf{X}, \psi)$. Note that $f_{R|b}$, though not as explicit as in the outcome-based selection model, constitutes an informative dropout mechanism as the dropout distribution depends on both $\mathbf{Y}_i^{obs}$ and $\mathbf{Y}_i^{miss}$ through the (unobserved) random effects (Hogan and Laird, 1997). To see this, note that, unless $f(\mathbf{R} \mid \mathbf{b}, \mathbf{X}, \psi) = f(\mathbf{R} \mid \mathbf{X}, \psi)$, the model for

dropout $\mathbf{R}$ depends on $\mathbf{Y}^{miss}$ as

$$f\left(\mathbf{R}\,|\,\mathbf{Y}^{obs},\mathbf{Y}^{miss},\mathbf{X},\psi\right) = \int f\left(\mathbf{R}\,|\,\mathbf{Y}^{obs},\mathbf{Y}^{miss},\mathbf{b},\mathbf{X},\psi\right) f\left(\mathbf{b}\,|\,\mathbf{Y}^{obs},\mathbf{Y}^{miss}\right) d\mathbf{b}$$
$$= \int f\left(\mathbf{R}\,|\,\mathbf{b},\mathbf{X},\psi\right) f\left(\mathbf{b}\,|\,\mathbf{Y}^{obs},\mathbf{Y}^{miss}\right) d\mathbf{b}.$$

There are circumstances where this approach seems more sensible, especially in longitudinal studies. The probability of missing a visit or dropping out for instance, can be related to rate of change over time of the outcome rather than directly to a specific (potentially missing) outcome, especially when individual measurements are highly variable (Hogan and Laird, 1997).

Wu and Carroll (1988) propose a method for jointly modelling the rate of change of a response variable and a primary informative right-censoring process in follow-up studies where the main focus is on contrasting mean slopes across treatment groups. For a primary right-censoring process they intend loss of follow-up caused by death or withdrawal although the proposed method can be easily extended to account for secondary right-censoring mechanisms.

The response variable is modelled using a linear mixed model with random intercepts and slopes and the probability of death or withdrawal is related to the subject-specific random coefficients using probit regression. The method is compared to simple Unweighted Least Squared (UWLS) and Generalized Least Squared estimates (GLS) of the rate of change; a simulation study shows that these lead to biased estimates of the group mean slopes and between-group differences in the presence of informative right-censoring.

Wu and Bailey (1989) build on these ideas. They demonstrate that, under the model of Wu and Carroll, the conditional expectation of the random slopes given the dropout time is a monotonic increasing (decreasing) function of the dropout time if the parameter that relates the latter to the random slope in

the probit model is negative (positive). Therefore, the subject-specific Ordinary Least Squares (OLS) estimates of the random slopes are related to the (subject-specific) dropout time using a polynomial function of degree L, where L is chosen using stepwise regression and analysis of covariance techniques. It is assumed that the variances of the OLS estimates vary according to time of dropout and treatment group. If these variances are known, a Linear Minimum Variance UnBiased estimator (LMVUB) of the mean group slope is obtained as a weighted average of OLS estimates within each group. A second estimator of the mean group slope is obtained as weighted average of the individual slope estimates with weights found to minimize the Mean Squared Error (Linear Minimum MSE estimator or LMMSE).

Wang-Clow et al. (1995) compare these estimation methods in a simulation study including maximum likelihood and complete-case analysis. Predictably, the results depend heavily on using the correct dropout model; in real situations, though, this cannot be determined from the data. Nevertheless, they report that UWLS estimates are most inefficient though unbiased under informative processes. As we would expect, ML gives the best results under a MAR process in terms of bias and standard errors but biased estimates under an informative dropout mechanism. LMVUB and LMMSE estimators perform well in terms of bias under a nonignorable dropout process but give large standard errors compared to ML and complete-case analysis. Follmann and Wu (1995) generalize the results of Wu and Bailey to other link functions and distributions within the exponential family. They show that the expectation of the conditional distribution of the random effects given the response indicators is a monotonic function of the response indicators if these are conditionally independent given the random effects. This is almost always the case for the kind of models considered here. Therefore an approximate conditional model can be fitted where the usual

marginal (zero) mean of the random effects in the response model is replaced with an approximation to the conditional expectation expressed as a linear function of the response indicators. The approach has been adapted to longitudinal binary data by TenHave et al. (1997).

Subsequently, Follmann and Wu (1999) show that, for the same approximate conditional model, the expectation of the conditional distribution of the random effects can be approximated by considering linear functions of the sufficient statistics of the missingness process. In the case of monotone missingness for instance, these are simply the time of dropout and the sum of all possible dropout times after this.

A different approach was proposed by Schluchter (1992) and DeGruttula and Tu (1994). They assume that the random effects and dropout indicator have a joint multivariate distribution. This allows model fitting using the EM algorithm treating the unobserved random coefficients as missing data. In the E-step, expectations of the complete-data missing sufficient statistics are obtained taking advantage of the assumed multivariate normality. A slight modification extends the approach to individuals that completed the follow-up and thus have unobserved dropout time; this involves finding conditional moments of the truncated normal distribution.

Similar models were considered by Touloumi et al. (1999). The response variable is modelled using a linear mixed model and the dropout indicator using a log-normal model. Like Schluchter (1992), Touloumi *et al.* assume that the log-survival time residuals and the subject-specific random intercept and slope have a trivariate normal distribution with the form of the covariance matrix governing the informativeness of the dropout mechanism (whether or not it allows dependence between the former and the latter).

Maximum restricted likelihood estimates are obtained using an adaptation of the

Restricted Iterative Generalized Least Squares (RIGLS) algorithm (Goldstein, 1986, 1989). Both the fixed and random effects design matrices are modified to take the parameters of the survival model into account. The RIGLS algorithm is then implemented on this enlarged model. A slight modification is made for patients with censored survival times and this involves a nested EM algorithm. At each iteration the E-step finds the conditional expectation of survival time given the data and the fact that it has to be greater than the observed censored time exploiting properties of the conditional moments of the truncated normal distribution. In the M-step new values of the parameters are obtained via RIGLS. The application of the model (in both the uninformative and informative versions) is illustrated using data on CD4 count data.

A simulation study is also performed, with true values based on the analysis of the actual data. Population slope estimates from the model that ignores the primary right-censoring process are biased. This is because, broadly speaking, these estimates are weighted averages of the individual slopes with weights proportional to the number of observations available on each subject (Wu and Carroll, 1988); this tends to overestimate or underestimate the population parameters by not giving much weight to patients that drop out prematurely and have fewer measurements taken.

Goldstein (1999) proposes a bivariate multilevel model for jointly modelling the response and missing-data indicator in repeated measures study. Different subject-specific random effects are considered for the outcome variable and the response indicator and these are assumed to be correlated. A small simulation study shows that the approach corrects the bias from a naïve approach which ignores the missing-data mechanism and works best if there is substantial variation at the individual level in the probability of missingness.

2.3 EM-type algorithms for missing data problems

In the previous Section, the EM algorithm has emerged as the most commonly used tool for dealing with missing data problems. Note the term "missing data" is applied here in a broader sense. It includes, for instance, latent variables; indeed the EM algorithm can be applied to a variety of problems provided that they can be formulated as missing data problems. As such the EM algorithm has an important role to play, since, when missing data are present, direct maximum likelihood estimation can became impossible because obtaining the likelihood of the observed data involves an intractable integral. Little and Rubin (1987) present examples of these difficulties in the case of multivariate data with general patterns of missingness.

Below we describe the EM algorithm. Apart from its "original" version due to Dempster et al. (1977) (hereafter DLR), other variants have been proposed in an attempt to overcome limitations, especially the difficulty of perfoming the E-step. These extensions are, broadly, the Stochastic, Monte Carlo and quasi Monte Carlo EM algorithms.

2.3.1 The EM algorithm

Denote by $\ell(\boldsymbol{\theta} \mid \mathbf{Y})$ the log-likelihood function of the complete data and let $f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \boldsymbol{\theta})$ be the conditional distribution of the missing data given the observed data. We wish to calculate ML estimates of $\boldsymbol{\theta}$ by maximising the observed log-likelihood $\ell(\boldsymbol{\theta} \mid \mathbf{Y}^{obs})$.

Using the factorization

$$f(\mathbf{Y} \mid \boldsymbol{\theta}) = f(\mathbf{Y}^{obs}, \mathbf{Y}^{miss} \mid \boldsymbol{\theta}) = f(\mathbf{Y}^{obs} \mid \boldsymbol{\theta}) f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \boldsymbol{\theta})$$

$\ell(\boldsymbol{\theta} \mid \mathbf{Y}^{obs})$ can be written as

$$\ell(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}) = \ell(\boldsymbol{\theta} \mid \mathbf{Y}) - \log f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \boldsymbol{\theta}). \tag{2.4}$$

Taking the expectation of both sides of (2.4) with respect to the conditional distribution $f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the current estimate of the vector $\boldsymbol{\theta}$, we write

$$\ell(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}) = Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) - H(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}), \tag{2.5}$$

where

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) = \int \ell(\boldsymbol{\theta} \mid \mathbf{Y}) f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}})\, d\mathbf{Y}^{miss}$$

and

$$H(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) = \int \log f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \boldsymbol{\theta}) f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}})\, d\mathbf{Y}^{miss}.$$

Notice that $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ play fundamentally different roles in both $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})$ and $H(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})$. Then we have:

THE EM ALGORITHM

1. E-step: at step t, find

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) = \int \ell(\boldsymbol{\theta} \mid \mathbf{Y}) f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}}^{(t)})\, d\mathbf{Y}^{miss};$$

2. M-step: find $\hat{\boldsymbol{\theta}}^{(t+1)}$ that maximises $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$;
3. iterate between step 1 and 2 until convergence.

Thus, the EM algorithm replaces maximisation of the observed log-likelihood with successive maximisations of the expected value of the complete data log-likelihood where the expectation is taken with respect to the conditional distribution of the missing data given the observed data and current parameter estimates.

DLR show that each step of the algorithm increases the likelihood of the observed data. In fact at the generic iteration $t+1$ we have

$$\begin{aligned}\ell(\hat{\boldsymbol{\theta}}^{(t+1)} \mid \mathbf{Y}^{obs}) - \ell(\hat{\boldsymbol{\theta}}^{(t)} \mid \mathbf{Y}^{obs}) = &\left[Q(\hat{\boldsymbol{\theta}}^{(t+1)} \mid \hat{\boldsymbol{\theta}}^{(t)}) - Q(\hat{\boldsymbol{\theta}}^{(t)} \mid \hat{\boldsymbol{\theta}}^{(t)})\right] \\ &- \left[H(\hat{\boldsymbol{\theta}}^{(t+1)} \mid \hat{\boldsymbol{\theta}}^{(t)}) - H(\hat{\boldsymbol{\theta}}^{(t)} \mid \hat{\boldsymbol{\theta}}^{(t)})\right].\end{aligned} \tag{2.6}$$

The first term on the right-hand side of (2.6) is non-negative from the M-step of the algorithm and the difference of the H functions can be shown to be negative. Thus, if the sequence $\{\hat{\boldsymbol{\theta}}^{(t)}\}$ converges, it converges to a local maximum of the observed log-likelihood $\ell_{obs} = l(\boldsymbol{\theta} \mid \mathbf{Y}^{obs})$. Furthermore, considering the second derivatives of both sides of (2.4) with respect to $\boldsymbol{\theta}$ and taking the expectation with respect to $f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}})$ we can write

$$I(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}) = -\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) + \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} H(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}).$$

The latter expression is an important result as it states that, in the presence of missing data, the *observed information* is the difference between the *complete information* and the *missing information.*

If the log-likelihood of the complete data is linear in the missing data then the E-step reduces to the intuitive idea of imputing mean values for the missing data. In general, if the response distribution belongs to the exponential family, implementation of the algorithm only requires substitution of the missing sufficient statistics with their expected values, again calculated with respect to the conditional distribution of the missing data given the observed values and current parameter estimates.

The EM algorithm has three well-documented limitations (Celeux et al., 1995; McCulloch, 1997; McLachlan and Krishnan, 1997).

First it can converge very slowly depending on the amount of missing information.

Secondly, the E-step can be difficult to carry out if high dimensional integration is involved or no algebraic form of the integral is available. This is particularly true when the algorithm is used to fit informative missing-data models. The conditional expectation of the missing data in these circumstances is known only up to a normalizing constant, resulting in an intractable E-step. Thirdly, the algorithm is only guaranteed to converge to local maxima, hence can be sensitive to the chosen initial values.

2.3.2 The Stochastic EM algorithm

Many authors have proposed modifications of the EM algorithm in an attempt to cope with the above problems, especially the second. One elegant algorithm is the Stochastic EM algorithm (SEM) (Dielbolt and Celeux, 1993; Bock, 1983) in which the expectation step is replaced by a stochastic step where values from the corresponding conditional distribution are sampled to form a pseudo-complete data set. Thus the algorithm is as follows:

THE STOCHASTIC EM ALGORITHM

1. S-step: at step t generate a pseudo-complete data set by drawing $\mathbf{Y}^{miss}$ from $f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}}^{(t)})$;
2. M-step: update $\boldsymbol{\theta}$ to $\hat{\boldsymbol{\theta}}^{(t+1)}$ as the ML estimate of the likelihood of the pseudo-complete data obtained in step 1;
3. iterate between step 1 and 2 until convergence.

Note that in this case convergence of parameters estimates is not pointwise as in the EM algorithm but rather in distribution. After observing convergence, the mean of the last few iterations is generally taken to be the final estimate; alternatively the value corresponding to the maximum of the observed log-likelihood could be used (see Ip in Gilks et al. (1996), Ch. 15). The SEM algortihm has additional advantages. Its stochastic nature prevents the algorithm getting stuck

around insignificant local maxima and its speed of convergence is generally more satisfactory (Celeux et al., 1995).

2.3.3 The Monte Carlo EM algorithm

Wei and Tanner (1990) propose a Monte Carlo implementation of the E-step where $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ is replaced by a average based on a sample of size $m\,(m \gg 1)$ from $f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}}^{(t)})$.

Thus the algorithm is: THE MONTE CARLO EM ALGORITHM

1. Sampling step: at step t form $m\,(m \gg 1)$ pseudo-complete data sets $\mathbf{Y}_1 \ldots \mathbf{Y}_m$ by drawing values from $f(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs}, \hat{\boldsymbol{\theta}}^{(t)})$;
2. Monte Carlo E-step: approximate $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ by

$$\hat{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) = \frac{1}{m} \sum_{j=1}^{m} l(\boldsymbol{\theta} \mid \mathbf{Y}_j);$$

3. M-step: find new estimates $\hat{\boldsymbol{\theta}}^{(t+1)}$ that maximise $\hat{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$;
4. iterate between steps 2 and 3 until convergence.

When $m = 1$, the MCEM algorithm reduces to the SEM algorithm. As $m \to \infty$ the MCEM algorithm behaves like the EM algorithm. The difficulty with the MCEM algorithm is the choice of m. Various authors (Wei and Tanner, 1990; McCulloch, 1997; Quintana et al., 1999) suggest choosing m to be small initially, to take advantage of the robustness of the SEM to starting values, and then increasing it.

As mentioned in Chapter 1, recently MCEM algorithms have been proposed for ML estimation in the context of Generalized Linear Mixed Models (GLMM) and latent variable models where the random effects or latent variables are viewed as missing data (McCulloch, 1997; Quintana et al., 1999; Shi and Lee, 2000).

2.3.4 The quasi-Monte Carlo EM algorithm

In general, the asymptotic rate of approximation of Monte Carlo integration is $O(m^{-1/2})$ where m is the sample size. In an attempt to improve on this, Pan and Thompson (1998) propose the use of quasi-Monte Carlo approximations to the E-step. The idea is to deterministically choose a set of Cumulative-Distribution-Function (CDF)-representative points which are used as integration nodes. These are chosen to minimize the *discrepancy* which is a measure of the absolute error made in the approximation.

For an integral on the q-dimensional unit cube C^q the discrepancy corresponding to a set of integration nodes P_m is

$$D(P_m) = \sup_{\mathbf{x} \in C^q} | U_m(\mathbf{x}) - U(\mathbf{x}) |$$

where $U(\mathbf{x})$ is the cumulative distribution function of the uniform distribution over C^q and $U_m(\mathbf{x})$ is the empirical cumulative distribution of the set of points P_m. Points with the smallest discrepancy lead to a rate of approximation of order $O((\log m)^{q-1}/m)$. Sets of points with discrepancy close to this optimum value have been tabulated and are available (Fang and Wang, 1994). Among these, some are known as Good Lattice Points (GLP). Notice that for large q it could take an impractically large sample size m for the asymptotic errors to be relevant. However, in empirical studies, quasi-Monte Carlo integration produces significantly smaller biases compared to crude Monte Carlo with much less computational effort (Pan and Thompson, 1998).

For $q = 2$, $m = 610$ GLP are represented in Figure 2.3(a). CDF-representative points $\mathbf{b}^*$ from the standard bivariate normal distribution $N_2(0, \mathbf{I}_2)$ can be obtained from the inverse distribution function of $N(0,1)$ on each coordinate of the GLP (Pan and Thompson, 1998). The latter can then be used to

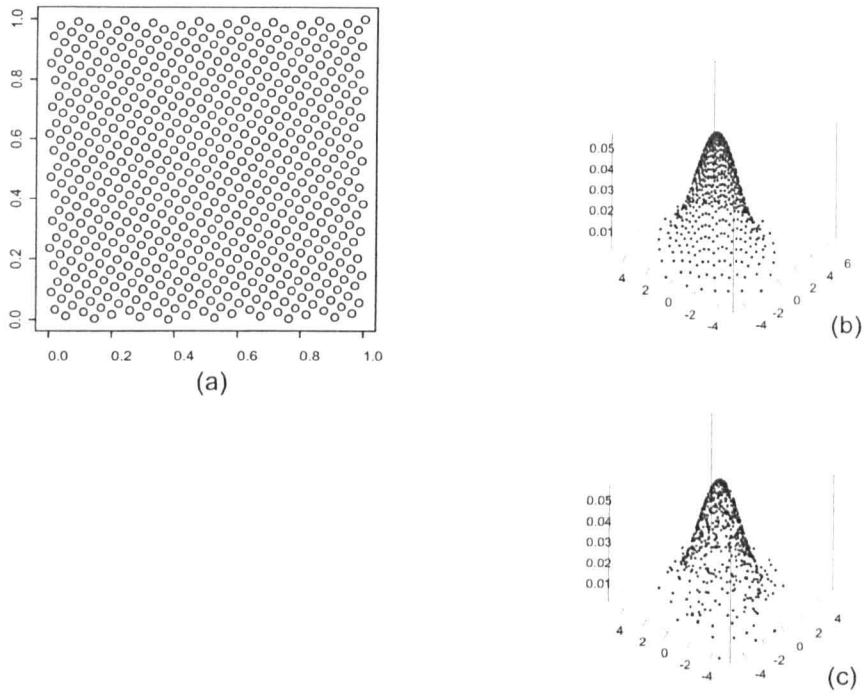

Figure 2.3: (a) Good Lattice Points on the unit square (m=610); (b) 'Random' points from a bivariate normal distribution obtained using (a); (c) Random points from a bivariate normal distribution obtained using the random number generator of Splus®.

obtain CDF-representative points $\mathbf{b}$ from the generic bivarite normal distribution $N_2(0, \Sigma)$ as $\mathbf{b} = \Sigma^{1/2}\mathbf{b}^*$ where $\Sigma^{1/2}$ is the Choleski decomposition of Σ. Figure 2.3(b) shows CDF-representative points from a normal distribution $N_2(0, \mathbf{D})$ where

$$\mathbf{D} = \begin{pmatrix} 3 & 0.8 \\ 0.8 & 3 \end{pmatrix}$$

alongside those obtained from the random number generator in Splus® (c).

The former are clearly more CDF-representative nodes of integration in the sense that the cover the area of integration in a more efficient way. We will explore this further in Section 3.2.

2.4 Discussion

The presence of missing data introduces ambiguity into the analysis that goes beyond the usual statistical imprecision. Such conceptual problems are compounded by additional technical problems. The easiest approach of ignoring the missing data process or, in a follow-up study, discarding subjects with incomplete observations makes strong assumptions about the missing-data mechanism that are frequently both inappropriate and misleading.

On the other hand, a model for the missing-data mechanism requires careful thought and should be developed in the light of what seems a plausible non-response process within the context of a particular study.

Outcome-based models are intuitively appealing as they relate the probability of non-response to the unseen value of the response variable. The results from this approach are usually straightforward to interpret. However, as argued in Subsection 2.2.4, in some circumstances random-coefficient-based models are a more sensible choice.

A related problem is the choice of fitting procedure. Many of the strategies discussed above are problem-specific although EM-type algorithms have emerged as very flexible tools.

In the next chapter, the MCEM and quasi-MCEM algorithms are used to fit random-coefficient-based models for informative dropout processes. The approach is illustrated with Activities of Daily Living (ADL) data from the Parkinson's disease trial.

Chapter 3

Monte Carlo and quasi-Monte Carlo EM algorithms for random-coefficient-based dropout models

In longitudinal studies of neurological disorders, patients are typically followed over several years and measurements of their mental and physical status are recorded at regular intervals in order to assess the progression of the disease and compare the efficacy of different treatments. Many patients, however, will not complete the study and if the reason for a patient's drop-out is related to an unseen outcome then, as mentioned in the previous Chapter, failing to model this could seriously bias the parameter estimates.

Generally in such studies, disease progression causes slow deterioration in an individual's physical and mental abilities which can lead to premature withdrawal. Therefore it seems sensible to relate the probability of dropping out to underlying disease progression, perhaps by expressing it as a function of the individual's own random variation about the overall mean rate of change. In Chapter 2, these models were termed nonignorable random-coefficient-based models (hereafter NIRCBM). They assume separate models for the response variable and the dropout indicator which share a common set of random coefficients.

In this Chapter we show how the Monte Carlo and quasi-Monte Carlo EM algorithms (MCEM and quasi-MCEM respectively), described in Subsection 2.2.3, can be used to fit NIRCB models (Verzilli and Carpenter, 2002a). The methods are illustrated using data from the Parkinson's disease trial described in Subsection 2.1.1.

We start by describing the models used for modelling the response variable (ADL scores) and the dropout mechanism separately; the algorithms used to fit them jointly are illustrated in Section 3.2 where results from a small simulation study are also shown. Section 3.3 presents the results from the application of the proposed approach to the real data and Section 3.4 gives concluding remarks.

3.1 Description of models

The outcome variable considered here consists of ADL scores from the Parkinson's disease trial of Subsection 2.1.1; Figure 2.1 on page 23 shows, at each time point, the mean ADL score for the Ropinirole and Levodopa arms using all available data. A random selection of patient profiles are also plotted in Figures 3.1 and 3.2 for the Ropinirole and Levodopa arm respectively. Here, we ignore data from the first two visits since, as mentioned in Subsection 2.1.1, treatment seems to attain its full effect only by the third visit (Figure 2.1). Thus, we model measurements made at 12, 24, 48, 72, 96, 120, 144, 168, 192, 216 and 240 weeks after randomization.

There were 83 patients still in the study at the third visit in the Levodopa arm and 167 in the Ropinirole arm. Of these, 31 (37%) and 80 (48%) did not complete the study respectively. Dropout rate did not vary greatly across the period considered (mean number of dropouts at each visit 11.1, SD 3.17) with a slight decrease towards the end of the study.

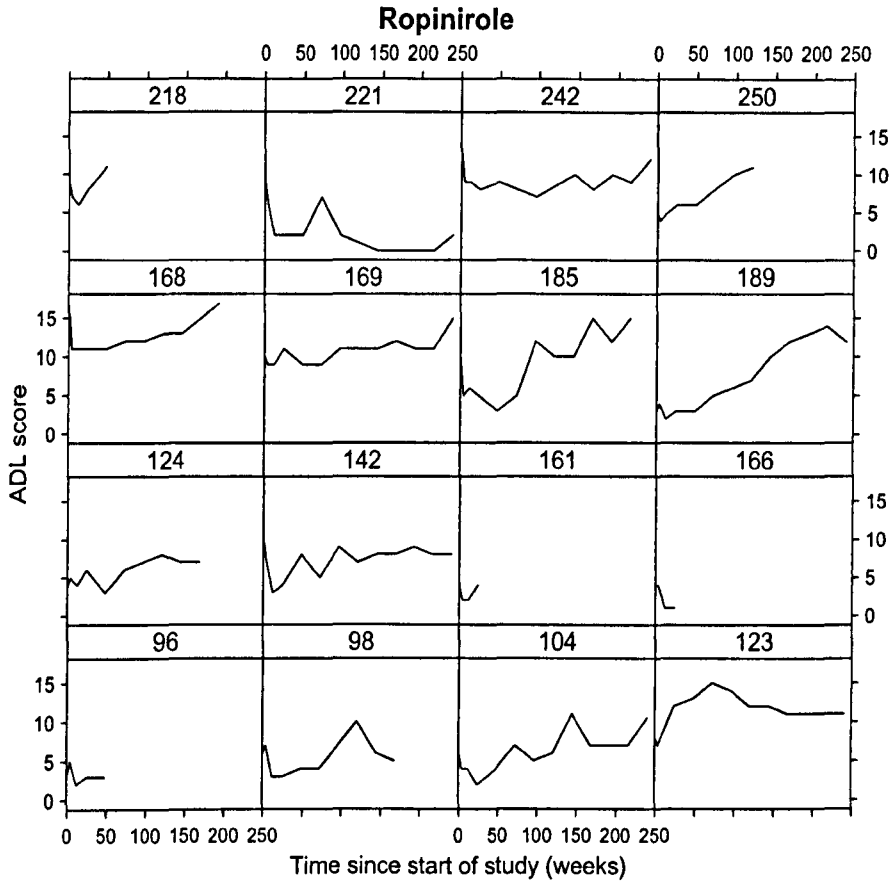

Figure 3.1: Random selection of individual profiles from the Ropinirole treatment arm.

For the generic subject i with n_i observations, let $\mathbf{y}_i^{obs}$ be the vector of observed ADL scores. Further, let $\mathbf{R}_i$ be a vector of dummy variables flagging whether or not the subject is still in the study at each visit with 1 indicating dropout. This has length $(n_i + 1)^{d_i} (n_i)^{1-d_i}$ where d_i equals 1 if the subject withdrew prematurely from the study and is 0 otherwise. After a patient drops out, we assume that the probability of re-entering the study is equal to zero. Few patients had interim missing data with at most two consecutive measurements missing. Here we consider all data available from each patient and therefore our definition of dropout time implies that no further measurements are available

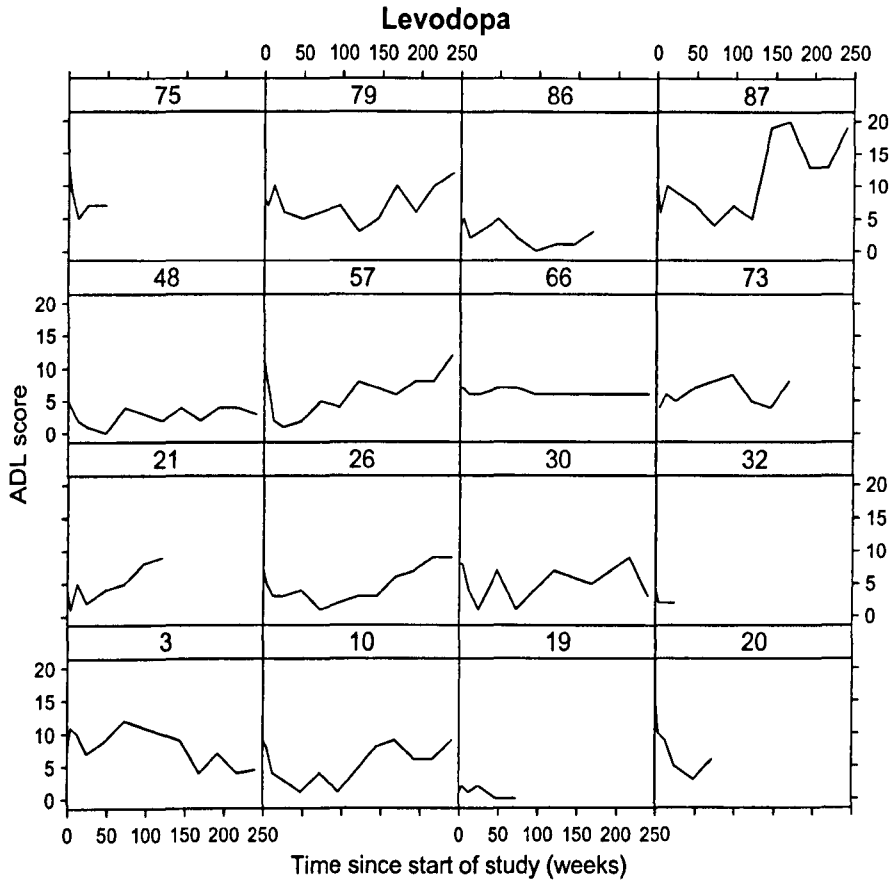

Figure 3.2: Random selection of individual profiles from the Levodopa treatment arm.

after this time point. We indicate with y_i^{miss} the first unseen measurement that would have been taken at visit $n_i + 1$ had the patient not dropped out. Recall from Subsection 2.1.4 that in NIRCBM the joint distribution of $\mathbf{y}_i^{obs}, y_i^{miss}$, $\mathbf{R}_i$ and the set of random effects $\mathbf{b}_i$ given some covariates $\mathbf{x}_i$ is factorized as

$$f_{y,R,b}\left(\mathbf{y}_i^{obs}, y_i^{miss}, \mathbf{R}_i, \mathbf{b}_i|\mathbf{x}_i\right) = f_{y|b}\left(\mathbf{y}_i^{obs}, y_i^{miss}|\mathbf{b}_i, \mathbf{x}_i\right) f_{R|b}\left(\mathbf{R}_i|\mathbf{b}_i, \mathbf{x}_i\right) f_b\left(\mathbf{b}_i|\mathbf{D}\right). \tag{3.1}$$

The observed likelihood is then obtained by integrating (3.1) over the distribution of y_i^{miss} and $\mathbf{b}_i$.

We further assume that the elements of both $\mathbf{y}_i$ and $\mathbf{R}_i$ are conditionally independent given the random effects, that is

$$f_{y|b}(\mathbf{y}_i|\mathbf{b}_i, \mathbf{x}_i) = \prod_{j=1}^{n_i} f_{y|b}(y_{ij}|\mathbf{b}_i, \mathbf{x}_i) \tag{3.2}$$

and

$$f_{R|b}(\mathbf{R}_i|\mathbf{b}_i, \mathbf{x}_i) = \prod_{j=1}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} f_{R|b}(R_{ij}|\mathbf{b}_i, \mathbf{x}_i)\,. \tag{3.3}$$

We now illustrate separately the models for the ADL scores and the dropout mechanism before describing the modified MCEM algorithm used to fit them jointly.

3.1.1 Linear mixed model for ADL scores

We first chose an appropriate form for the covariance matrix by fitting various structures with a saturated model for the mean structure. The saturated mean structure is obtained by assuming different parameters for the mean ADL score at each visit time within each treatment group.

Table 3.1 shows the results where dashes distinguish the models used for the between and within subject sources of variation. Although the heterogeneous Toeplitz structure (see Verbeke and Molenberghs (1997)) seems to fit well, the more parsimonious covariance structure corresponding to random intercepts and random slopes alone exhibits an acceptable fit (eighth row of Table 3.1). This is confirmed by Figure 3.3 where the variance functions corresponding to the different structures are plotted alongside the sample variances. Therefore, mean ADL scores have been modelled using a linear mixed model with intercepts and slopes varying randomly across subjects. Quadratic trends were also considered but did not improve the fit significantly.

Model	−2 Log-lik.	AIC[a]	SBC[b]	df
Unstructured	9915	-5035	-5253	78
Heter. Toeplitz	9990	-5018	-5082	23
RI[c]/RS[d] – Toeplitz	10053	-5041	-5083	16
RS – Toeplitz	10057	-5041	-5078	14
RI/RS – Exp. Ser. Corr.	10155	-5049	-5065	5
RI/RS – Gauss. Ser. Corr.	10155	-5083	-5100	5
RI/RS – Power Ser. Corr.	10087	-5048	-5062	5
RI/RS	10155	-5081	-5092	4
RI	10489	-5246	-5252	2

Table 3.1: Results of fitting different covariance structures to the ADL score data with a saturated mean structure. [a]Aikake Information Criterion; [b]Swartz Bayesian Criterion — high values indicate a better fit; [c]random intercepts; [d]random slopes.

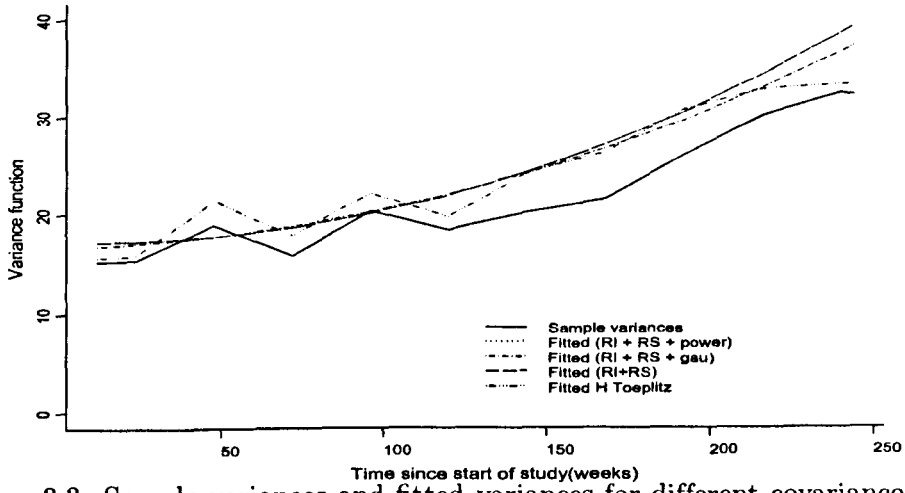

Figure 3.3: Sample variances and fitted variances for different covariance structures. See Table 3.1 for explanation of abbreviations in legend.

Finally, the variance structure is assumed the same across the two arms.

Baseline ADL score has been included as a covariate. We found no statistically significant overall treatment-by-time interaction (p=0.97) thus the model considers a common fixed slope across the two treatment arms.

For the generic i-th subject and j-th visit, the model is

$$\begin{aligned} y_{ij} &= \beta_0 + b_{i0} + (\beta_1 + b_{i1})\text{week}_j + \beta_2 \text{treat}_i \\ &\quad + \beta_3 \text{BaseADL}_i + e_{ij} \end{aligned} \tag{3.4}$$

where $i = 1, \ldots, 250$, $j = 1, \ldots, n_i$, $treat_i = \{1(\text{Ropinirole}), 0(\text{Levodopa})\}$, $e_{ij} \sim N(0, \sigma^2)$, $\mathbf{b}_i \sim N_2(0, \mathbf{D})$ and

$$\mathbf{D} = \begin{pmatrix} \sigma^2_{b_{i0}} & \sigma_{b_{i0},bi1} \\ \sigma_{b_{i0},bi1} & \sigma^2_{b_{i1}} \end{pmatrix}.$$

Parameter estimates for model (3.4) are reported in the second column of Table 3.4 on page 64.

3.1.2 A random-coefficient-based model for the dropout mechanism

Simple preliminary analyses can be useful for exploring the dropout process (Carpenter et al., 2002). Figure 3.4 shows the mean ADL scores (95% CI) for patients still in the study at the next visit and for those who are about to drop out. With few exceptions, the latter show a higher mean ADL score than the former, which suggests that patients with higher mean ADL scores are more likely to withdraw from the study prematurely. This is a clinically plausible scenario since high ADL scores indicate increased difficulty with the activities of daily life.

Logistic regression has been used to identify possible predictors of dropout by five years from baseline covariates, person-specific intercept and slope and the

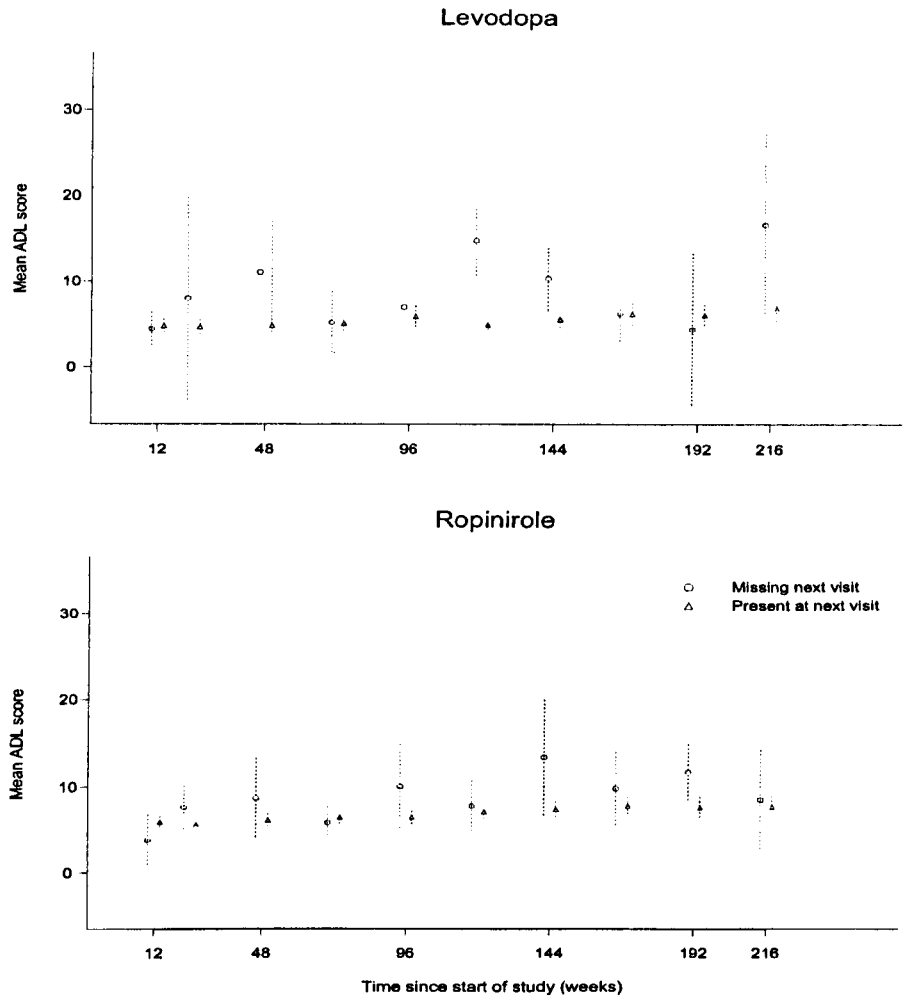

Figure 3.4: Mean ADL score by status of patients at next visit. Dotted lines represent 95%CI.

last observed ADL score (Table 3.2). The results indicate that the initial level of ADL and the rate of progression over time are the principal predictors; these being the only ones retained in a stepwise selection of covariates.

Since we have a monotone pattern of missingness and long intervals between successive measurements, we model the dropout mechanism using a discrete time proportional hazards model (Aitkin et al., 1990). We include the explanatory variables highlighted in the previous paragraph, namely baseline ADL and the subject-specific random slope b_{i1}.

Variable	OR	95 % CI		P-value
		Lower	Upper	
Levodopa/Ropinirole	0.65	0.38	1.12	0.12
Age	1.02	0.99	1.04	0.25
Age at onset of desease	1.01	0.98	1.04	0.39
Baseline ADL	1.06	1.01	1.12	0.02
Last observed ADL	0.99	0.95	1.04	0.67
PD stage[a]	1.49	0.88	2.53	0.14
Selegiline[b]	1.23	0.75	2.03	0.41
Male/Female	1.22	0.73	2.05	0.45
Random intercept[c]	1.14	1.01	1.28	0.03
Random slope[d]	1.26	1.06	1.49	0.01

Table 3.2: Parameter estimates from univariate logistic regression for the probability of dropout by year 5. [a]Stage of PD disease measured on Hoehn and Yahr scale; [b]Concomitant treatment with Selegiline (1=yes, 0=no); [c]As calculated from model (3.4); [d]As calculated from model (3.4) and per 0.01 increase.

The generic i-th subject either drops out at time $t_i = \text{week}_j$ or is finally censored at time $t_i = \text{week}_{11}$. Note that the discrete time proportional hazards model corresponds to a generalized linear model where the censoring indicator has a Bernoulli distribution with complementary log-log link function.

In fact, denoting by f_j, s_j and h_j the probability mass, survivor and hazard function of the resulting discrete survival distribution, we have

$$f_j = s_j - s_{j+1},$$
$$h_j = \frac{f_j}{s_j},$$
$$h_j = \frac{s_j - s_{j+1}}{s_j},$$

$$\frac{s_{j+1}}{s_j} = 1 - h_j,$$
$$s_{j+1} = \prod_{r=1}^{j} \frac{s_{r+1}}{s_r} = \prod_{r=1}^{j} (1 - h_r).$$

The contribution to the likelihood for subject i is then

$$L_i = f_{t_i}^{R_{it_i}} s_{t_i+1}^{1-R_{it_i}} = \prod_{j=1}^{t_i} h_{ij}^{R_{ij}} (1 - h_{ij})^{1-R_{ij}}$$
$$= \prod_{j=1}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} h_{ij}^{R_{ij}} (1 - h_{ij})^{1-R_{ij}},$$

and for the proportional hazards model we have

$$h_{ij} = 1 - \frac{s_{ij+1}}{s_{ij}}$$
$$= 1 - \exp\left[-\exp\left(\boldsymbol{\beta}'\mathbf{x}_i\right)\left\{H_0\left(t_{j+1}\right) - H_0\left(t_j\right)\right\}\right]$$
$$= 1 - \exp\left[-\exp\left(\boldsymbol{\beta}'\mathbf{x}_i + \gamma_j\right)\right],$$

where $\gamma_j = \log\left\{H_0\left(t_{j+1}\right) - H_0\left(t_j\right)\right\}$.

The model for the dropout mechanism can then be written as

$$Pr\left(R_{ij} = 1\right) = \Pr\left(t_i = \text{week}_j \middle| t_i > \text{week}_{j-1}\right) = h_{ij}$$
$$= 1 - \exp\left[-\exp\left(\alpha_0 + \alpha_1 \text{treat}_i + \alpha_2 b_{i1} + \alpha_3 \text{BaseADL}_i + \gamma_l\right)\right], \tag{3.5}$$

where $i = 1, \ldots, 250$, $j = 1, \ldots, (n_i + 1)^{d_i} (n_i)^{1-d_i}$ and γ_l are a set of contrasts for week_l, $l = 2, \ldots, t_i$.

3.2 Monte Carlo EM and quasi-Monte Carlo algorithms for random-coefficient-based dropout models

Theoretically, the EM algorithm could be used to find the ML estimates of the parameters in (3.4) and (3.5). However, the E-step would involve finding the expectation of the complete data log-likelihood

$$E_{b|y,R}\left[\sum_{i=1}^{250}\log f_{y|b}\left(\mathbf{y}_i|\mathbf{b}_i,\mathbf{x}_i,\boldsymbol{\beta},\sigma^2\right)f_{R|b}\left(\mathbf{R}_i|\mathbf{b}_i,\mathbf{x}_i,\boldsymbol{\alpha},\boldsymbol{\gamma}\right)f_b\left(\mathbf{b}_i|\mathbf{D}\right)\right] \tag{3.6}$$

where the expectation is with respect to the distribution of the random effects $\mathbf{b}_i$ (the missing data in this perspective) given the data, the vector of censoring indicators $\mathbf{R}_i$ and current estimates of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, σ^2 and $\mathbf{D}$. The intractable conditional distribution of $b|y,R$ makes direct calculation of this expectation impossible; hence the basic EM algorithm cannot be used.

As discussed in Subsection 2.2.3, McCulloch (1997) proposes a Monte Carlo implementation of the E-step in the context of generalized linear mixed models. A nested Metropolis-Hastings (MH) algorithm gives samples of the random effects $\mathbf{b}_i$ which are then used to calculate the Monte Carlo approximation to (3.6). This approach can be readily modified to incorporate an informative dropout mechanism. Our target distribution can be written as

$$f\left(\mathbf{b}_i|\,\mathbf{y}_i,\mathbf{R}_i,\mathbf{x}_i\right)=\frac{f_{y|b}\left(\mathbf{y}_i|\,\mathbf{b}_i,\mathbf{x}_i\right)f_{R|b}\left(\mathbf{R}_i|\,\mathbf{b}_i,\mathbf{x}_i\right)f_b\left(\mathbf{b}_i|\,\mathbf{D}\right)}{f_{y,R}\left(\mathbf{y}_i,\mathbf{R}_i|\,\mathbf{x}_i\right)}. \tag{3.7}$$

To sample from (3.7) using a MH algorithm involves choosing a proposal distribution $h\left(\mathbf{b}_i\right)$ and accepting candidate values $\mathbf{b}_i^{(k)}$ with probability

$$\min\left(1,\frac{f_{y|b}\left(\mathbf{y}_i|\,\mathbf{b}_i^{(k)},\mathbf{x}_i\right)f_{R|b}\left(\mathbf{R}_i|\,\mathbf{b}_i^{(k)},\mathbf{x}_i\right)f_b\left(\mathbf{b}_i^{(k)}|\,\mathbf{D}\right)h\left(\mathbf{b}_i^{(k-1)}\right)}{f_{y|b}\left(\mathbf{y}_i|\,\mathbf{b}_i^{(k-1)},\mathbf{x}_i\right)f_{R|b}\left(\mathbf{R}_i|\,\mathbf{b}_i^{(k-1)},\mathbf{x}_i\right)f_b\left(\mathbf{b}_i^{(k-1)}|\,\mathbf{D}\right)h\left(\mathbf{b}_i^{(k)}\right)}\right).$$

Taking $h(.) \equiv f_b(.)$ i.e. choosing the candidate distribution to be the marginal distribution of the random effects, simplifies considerably both the processes of drawing new candidate values and calculating the acceptance probability. The latter reduces to the ratio of the likelihoods corresponding to $\mathbf{b}_i^{(k)}$ and $\mathbf{b}_i^{(k-1)}$.

The MCEM algorithm is thus:

1. Initialization: given initial values for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, σ^2 and $\mathbf{D}$ obtain a sample from the target distribution (3.7) using the MH algorithm.
2. M-step: given a sample of size m from 1., the maximisation step reduces to finding the maximum of a Monte Carlo approximation of (3.6), that is

$$\max \left[\frac{1}{m} \sum_{k=1}^{m} \sum_{i=1}^{250} \ln f_{y|b}\left(\mathbf{y}_i | \mathbf{b}_i^{(k)}, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2\right) f_{R|b}\left(\mathbf{R}_i | \mathbf{b}_i^{(k)}, \mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) f_b\left(\mathbf{b}_i^{(k)} | \mathbf{D}\right) \right].$$

From (3.2) new estimates of $\boldsymbol{\beta}$ are obtained by fitting a linear model regressing the $m \sum_{i=1}^{250} n_i$ independent observations $y_{ij} - b_{i0}^{(k)} - b_{i1}^{(k)}\text{week}_j$ on the covariates in (3.4) where $k = 1, \ldots, m$. Similarly from (3.3) new estimates of the vector $\boldsymbol{\alpha}$ of parameters of the dropout model can be obtained fitting a generalized linear model with complementary log-log link function to the expanded data set. Finally the covariance matrix of the sampled $\mathbf{b}$'s yields new estimates of the elements of $\mathbf{D}$.

We have found that the performance of the MCEM algorithm is improved by allowing a long burn-in phase in the MH (typically 1500 iterations). The sample size m kept at each iteration can be moderate at the beginning and can be increased as convergence occurs as discussed in Chapter 2. We started with a sample size of 50 and increased it to 1000 by iteration 100 as, by then, parameter estimates began to show stability.

3. Sampling step: given current estimates of relevant parameters from 2. draw

m new values of $\mathbf{b}_i$ using the MH algorithm.

4. Iterate between 2. and 3. until convergence.

An alternative approach is to use the quasi-Monte Carlo EM algorithm described in Subsection 2.3.4. This avoids the nested MH sampling step. Given current estimates of the elements of $\mathbf{D}$ at iteration t, $\hat{\mathbf{D}}$, CDF-representative points are obtained from $f(\mathbf{b}|\hat{\mathbf{D}})$ as described in Subsection 2.3.4 which are then used to approximate (3.6) as

$$\sum_{k=1}^{m}\sum_{i=1}^{250} w_i^{(k)} \ln f_{y|b}\left(\mathbf{y}_i|\,\mathbf{b}_i^{(k)}, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2\right) f_{R|b}\left(\mathbf{R}_i|\,\mathbf{b}_i^{(k)}, \mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) f_b\left(\mathbf{b}_i^{(k)}|\,\mathbf{D}\right)$$

where, from (3.1), the weights $w_i^{(k)}$ are defined as

$$w_i^{(k)} = \frac{f_{y|b}\left(\mathbf{y}_i|\,\mathbf{b}_i^{(k)}, \mathbf{x}_i, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) f_{R|b}\left(\mathbf{R}_i|\,\mathbf{b}_i^{(k)}, \mathbf{x}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}\right)}{\sum_{k=1}^{m} f_{y|b}\left(\mathbf{y}_i|\,\mathbf{b}_i^{(k)}, \mathbf{x}_i, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) f_{R|b}\left(\mathbf{R}_i|\,\mathbf{b}_i^{(k)}, \mathbf{x}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}\right)}.$$

Again, the assumption of conditional independence of the y_{ij} and R_{ij} leads to a straightforward M-step taking the weights $w_i^{(k)}$ into account.

3.2.1 Calculation of standard errors

Standard errors have been calculated using a conditional parametric bootstrap approximation of Louis' formula as suggested by Diebold and Ip (in Gilks et al. (1996) editors).

The Louis identity relates the observed log-likelihood to the complete data log-likelihood,

$$-\ddot{\ell}_{obs} = E\left[-\ddot{\ell}_{compl}\right] - \mathrm{cov}\left[\dot{\ell}_{compl}\right], \tag{3.8}$$

where $\ddot{\ell}$ and $\dot{\ell}$ refer to the Hessian matrix and score vector of the log-likelihood function respectively and the expectations are with respect to the conditional distribution (3.7). In this setting, conditioning of the vector $\mathbf{b}_i$, the distributions of the responses y_{ij}, the dropout indicators R_{ij} and the marginal distribution of the $\mathbf{b}_i$ are mutually independent. This results in $\ddot{\ell}_{compl}$ being a block diagonal matrix with three blocks corresponding to the parameters of the model for the response model, dropout indicator and marginal distribution of the random effects. More details and expressions for $\ddot{\ell}$ and $\dot{\ell}$ are given in Appendix A.

3.2.2 A simulation study

We performed a simulation study to assess the performance of the proposed method. We considered $N = 200$ subjects in two groups of size 100 and $T = 5$ time points from the following model

$$y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})t_j + \beta_2 \text{group}_i + \beta_3 \text{group}_i t_j + e_{ij}, \tag{3.9}$$

where $i = 1, \ldots, 200$, $j = 1, \ldots, n_i$, $\text{group}_i = 0, 1$ and e_{ij} and $\mathbf{b}_i$ are as in (3.4).

In the model for the dropout process the probability of dropping out at time j is

$$\Pr(R_{ij} = 1) = 1 - \exp(-\exp(\alpha_0 + \alpha_1 b_{i0} + \alpha_2 b_{i1} + \alpha_3 \text{group}_i b_{i1} + \gamma_l)) \tag{3.10}$$

where $i = 1, \ldots, 200$, $j = 1, \ldots, t_i$ and γ_l are a set of contrasts as in (3.5), $l = 2, \ldots, t_i$. The true values of α_2 and α_3 were fixed at 0.20 and 3.00 implying that subjects whose response increased at a greater rate had a higher probability of dropping out and that the dependence of the dropout process on subject-specific slopes differs across the two groups.

Parameters	True values	Estimates (SE)	
		Missing at random	MCEM with informative dropout model
β_0 (Intercept)	1.00	1.04 (0.10)	1.03 (0.10)
β_1 (Time)	2.00	1.45 (0.10)	1.84 (0.04)
β_2 (Group)	3.00	3.04 (0.19)	3.08 (0.14)
β_3 (Time-by-group)	2.00	1.62 (0.16)	1.99 (0.10)
α_0 (Intercept)	−1.00		−1.19 (0.24)
α_1 (Random intercept)	1.00		1.20 (0.65)
α_2 (Random slope)	0.20		0.18 (0.33)
α_3 (Random slope-by-group)	3.00		3.27 (1.64)
γ_2	1.00		1.24 (0.48)
γ_3	1.00		1.14 (0.67)
γ_4	1.00		1.12 (0.78)
γ_5	1.00		1.35 (0.89)
σ^2	1.00	1.05 (0.07)	1.02 (0.03)
σ^2_{int}	1.00	0.82 (0.14)	0.86 (0.09)
σ^2_{slope}	1.00	0.43 (0.09)	0.78 (0.10)
$\sigma_{int,slope}$	0.50	0.12 (0.05)	0.38 (0.04)

Table 3.3: Parameter estimates for models (3.9) and (3.10).

We first fitted a model which ignores the dropout process. As expected (column 3 of Table 3.3), estimates for the overall slope and the interaction parameter β_3 are both biased downwards as are those of the variance components . By contrast, using the MCEM algorithm to fit the informative dropout model removes much of the bias.

The quasi-Monte Carlo EM algorithm leads to similar results. Figure 3.5

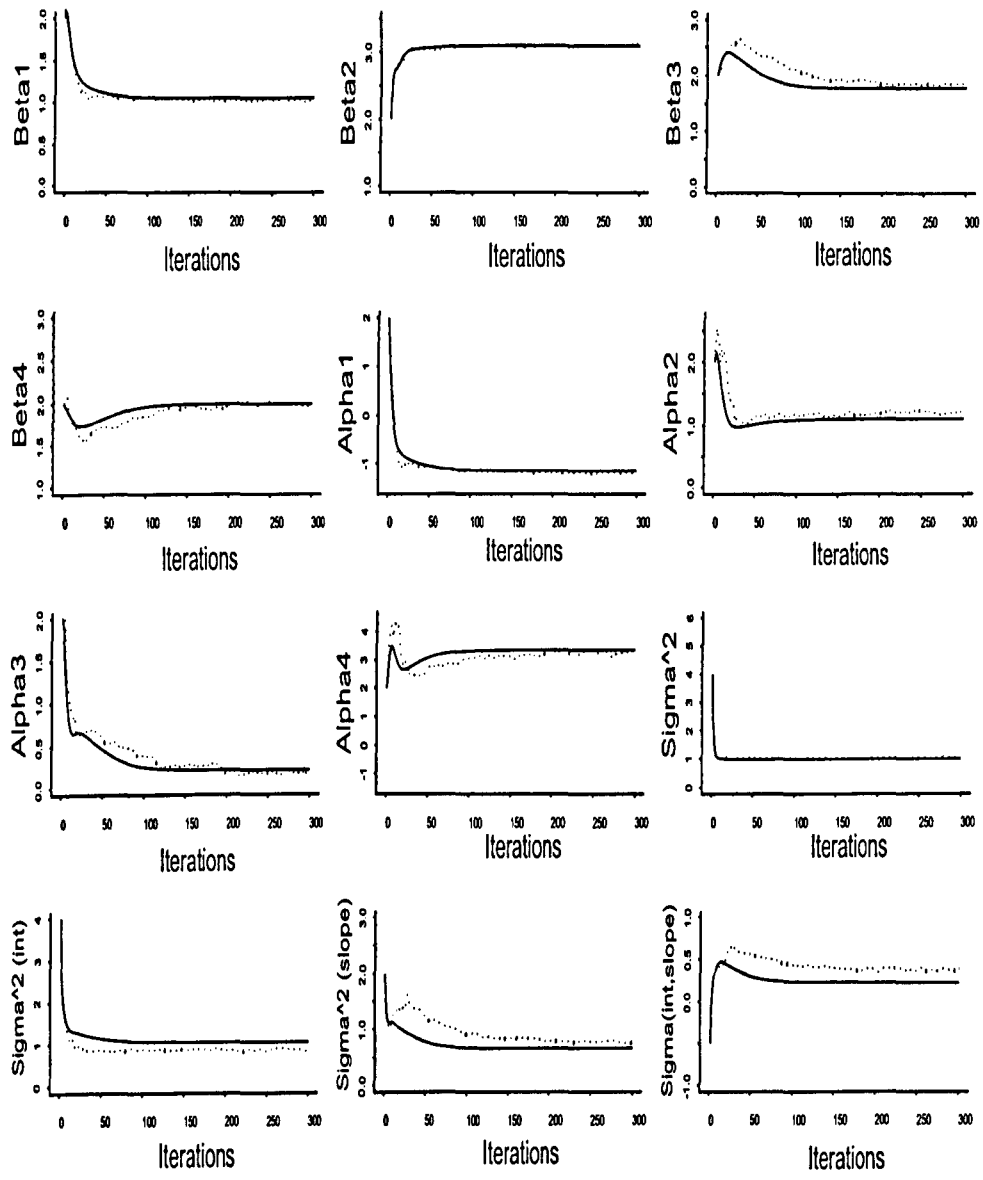

Figure 3.5: Sequences of parameter estimates for models (3.9) and (3.10) jointly fitted using a MCEM algorithm (dotted line) and a quasi-MCEM algorithm (solid line).

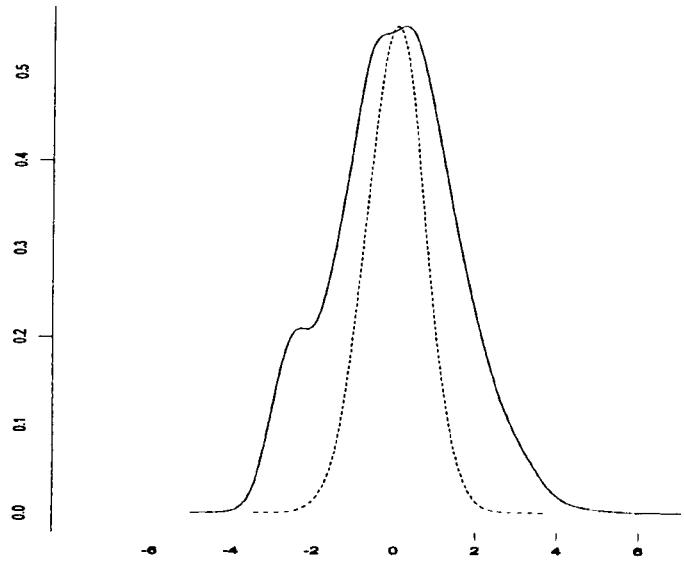

Figure 3.6: Density plot of random slopes from (3.7) sampled using the MH algorithm described in Section 3.2 (solid line) and from a MAR analysis (dotted line).

shows sequences of parameter estimates using the two approaches. The only noticeable feature is the underestimation of the covariance of the random effects in the qMCEM algorithm. Furthermore, in the latter, the computational time saved by avoiding the nested M-H algorithm is outweighted by the fact that a large sample is required in order to reproduce the features of (3.7) using the weights w_i. Thus, in the application to the real data set we used the MCEM algorithm.

Figure 3.6 shows the density plot of random slopes drawn using the MCEM algorithm. The procedure adjusts the missing at random analysis by accepting more small positive and large negative random deviations from the population average, corresponding to completers and withdrawals respectively, thus correcting the estimate of the fixed effects upwards.

3.3 Application to Parkinson's Disease trial

In the application to the Parkinson's disease data set, the main focus has been on assessing the sensitivity of the results to different clinically plausible assump-

tions about the dropout process. Results from a missing at random analysis which ignores the dropout process are shown in the second column of Table 3.4.

We first tested the significance of a random slope-by-treatment interaction term in the dropout model; that is, we tested whether the dependence of the dropout mechanism on the latent progression of the illness varied across the two treatment arms. For this purpose we introduced an extra term α_4 in (3.5) and a treatment-by-time interaction term β_4 in (3.4). Results are reported in the fourth column of Table 3.4.

The dependence of the dropout process on the random slopes seems particularly significant with α_2 = 1.48 (0.65) causing the overall time trend β_2 to increase when compared to the MAR estimate. However, there is no evidence that the dropout mechanism varies across the two arms; the interaction is not statistically significant (α_4 = -0.68 (0.69)) and neither is β_4. The treatment effect remains almost unchanged.

We thus fit simpler models (3.4) and (3.5) with no interaction terms. The main feature is again an increase in the overall time trend when compared to the missing at random analysis, though this is not statistically significant. Treatment effect and other fixed parameter estimates remain virtually unchanged.

The conclusions from the MAR analysis for this data do not appear to be very sensitive to the MAR assumptions. The exception is the slight underestimate of the rate of increase of the ADL score in the MAR model.

3.4 Discussion

In this Chapter we have shown how random-coefficient-based dropout models can be fitted using Monte Carlo and quasi-Monte Carlo versions of the EM algorithm. The approach allows great flexibility in both the response model, which can be continuous or discrete, and in the model for the dropout process which, we feel,

Parameters	Estimates (SE)		
	MAR^a	$M1^b$	$M2^c$
β_0 (Levodopa)	4.332 (0.312)	4.387 (0.258)	4.359 (0.317)
β_1 (Time)	0.017 (0.002)	0.020 (0.002)	0.021 (0.002)
β_2 (Ropinirole-Levodopa)	1.183 (0.377)	1.105 (0.312)	1.126 (0.263)
β_3 (Baseline ADL)	0.459 (0.041)	0.457 (0.035)	0.456 (0.055)
β_4 (Time-by-Treatment)			−0.001 (0.001)
α_0 (Intercept)		−10.193 (8.580)	−10.226 (8.301)
α_1 (Ropinirole-Levodopa)		0.325 (0.215)	0.437 (0.268)
α_2 (Random slope)		0.937 (0.291)	1.482 (0.652)
α_3 (Baseline ADL)		0.046 (0.019)	0.046 (0.019)
α_4 (Random slope-by-treatment)			−0.684 (0.686)
γ_2		6.801 (8.583)	6.735 (8.303)
γ_3		7.176 (8.582)	7.113 (8.301)
γ_4		7.102 (8.583)	7.039 (8.302)
γ_5		7.328 (8.582)	7.263 (8.302)
γ_6		6.814 (8.586)	6.759 (8.305)
γ_7		7.357 (8.583)	7.299 (8.302)
γ_8		6.696 (8.588)	6.643 (8.303)
γ_9		7.363 (8.584)	7.314 (8.304)
γ_{10}		7.356 (8.584)	7.307 (8.307)
γ_{11}		6.990 (8.587)	6.940 (8.308)
σ^2	5.441 (0.205)	5.458 (0.197)	5.463 (0.198)
σ^2_{int}	6.244 (0.811)	5.948 (0.744)	5.924 (0.668)
σ^2_{slope}	0.228 (0.030)	0.246 (0.025)	0.248 (0.018)
$\sigma_{int,slope}$	−0.151 (0.198)	−0.068 (0.156)	−0.060 (0.116)

Table 3.4: Application to the PD trail: parameter estimates under different assumptions about the dropout mechanism. [a]Missing At Random. [b]Models (3.4) and (3.5). [c]Models (3.4) and (3.5) with different slopes across the two arms and random slope-by-treatment group interaction in the dropout model.

should reflect plausible clinical hypotheses. Thus, the sensitivity of inference to assumptions about the dropout mechanism can be assessed. If the inference is sensitive to MAR, then the conclusions of the analysis depend on inherently untestable assumptions about the dropout process. Otherwise, as in the example in Section 3, a sensitivity analysis confirms that the conclusions are robust to assumptions about the dropout mechanism. In the next Chapter, more 'formal' approaches to sensitivity analysis for missing data problems, based on generalizations of Cooks' distance (Cook, 1986), will be presented and discussed.

Chapter 4

Sensitivity analysis

In the previous Chapter we argued that the reason for jointly modeling the response variable and the dropout mechanism is to assess the sensitivity of the conclusions from a MAR analysis to plausible assumptions about the dropout or missing-data mechanism.
Verbeke and Molenberghs (2000) suggest a 'formal' sensitivity analysis in the context of informative outcome-based dropout models; they consider the Diggle and Kenward model (see Subsection 2.2.3) and use local influence to explore the sensitivity of the MAR model to informative dropout. This approach complements the methods of the previous Chapter.

Sensitivity analysis using local influence is described in more detail in the first Section of this Chapter where we derive an extension to NIRCB models. In the second Section, a sampling-based sensitivity analysis for NIRCB models is presented, whereby rather that estimating the parameters relating the probability of dropping out to the subject-specific random effects in the dropout model, these are allowed to vary over sensible ranges. The resulting effect on the parameters of interest is then measured using finite-element response surface methods.

As mentioned in Chapter 2, in the Parkinson's disease trial patients could receive open-label Levodopa supplementation if their symptoms were not ad-

equately controlled by the randomised treatment. There are close similarities between missing data and rescue medication problems as in both cases, the treatment effect that would have been observed (had all data been collected or had patients remained on the allocated treatment, respectively) are potentially biased. Thus, in the final Section of this Chapter, the sampling-based sensitivity analysis is used to assess the effect of rescue medication in the Parkinson's Disease trial. Results from using other methods based on summary statistics (White et al., 2001) and Robins' Inverse Probability of Censoring Weighted estimators and Structural Nested Mean and Distribution models are also considered (Robins and Finkelstein, 2000; Robins and Greenland, 1994; Robins, 1992, 1998).

4.1 Sensitivity analysis using Local Influence

The idea behind local influence is to investigate how stable the parameter estimates from a particular model are to small perturbations of the data. Cook (1986) introduced the concept of likelihood displacement as a measure of stability. The objective is to assess the robustness of the model as a q-dimensional vector of weights $\mathbf{w}$ varies in $\mathbb{R}^q$. In what follows, $q \equiv n$, the number of subjects. In particular, the model tested corresponds to a particular value of $\mathbf{w}$ which we indicate as $\mathbf{w}^*$.

For the random-coefficient-based model (3.5), the vector of weights $\mathbf{w} = (w_1, \ldots, w_n)$ relates to the subject-specific random slope b_{i1} as

$$\begin{aligned} \Pr(R_{ij} = 1) &= \Pr(t_i = \text{week}_j | t_i > \text{week}_{j-1}) = h_{ij} \\ &= 1 - \exp\left[-\exp(\alpha_0 + \alpha_1 \text{treat}_i + \alpha_2 \text{BaseADL}_i + w_i b_{i1} + \gamma_l)\right]. \end{aligned} \tag{4.1}$$

The model tested is the one corresponding to a MAR analysis with $\mathbf{w} = \mathbf{w}^* = \mathbf{0}$, a vector of zeros, and log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{w}^*)$. When $\mathbf{w} \neq \mathbf{0}$, model (4.1) signifies an informative mechanism with log-likelihood function indicated as $\ell(\boldsymbol{\theta}|\mathbf{w})$. In particular when $\mathbf{w}$ has only one non-zero entry in position i, this corresponds to a situation where we allow the i-th subject to drop out non-randomly.

The likelihood displacement is defined then as a function of $\mathbf{w}$. If $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_w$ are the ML estimators of $\boldsymbol{\theta}$ corresponding to $\mathbf{w}^*$ and $\mathbf{w}$ respectively, the likelihood displacement function is

$$g(\mathbf{w}) = 2\left[\ell(\hat{\boldsymbol{\theta}}|\mathbf{w}^*) - \ell(\hat{\boldsymbol{\theta}}_w|\mathbf{w})\right], \tag{4.2}$$

and can be used to assess the influence of small perturbations of the weights $\mathbf{w}$ on the parameters of interest.

The vector $\mathbf{w}$ and the corresponding value of $g(\mathbf{w})$ identify an *influence surface* in $\mathbb{R}^{n+1}$ as $\mathbf{w}$ varies in $\mathbb{R}^n$ referred to as $\alpha(\mathbf{w}) = (\mathbf{w}, g(\mathbf{w}))$. From (4.2) it follows that this surface has a minimum of zero at $\mathbf{w} = \mathbf{w}^*$. Interest then lies in exploring the behaviour of the surface $\alpha(\mathbf{w})$ in the vicinity of $\mathbf{w}^*$, namely how it deviates from its tangent plane around the critical point $\mathbf{w}^*$; this can be done by considering the normal curvature of the surface in $\mathbf{w}^*$ when we move away from it along any arbitrary direction $\mathbf{v} \in S^n(0,1)$, the unit hypershere.

When n=2, i.e. in terms of (4.1), there are only two subjects, the normal curvature of the surface $\alpha(\mathbf{w})$ can be represented graphically. Figure 4.1 plots a generic surface $\alpha(\mathbf{w}) = (\mathbf{w}, g(\mathbf{w})) = ((w_1, w_2), g(w_1, w_2))$ in $\mathbb{R}^3$. The plane $\mathbf{Z}$, spanned by the generic direction $\mathbf{v}$ and the norm $\mathbf{N}$ at $\mathbf{w}^* = (0,0)$, cuts the surface $\alpha(\mathbf{w})$ and identifies a normal section. The normal curvature is then defined as the projection of the curvature vector onto the norm $\mathbf{N}$. As we rotate the normal plane around the norm at $\mathbf{w}^*$, the value of the normal curvature changes smoothly

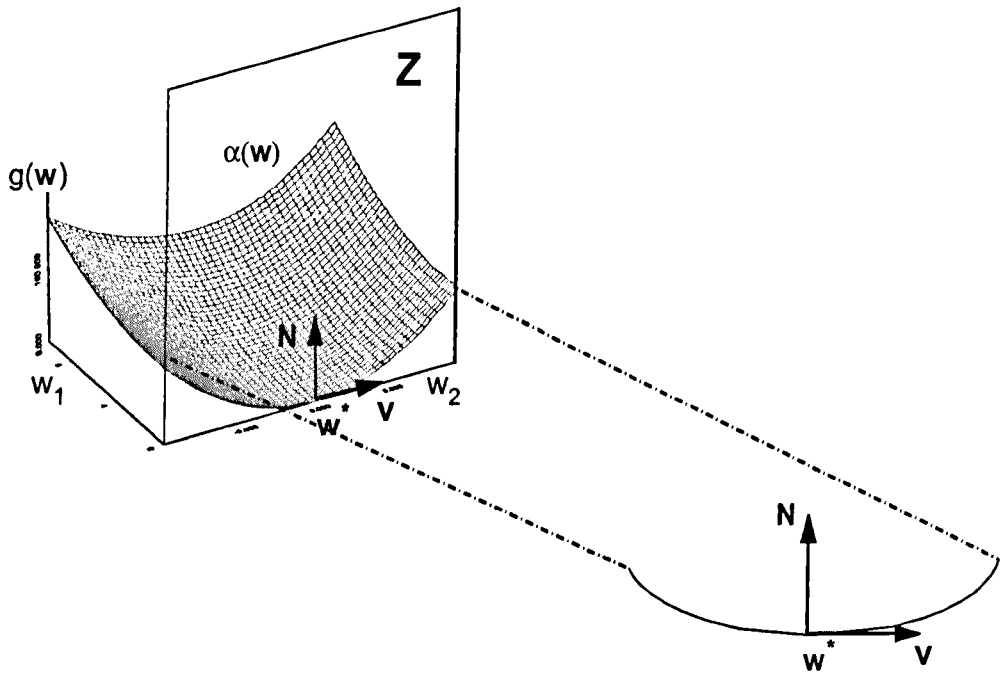

Figure 4.1: Normal section at $\mathbf{w}^*$ of the influence graph α, identified by the plane Z which is spanned by a generic direction $\mathbf{v}$ and the norm $\mathbf{N}$ at $\mathbf{w}^*$.

and reaches a maximum and a minimum.

Directions $E_1 = (1, 0)$ and $E_2 = (0, 1)$ are interesting in that they correspond to perturbing the MAR model by allowing the two subjects to drop out non-randomly. The normal curvature in this case measures the rate of increase in log-likelihood (ratio) around $\mathbf{w}^*$ when the subject-specific weights w_1 and w_2 change by infinitesimal quantities. Another interesting direction is the direction of maximum curvature. It identifies the value of $\mathbf{w}$ and therefore the combination

of subject weights that will result in maximum displacement.

In differential geometry, the normal curvature of the influence graph α in the direction $\mathbf{v}$ at the point $\mathbf{w} = \mathbf{w}^*$ is defined as the ratio of the *second* and *first fundamental forms* of $g(\mathbf{w})$(Poon and Poon, 1999), i.e. the scalar

$$C_v = \frac{\mathbf{II}(\mathbf{v}, \mathbf{v})}{\mathbf{I}(\mathbf{v}, \mathbf{v})} = \frac{\mathbf{v}^T \mathbf{II} \mathbf{v}}{\mathbf{v}^T \mathbf{I} \mathbf{v}}. \tag{4.3}$$

Here $\mathbf{I}$ and $\mathbf{II}$ are $n \times n$ matrices with elements

$$\mathbf{I}_{ij} = \delta_{ij} + \frac{\partial^2 g}{\partial w_i \partial w_j}, \tag{4.4}$$

and

$$\mathbf{II}_{ij} = \frac{1}{(1 + |\nabla(g)|^2)^{1/2}} \frac{\partial^2 g}{\partial w_i \partial w_j}, \tag{4.5}$$

where in (4.5) $\nabla(g)$ is the $n \times 1$ vector of first derivatives of g with respect to $\mathbf{w}$ and $\delta_{ij} = 1$ if $i = j$ and is 0 otherwise.

Expression (4.3) can be written as

$$C_v = \frac{\mathbf{v}^T \mathbf{H}_g \mathbf{v}}{\mathbf{v}^T (\mathbf{I}_n + \nabla(g) \nabla(g)^T) \mathbf{v} (1 + |\nabla(g)|^2)^{1/2}}_{|\mathbf{w}=\mathbf{w}^*} \tag{4.6}$$

where the generic element of the $n \times n$ matrix $\mathbf{H}_g$ is $\mathbf{H}_{g_{ij}} = \frac{\partial^2 g}{\partial w_i \partial w_j}$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Notice that when $\mathbf{w} = \mathbf{w}^*$, $\nabla(g) = 0$ and since $\mathbf{v} \in S^n(0, 1)$ implies $\mathbf{v}^T \mathbf{I}_n \mathbf{v} \equiv 1$ expression (4.6) simplifies to

$$C_v = \mathbf{v}^T \mathbf{H}_g \mathbf{v}|_{\mathbf{w}=\mathbf{w}^*}. \tag{4.7}$$

From (4.2), applying the chain rule, (4.7) can be written as

$$C_v = -2[\mathbf{v}^T \Delta'(\ddot{\ell})^{-1} \Delta \mathbf{v}] \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \mathbf{w}=\mathbf{w}^*}, \tag{4.8}$$

where $\ddot{\ell}$ and Δ are $p \times p$ and $p \times n$ matrices with generic elements $\ddot{\ell}_{ij} = \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{w})}{\partial \theta_i \partial \theta_j}$ and $\Delta_{ij} = \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{w})}{\partial \theta_i \partial w_j}$ evaluated at $\mathbf{w} = \mathbf{w}^* = \mathbf{0}$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, the estimates from a MAR analysis.

Poon and Poon (1999) suggest the use of conformal normal curvatures to assess local influence. Unlike the normal curvature which can take any value, they show that the conformal normal curvature ranges between 0 and 1 and can therefore be used as a more objective measure of influence.

The conformal normal curvature is defined as

$$B_{\mathbf{v}} = \frac{\mathbf{II}(\mathbf{v}, \mathbf{v})}{\mathbf{I}(\mathbf{v}, \mathbf{v}) \sqrt{tr\mathbf{II}^2}} \Big|_{\mathbf{w}=\mathbf{w}^*}$$

where, since $\mathbf{II}$ is a symmetric matrix, the trace in the denominator is equal to the sum of its squared eigenvalues i.e. $tr\mathbf{II}^2 = \sum_{i=1}^{n} \lambda_i^2$.

As with the normal curvature, at $\mathbf{w} = \mathbf{w}^*$ the expression of the conformal normal curvature simplifies to

$$B_{\mathbf{v}} = -\frac{\mathbf{v}^T \Delta^T (\ddot{\ell})^{-1} \Delta \mathbf{v}}{\sqrt{tr\{\mathbf{v}^T \Delta^T (\ddot{\ell})^{-1} \Delta \mathbf{v}\}^2}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \mathbf{w}=\mathbf{w}^*} \tag{4.9}$$

which involves the same quantities needed to calculate normal curvatures.

In terms of (4.9), the direction of maximum curvature is given by the eigenvector corresponding to the maximum eigenvalue of the quadratic form $\Delta^T (\ddot{\ell})^{-1} \Delta$ in the numerator.

The conformal normal curvature has the following interesting properties (Poon and Poon, 1999):

- for any direction $\mathbf{v} \in S^n(0,1)$ we have that $0 \leq B_{\mathbf{v}} \leq 1$
- indicating with $\mathbf{e}_{max}$ the eigenvector corresponding to the maximum eigenvalue (i.e. the direction of maximal displacement) with generic element $e_{max,j}$, the contribution of basic subject-specific perturbation vectors $\mathbf{E}_i$ that have the only non-zero entry equal to one in position i, can be measured considering how close they are to $\mathbf{e}_{max}$. If the contributions of all subjects to the maximal displacement are equal, then $|e_{max,j}| = \frac{1}{\sqrt{n}}$ $j = 1, \ldots, n$.
  When $n = 2$, this can be easily seen. Suppose the direction of maximum conformal normal curvature is $\mathbf{e}_{max} = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ then the two subjects are equally influential as their basic perturbation vectors $\mathbf{E}_1 = (0,1)$ and $\mathbf{E}_2 = (1,0)$ are equidistant from $\mathbf{e}_{max}$ and $|e_{max,j}| = |\frac{\sqrt{2}}{2}| = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{2}}$ $, j = 1, 2$.
- the *total contribution* of basic perturbation vectors can be measured. This measures the closeness of $\mathbf{E}_j$ to all directions identified by the normalized eigenvectors $\mathbf{e}_i$ $i = 1, \ldots, n$ of $\Delta^T(\ddot{\ell})^{-1}\Delta$ and not just $\mathbf{e}_{max}$.
  It can be shown that if the basic perturbation vectors have the same total contribution this is equal to

$$\widetilde{B} = \frac{tr(\mathbf{\Pi})}{n\sqrt{tr(\mathbf{\Pi}^2)}}. \tag{4.10}$$

  Therefore $\widetilde{B}$ can be used as an objective benchmark to judge the total influence of each subject.

4.1.1 Assessment of local influence using conformal normal curvatures in random-coefficient-based models

The quantities involved in the calculation of the conformal normal curvatures (4.9) are the elements of the Hessian matrix of the log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{w})$ of the perturbed model. Since these quantities are evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{MAR}$ and $\mathbf{w} = \mathbf{w}^* = \mathbf{0}$, they can be obtained using the parametric bootstrap approximation to Louis' formula of Subsection 3.2.1. In particular, the quantities on the right-hand side of expression (4.9) have been approximated using quasi-Monte Carlo integration considering a set of cumulative representative points from (3.7) as described in Subsection 2.3.4 with all parameters fixed at their MAR estimates. Expressions for Δ_{ij} are reported in Appendix B.

It should be pointed out that, with the local influence approach to sensitivity analysis illustrated here, each patient in turn is allowed to dropout non-randomly that is, the log-likelihood displacement is measured for directions $\mathbf{E}_i$, $1, \ldots, 250$ in turn. A global measure of influence could be obtained by averaging the curvatures corresponding to uniformly distributed directions in the unit hypersphere $S^n(0, 1)$.

Figure 4.2 shows the conformal normal curvatures $B_{\mathbf{E}_i}$ for patients in the two treatment arms obtained from expression (4.9). Different symbols refer to completers and patients lost to follow up. The dotted line corresponds to $\widetilde{B}$ of expression (4.10) which, as mentioned above, can be used to judge the total influence of each patient. Relatively few patients are found to be influential, as judged by being above the reference benchmark $\widetilde{B}$; this supports the results of Section 3.3 where estimates from a MAR model did not appear sensitive to the MAR assumption.

In both arms, completers seem to have a greater influence. Thus allowing completers (i.e. those patients with most data) to 'drop out' non-randomly results in a larger perturbation of the MAR model or, equivalently, larger displacement of

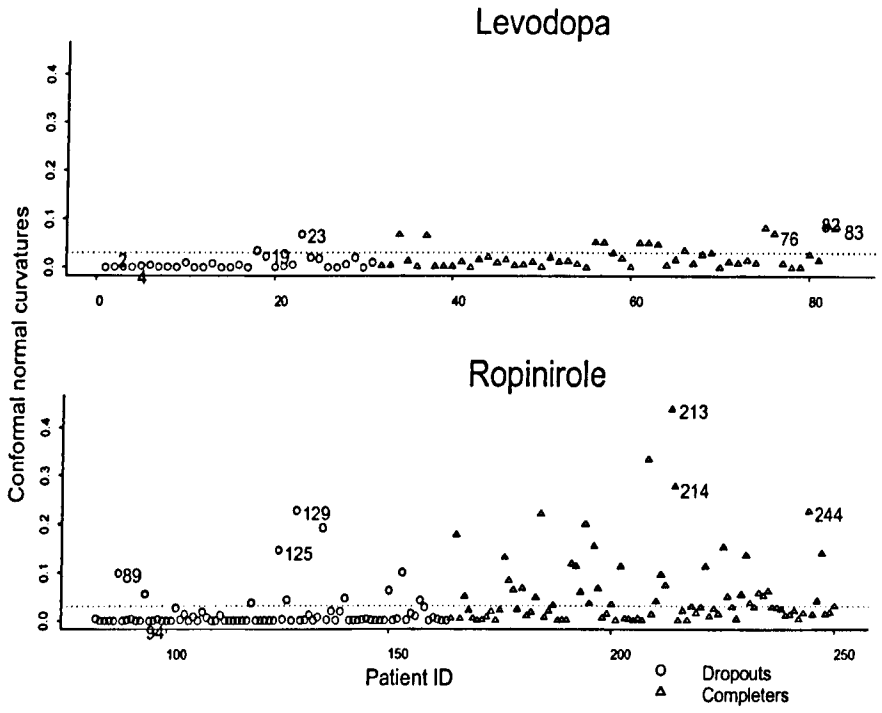

Figure 4.2: Conformal normal curvatures for patients in the PD trial. Circles and triangles correspond to dropouts and completers respectively. The dotted line represents the benchmark $\widetilde{B}$ (see text for explanation). The cases with high influence are labelled.

the likelihood function around $\mathbf{w}^*$. This is explained by completers contributing more data to the analysis than dropouts. Inspection of those patients with largest influence shows that more patients in the Ropinirole arm tend to remain in the study despite a relatively steep positive rate of increase in ADL scores, compared to completers in the Levodopa arm. This can be seen in Figure 4.3 where individual profiles corresponding to some of the completers identified in Figure 4.2 are plotted. Qualitatively this latter feature is consistent with the sign of the coefficient α_4 for the treatment-by-random slope interaction term reported in Table 3.4 though this was not statistically significant — i.e. there is less evidence in the Ropinirole arm that the dropout mechanism depends on subject-

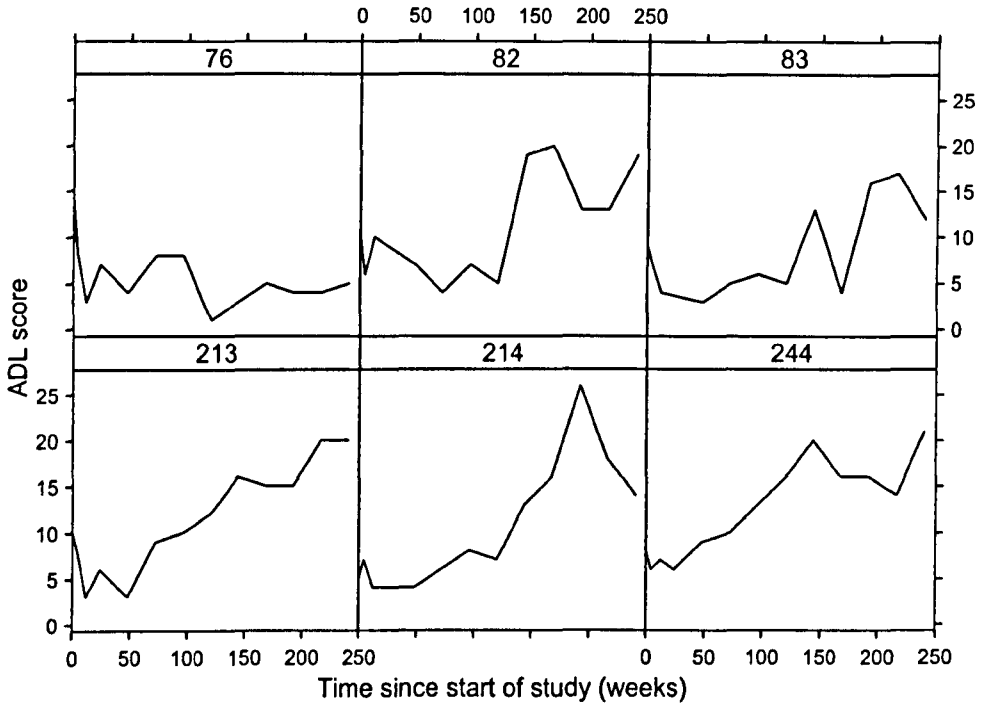

Figure 4.3: Profiles of influential patients, identified in Figure 4.2, who completed follow-up.

specific random slopes. As for those patients who do not complete the follow-up, the most influential tend to have steeper rate of change and tend to deviate more from the overall mean slope β_3 than the less influential ones (Figure 4.4). These results are consistent with those found in Section 3.3.

4.2 Sampling-based sensitivity analysis using finite-elements methods

The results from the joint modelling approach of Chapter 3 (including assessment of the statistical significance of the coefficients in the dropout model) depend on the assumption of normality of the data and random effects. This assumption is untestable for the missing data as no data were collected.

The idea behind a sampling-based sensitivity analysis is to avoid explicit estimation of the parameters in (3.5) that relate the random effects to dropout. Instead,

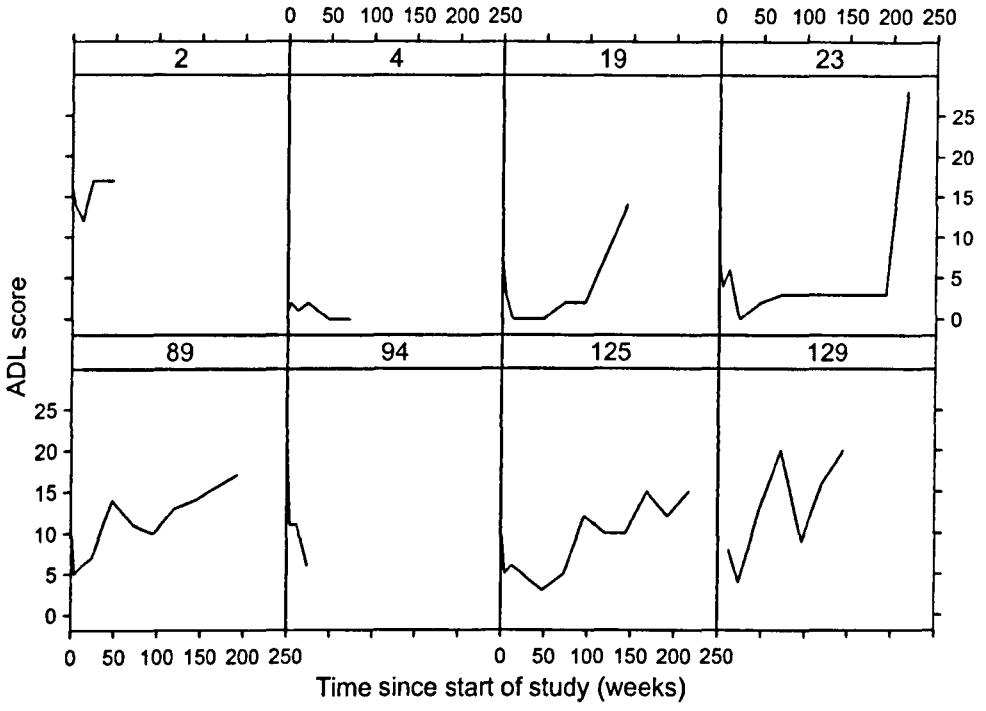

Figure 4.4: Profiles of eight patients, identified in Figure 4.2, who did not complete follow-up.

they are allowed to take values over intervals or regions defined by the analyst. The response and dropout models are then jointly fit using the methods of the previous Chapter with the sensitivity parameters in the dropout model fixed at selected values; this produces a 'response surface' for the estimate of interest, say treatment effect or the overall rate of disease progression over time.

This 'black-box' mapping from the domain of the sensitivity parameters in the dropout model to the set of possible estimates of the coefficients of interest in the response model is dictated primarily by the non-standard model fitting procedure used here. Other methods for assessing the sensitivity of the results could in fact be used if one were able to obtain a closed-form expression relating the estimator of the parameter of interest to the informative parameters in the dropout model (Saltelli et al., 2000).

We rewrite the random-coefficient-based model for the dropout mechansim of Subsection 3.1.2 as

$$h_{ij} = \Pr(t_i = \text{week}_j | t_i > \text{week}_{j-1}) =$$
$$= 1 - \exp\big[-\exp\big(\alpha_0 + \alpha_1 \text{treat}_i + \alpha_2 \text{BaseADL}_i$$
$$+ \psi \text{I}(\text{treat}_i{=}1) b_{i1} + \eta \text{I}(\text{treat}_i{=}0) b_{i1} + \gamma_l\big)\big] \quad (4.11)$$

where $\text{treat}_i = \{1(\text{Ropinirole}), 0(\text{Levodopa})\}$, $i = 1, \ldots, 250$, $j = 1, \ldots, t_i$, $t_i =$ dropout time, and γ_l are the set of contrasts for $\text{week}_l, l = 2, \ldots, t_i$. The introduction of the terms ψ and η in the expression above allows the approximate log hazard ratio of dropping out to depend on the subject-specific random slopes differently in the two treatment arms.

From the results in Table 3.4, a sensible domain for ψ and η was chosen to be the unit square.

The next step is to sample values of ψ and η at which to evaluate the response and dropout models jointly. This can be done in different ways. The simplest approach is random sampling; as no particular prior distribution is assumed for ψ and η, this reduces to sampling from a bivariate uniform distribution on the unit square.

Random sampling can be very inefficient in terms of coverage of sample space and for this reason is particularly not well suited for cases where the underlying models are expensive to evaluate as here. A better coverage can be obtained using other sampling methods such as importance sampling and Latin hypercube sampling (McKay et al., 1979).

Latin hypercube sampling in particular, ensures coverage of the range of each variable: domains are first divided into intervals of equal probability and then, for each variable, a random value is selected from each interval and paired at

random and without replacement with values obtained for the other variables. A third approach, used here, is finite–element methods coupled with Lagrangian polynomial interpolation of the resulting response surface.

4.2.1 Lagrangian polynomial interpolation

For simplicity, consider the one dimensional case first. Suppose we wish to interpolate an (unknown) function $u(\psi)$ over the interval (a, b) using a polynomial of degree n. In our setting for example, $u(\psi)$ would relate changes in estimates of a particular coefficient in the response model to changes in ψ in the dropout model if this were characterized by a single sensitivity parameter.
The domain is first divided into subdomains (finite elements) of equal width. Within each subdomain, the values of the function at $n+1$ equally spaced nodal points $\psi_0, \psi_1, \ldots, \psi_n$, are obtained and denoted by $u_1, u_2, \ldots, u_{n+1}$. The Lagrangian interpolation of the target function at $\psi \in (a, b)$ is then given by

$$u(\psi) = \phi_1^{(n)}(\psi_1)u_1 + \phi_2^{(n)}(\psi_2)u_2 + \cdots + \phi_{n+1}^{(n)}(\psi_{n+1})u_{n+1} \tag{4.12}$$

where $\phi_i^{(n)}$, $i = 1, \ldots, n+1$ are the shape (basis) Lagrangian function of degree n

$$\phi_k^{(n)}(\psi) = \prod_{m=0,\, m\neq k}^{n} \frac{\psi - \psi_m}{\psi_k - \psi_m}. \tag{4.13}$$

Thus, the value of the target function at any point in (a, b) is calculated as a weighted sum of the value of the function at the nodal points with weights given by expression (4.13).

The case where the interpolating polynomial is of degree 2 is shown in Figure 4.5. The function to be interpolated over the interval (a, b) is shown on the left. The domain is divided into 3 subdomains of equal length (this number could be incresed to get an improved approximation). Within each subdomain, the value

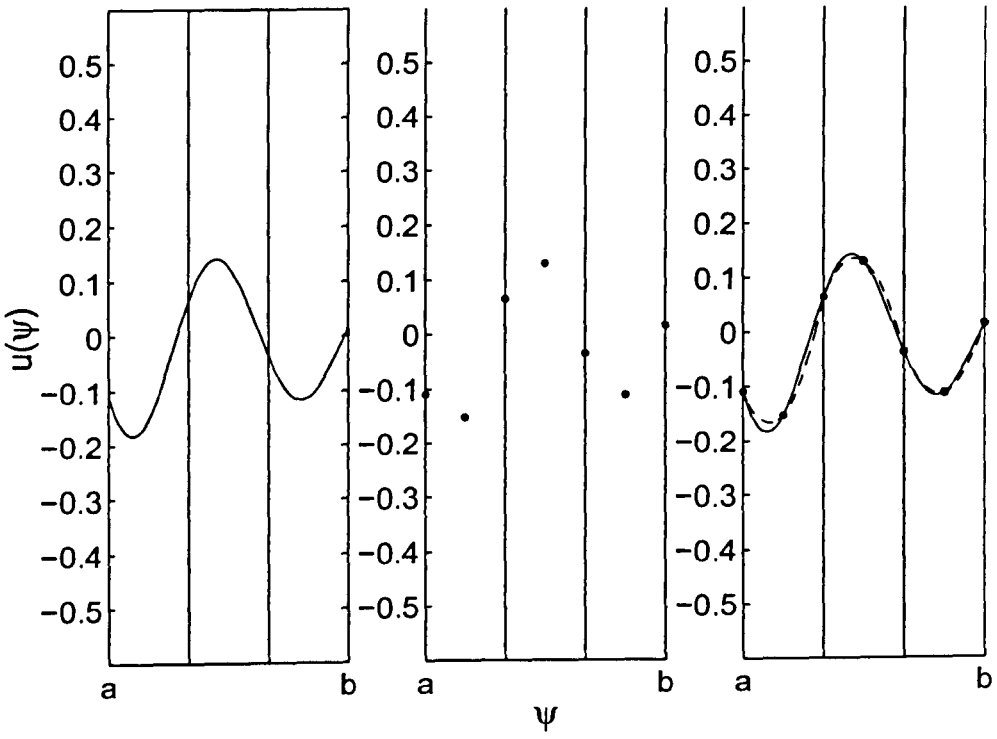

Figure 4.5: Lagrangian interpolation (dotted line in right panel) of function on the left on domain (a, b) using quadratic shape functions within each subdomain.

of the function is evaluated at $n + 1 = 3$ points (in practice only once for the subdomain in the middle). Finally, the original function and the approximation using Lagrangian interpolation (dotted line) are shown on the right.

2-Dimensional shape functions are constructed as the product of the corresponding 1-dimensional functions i.e.

$$\phi_i(\psi, \eta) = \prod_{m=0,\, m \neq k}^{n_1} \frac{\psi - \psi_m}{\psi_k - \psi_m} \prod_{m=0,\, m \neq k}^{n_2} \frac{\eta - \eta_m}{\eta_k - \eta_m}, \tag{4.14}$$

where n_1 and n_2 depend on the degree of the interpolating function for each coordinate. For example, with a cubic interpolating function in each coordinate i.e. with $n_1 = n_2 = 3$, we need to obtain the values of the target function $u(\psi, \eta)$

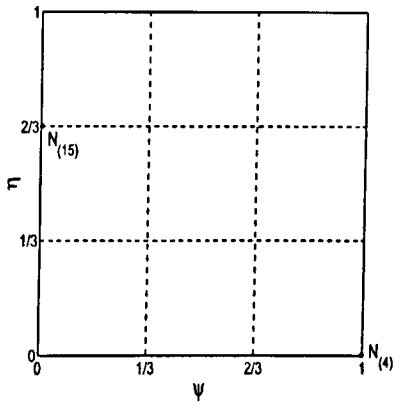

Figure 4.6: Grid of nodal points for 2-dimensional Lagrangian interpolation cubic in each coordinate.

at 16 nodal points within each subdomain (finite element) on the regularly spaced grid shown in Figure 4.6. The functions (4.14) are referred to in this case as bicubic shape functions. The interpolating surface is then constructed as

$$u(\psi, \eta) = \sum_{i=1}^{16} \phi_i(\psi, \eta) u_i, \tag{4.15}$$

which is again a weighted sum of the values taken by target function at the nodal points. Bicubic shape functions corresponding to nodes $N_{(4)} = (1, 0)$ and $N_{(15)} = (2/3, 1)$ are shown in Figure 4.7. Notice how the weight given to the value of the target function at the nodal points is higher the closer the interpolating point is to the coordinates of the node itself with weight equal to one if the point coincides with the node. In our setting, the method is used to interpolate the response surface corresponding to a chosen parameter in (3.4) as the values of ψ and η in (4.11) vary on the unit square.

In summary, the steps involved are

- select nodal pairs ψ_i, η_j on the regularly spaced grid, $i = 1, \ldots, n_1, j = 1, \ldots, n_2$;
- jointly fit the dropout and response models using the MCEM algorithm with

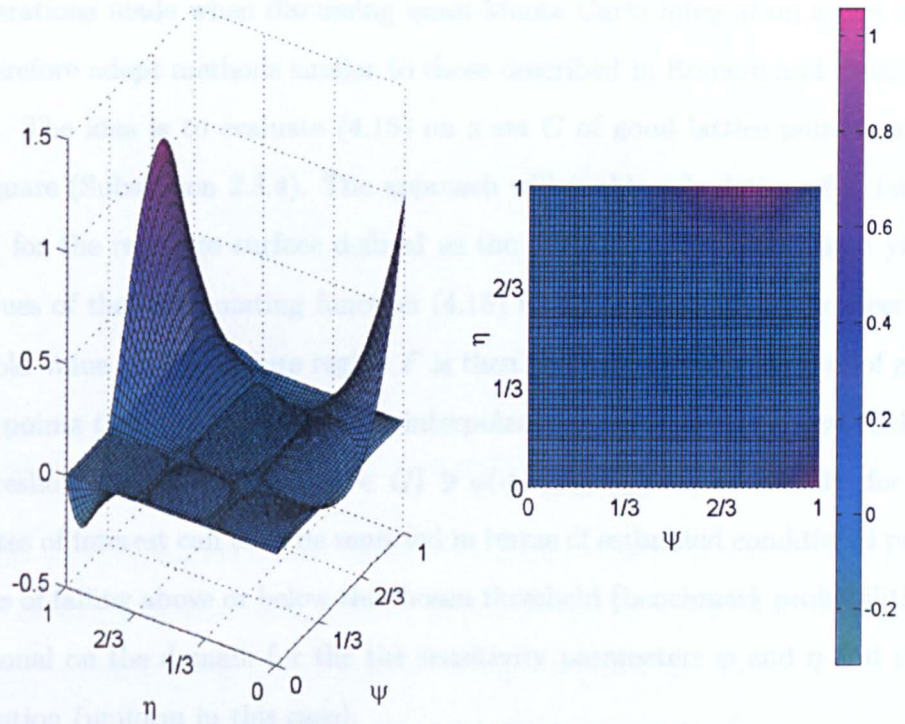

Figure 4.7: Bicubic shape function for nodes $N_{(4)} = (1, 0)$ and $N_{(15)} = (2/3, 1)$ in Figure 4.6. The weight given to the value of the target function at the nodal points is higher the closer the interpolating point is to the node itself with weight equal to one if the point coincides with the node.

ψ and η fixed at the selected values;

- obtain the value of the target function (estimate of treatment effect, overall rate of chance, etc.);
- repeat for all pairs, and
- interpolate the values thus obtained using the Lagrangian shape functions as in (4.15).

4.2.2 Estimation of benchmark probabilities

The interpolating surface (4.15) could be evaluated on pairs (ψ, η) chosen using simple, importance or Latin hypercube sampling from the unit square. However, as far as the optimal coverage of the sampling domain is concerned, the same

considerations made when discussing quasi-Monte Carlo integration apply here. We therefore adopt methods similar to those described in Romero and Bankston (1998). The idea is to evaluate (4.15) on a set G of good lattice points on the unit square (Subsection 2.3.4). The approach will enable calculation of a 'failure region' for the response surface defined as the portion of the unit square yielding values of the interpolating function (4.15) that fall above or below a certain threshold value T. The failure region F is then estimated as the fraction of good lattice points that yield values of the interpolating surface falling above (below) the threshold i.e. as $\hat{F} = \{(\psi, \eta) \in G\} \ni \phi(\psi, \eta) \geq (\leq) T$. The results for the estimates of interest can then be reported in terms of estimated conditional probabilities of falling above or below the chosen threshold (benchmark probabilities), conditional on the domain for the the sensitivity parameters ψ and η and their distribution (uniform in this case).

4.2.3 Application to the Parkinson's Disease trial

Recall that, in the application to data from the Parkinson's disease trial, from the results shown in Table 3.4, we consider the unit square as the domain for ψ and η in (4.11). This was divided into four finite subdomains. Within each subdomain biquadratic Lagrangian interpolation was used. Thus, the methods described here required fitting models (3.4) and (4.11) jointly with ψ and η fixed at the 25 values on the grid in Figure 4.8. Of course, a larger number of subdomains would improve the approximation at the expense of increased computational burden; however this is substantial for the model fitting procedure considered here.

In the Parkinson's disease trial, interest lies in particular on assessing the sensitivity of the results for β_1 (overall rate of change over time) and β_2 (average treatment difference between Ropinirole and Levodopa). Results for β_1 are shown in Figure 4.9. The darker plane represents the interpolating surface for parameter

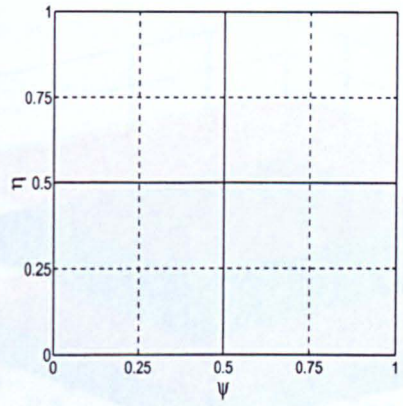

Figure 4.8: Application to the Parkinson's disease trial: grid of nodal points for 2-dimensional Lagrangian interpolation quadratic in each coordinate.

estimates and the shaded planes the interpolating surfaces for the corresponding 95% confidence intervals. As expected, higher values for ψ and η lead to higher estimated values for the overall rate of change as non-completers tend to have steeper positive slopes are are given more weight as discussed in Section 3.2.2. On the other hand, results for β_2 are not sensitive to changes in ψ and η (Figure 4.10). In both cases estimates remain statistically significant, so benchmark probabilities have not been calculated.

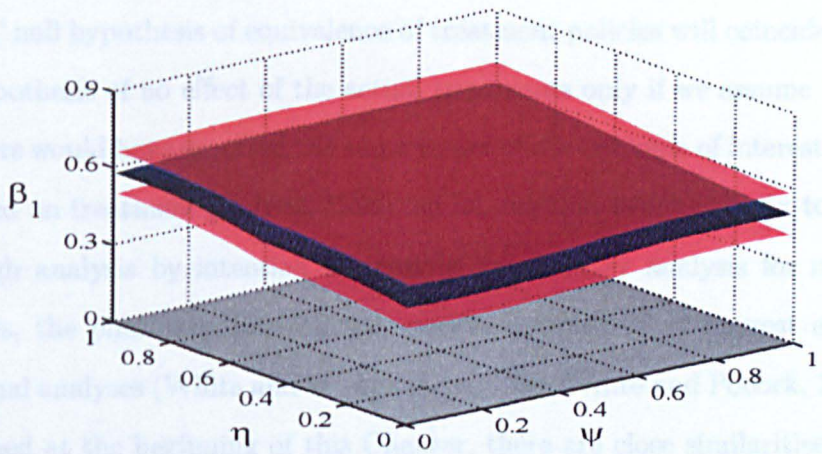

Figure 4.9: Estimated response surface for β_1 (overall rate of change over time) and corresponding 95% CIs, as ψ and η in (4.11) vary over the unit square.

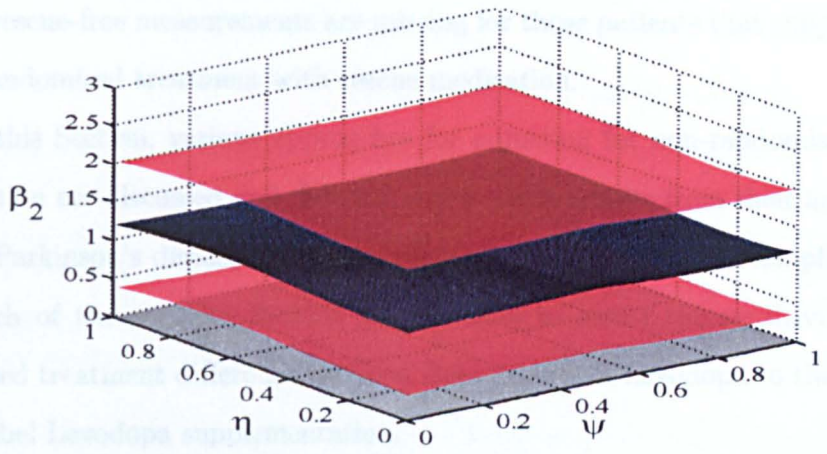

Figure 4.10: Estimated response surface for β_2 (average treatment difference between Ropinirole and Levodopa) and corresponding 95% CIs as ψ and η in (4.11) vary over the unit square.

4.3 Sensitivity analysis for the effect of Levodopa supplementation

When some patients deviate from the assigned treatment regime, a pragmatic, intention to treat (ITT) analysis of treatment policies can yield different results from an explanatory per protocol analysis (Schwartz and Lellouch, 1967). In fact, the ITT null hypothesis of equivalence of treatment policies will coincide with the null hypothesis of no effect of the actual treatments only if we assume that non-compliers would have reported the same values of the outcome of interest had they remained on treatment (Robins, 1998); an assumption which is likely to be false. Although analysis by intention to treat is the primary analysis for regulatory purposes, the pharmaceutical efficacy may sometimes be of interest as part of additional analyses (White and Goetghebeur, 1998; White and Pocock, 1996). As mentioned at the beginning of this Chapter, there are close similarities between missing data and treatment non-compliance problems: ignoring the missing data or the rescue process is likely to give biased parameter estimates of the true treatment effect. Likewise, any adjustment will depend on untestable assumptions

as the rescue-free measurements are missing for those patients that supplemented their randomised treatment with rescue medication.

In this Section, various approaches for adjusting for non-randomised rescue medication are discussed and, for each one in turn, results from their application to the Parkinson's disease trial are presented. In particular, the sampling-based approach of the previous Section will be used to assess the sensitivity of the estimated treatment difference between Ropinirole and Levodopa to the effect of open-label Levodopa supplementation.

Before proceeding any further, some additional notation is required. For subject $i = 1, \ldots, 250$ in treatment arm $Z \in \{1, 0\}$, (1=Levodopa, 0=Ropinirole), let Y_i denote the random variable corresponding to the outcome of interest at the end of follow-up period (week 244) and indicate with $R_{w_{ki}}$ dummy variables flagging whether or not patient i was taking supplementary open-label Levodopa at week $w_k \in \{12, 24, 48, 72, 96, 120, 144, 168, 192, 216, 244\}$, where $R_{w_{ki}} = 0$ if the patient is on the randomised regime in the interval $(w_k, w_{k+1}]$ and $R_{w_{ki}} = 1$ if he receives supplemented Levodopa, $k = 0, \ldots, 10$. As in Robins (1998), indicate with $\overline{\mathbf{R}}_{w_{ki}}$ the subject-specific history of treatment received by patient i up to and including week k.

The objective will then be to estimate and compare mean ADL score at week $244 = w_K$ across the two arms had all patients remained on their assigned treatment which we indicate as

$$E[Y(\overline{\mathbf{R}}_{w_K} = \mathbf{0})|Z]. \tag{4.16}$$

4.3.1 Methods based on summary statistics

The analysis of the Parkinson's disease data using summary statistics compares each subject's mean of post-randomisation ADL scores (from visit 3 onwards) acr-

Method	Unadjusted for Levodopa supplementation		Adjusted for Levodopa supplementation	
	Treatment difference		Treatment difference	
	Est. (95% CI)	P-val.	Est. (95% CI)	P-val.
ANCOVA	0.92 (0.10,1.74)	0.024	0.78[a] (0.03,1.52)	0.046
IPCW			2.2 (−8.94,12.95)[b]	

Table 4.1: Estimated treatment difference (Ropinirole-Levodopa) using the methods described in Subsection 4.3.1 and 4.3.2. [a] Censoring at rescue; [b] Bootstrap CI (percentile method). Higher ADL scores indicate worse condition.

oss the two arms using ANCOVA, with the response at randomisation used as covariate (Frison and Pocock, 1992). A subject's responses only contribute to this mean until he/she starts taking open-label Levodopa or withdraws from the trial. The treatment effect is only marginally statistically significant and correction seems to favour Ropinirole (first row of Table 4.1). However, this method is a crude adjustment prone to selection bias (White et al., 2001).

4.3.2 Inverse Probability of Censoring Weighted (IPCW) estimator

Under the rather strong assumption of rescue at random, that is, in our case, assuming that for each patient within each treatment arm the probability of receiving an additional open-label dose of Levodopa in the interval $(w_k, w_{k+1}]$ does not depend on the ADL score that he would shown under compliance at week $w_K = 244$ i.e. at the end of the follow-up, (4.16) is identifiable and is given by the mean ADL score for the compliers. This because we are assuming that the (unobservable) rescue-free ADL scores for non-compliers constitute a random sample of the (observed) ADL scores among compliers. Algebraically,

$$R_{wk} \perp Y(\overline{\mathbf{R}}_{w_K} = \mathbf{0}) | \overline{\mathbf{R}}_{w_{k-1}}, Z \qquad k = 1, \ldots, K. \tag{4.17}$$

As mentioned before, this assumption is unlikely to hold as, in general, it is reasonable to assume that non-compliers would have shown an even worse ADL score without rescue.

The idea underlying Robins' IPCW estimator approach is to consider data on auxiliary covariates that are good predictors of both the counterfactual value $Y(\overline{\mathbf{R}}_{w_K} = \mathbf{0})$ and the rescue process so that, conditioning on the past history of these covariates, (4.17) is approximately true. The value at week k and the history up to week k of these covariates for subject i are indicated with $L_{w_k i}$ and $\overline{\mathbf{L}}_{w_k i}$ respectively. The fundamental assumption of *sequential explainable non-random non-compliance* can then be written as

$$R_{wk} \perp Y(\overline{\mathbf{R}}_{w_K} = \mathbf{0}) | \overline{\mathbf{R}}_{w_{k-1}} = 0, \overline{\mathbf{L}}_{w_k}, Z. \tag{4.18}$$

Effort should be put in collecting data on covariates which are likely to make (4.18) hold.

Under (4.18), (4.16) is identifiable and can be estimated using the G-computation algorithm as described in Robins (1998). The latter can be written as the Inverse Probability of Censoring Weighted (IPCW) estimator

$$E\left[\frac{Y * I(\overline{\mathbf{R}}_{w_k} = \mathbf{0}, Z)}{\prod_{k=0}^{K} \pi(w_k)} \middle| Z\right], \tag{4.19}$$

where in the expression above $\pi(w_k) = Pr[R_{w_k} = 0 | \overline{\mathbf{R}}_{w_{k-1}} = \mathbf{0}, \overline{\mathbf{L}}_{w_k}, Z]$ is the probability that compliers up to week w_k remain on treatment in the interval $(w_k, w_{k+1}]$ given their past history of covariates $\mathbf{L}$. Therefore $\prod_{k=0}^{K} \pi(w_k)$ is the probability of remaining on treatment throughout conditional on covariate and treatment history.

Considering a logistic model for $\pi(w_k)$ with vector of parameters α, the IPCW estimator of (4.16) is

$$\frac{\sum Y_i * I(\overline{\mathbf{R}}_{w_K i} = \mathbf{0}, Z)/\prod_{k=0}^{K} \pi(w_k, \hat{\alpha})}{\sum I(\overline{\mathbf{R}}_{w_K i} = \mathbf{0}, Z)/\prod_{k=0}^{K} \pi(w_k, \hat{\alpha})}, \tag{4.20}$$

where summation is over the compliers and $\hat{\alpha}$ are ML estimates. Thus, the IPCW estimator is a weighted average of the observed scores among compliers.

Until now we have ignored the additional problem constituted by dropouts. Robins derives an extension of (4.20) which takes censoring by loss to follow-up into account. For subject i, a second dummy variable $Q_{w_k i}$ represents the censoring process and is modelled using logistic regression with parameters ϕ. Expression (4.20) is modified and becomes

$$\frac{\sum Y_i * I(\overline{\mathbf{R}}_{w_K i} = \mathbf{0}, \overline{\mathbf{Q}}_{w_K i} = \mathbf{0}, Z)/\prod_{k=0}^{K} \pi(w_k, \hat{\alpha})\lambda(w_k, \hat{\phi})}{\sum I(\overline{\mathbf{R}}_{w_K i} = \mathbf{0}, \overline{\mathbf{Q}}_{w_K i} = \mathbf{0}, Z)/\prod_{k=0}^{K} \pi(w_k, \hat{\alpha})\lambda(w_k, \hat{\phi})} \tag{4.21}$$

where $\lambda(w_k, \hat{\phi})$ is the fitted probability of not being censored at week w_k given $\overline{\mathbf{L}}_{w_k}$ among those patients still in the study at week w_{k-1}. It is hoped that, as for the rescue process, conditioning on covariates $\mathbf{L}$ dropout is at random in the sense of Rubin (1976).

For the Parkinson's disease trial, the logistic models for the probability of rescue/censoring by loss to follow-up in the interval $(w_k, w_{k+1}]$ include the following covariates:

- A_{w_k} = predicted ADL score at visit w_k from a linear model fitted on available scores up to and including week w_k;
- M_{w_k} = predicted Motor score at visit w_k from a linear model fitted on available scores up to and including week w_k;
- D_{w_k} = Diskynesia by visit w_k, and

- S_{w_k} = On supplemented Levodopa at visit w_k (in the model for the dropout process only).

The adjusted treatment difference increases to 2.2 compared to the ITT analysis and this seems plausible given that, as mentioned in the previous Section, Levodopa supplementation is more frequent in the Ropinirole arm. However, a 95% bootstrap confidence interval of $(-8.94, 12.95)$ for this difference does not support any definite conclusion as to the effect of treatment in the absence of rescue. The large confidence interval in particular reflects the fact that many patients received supplemental Levodopa. Also, it is much wider than the corresponding interval from the ITT analysis as it takes into account the extra uncertainty deriving from simultaneous adjustment for the rescue and dropout processes.

4.3.3 Structural nested distribution and mean models

A major drawback of the IPCW estimator is that, although the models for the rescue and censoring by loss to follow-up processes take into account all available data before the patient leaves the randomised treatment regime or drops out, the estimator itself is a weighted average of the final observation among completers remaining on treatment. As such, it can be very inefficient when, as in our case, the number of completers that are also compliers is small (less than 30 in both the Levodopa and Ropinirole arm out of the 83 and 167 patients randomised, respectively).

Structural Nested Distribution Models (SNDM) on the other hand, consider all available data at the end of the follow-up period. However they make further modelling assumptions.

We specify a surrogate random variable $H(\psi_z)$ for the (possibly) counterfactual outcome $Y(\overline{\mathrm{R}}_{w_K} = \mathbf{0})$ such that the former and the latter have the same distribution, $Z \in \{1, 0\}$. A simple form for H presented by Robins and used in the

application to the Parkinson's disease trial is

$$H_i(\psi) = Y_i - \sum_{k=0}^{K} \psi_z R_{w_k i}. \tag{4.22}$$

$H_i(\psi_z)$ can be interpreted as the rescue-free outcome for patient i, from which the cumulative effect of being on rescue is subtracted to obtain the observed outcome, Y_i. It can be shown that, fot the true value of ψ_z, under (4.18) we also have

$$R_{wk} \perp H | \overline{\mathbf{R}}_{w_{k-1}}, \overline{\mathbf{L}}_{w_k}, Z. \tag{4.23}$$

G-estimation of ψ_z consists then in a iterative process where we search for the value of ψ_z which makes the parameter θ relating R_{wk} to $H(\psi_z)$ in the logistic model for the rescue process equal to zero. The latter is

$$\text{logit}\,\pi(w_k, \boldsymbol{\alpha}, \theta) = \alpha_k + \alpha_{11} A_{w_{k-1}} + \alpha_{12} M_{w_{k-1}} + \alpha_{13} D_{w_{k-1}} + \theta H(\psi_z) \tag{4.24}$$

for $k = 0, \ldots, 10$. A 95% confidence interval for ψ is given by the set of values for which a 5% Wald test of θ equal zero in (4.24) does not reject (Robins, 1998). If a clinically plausible value of ψ is found and is statistically significant, corrected estimates of mean score within each arm are obtained as averages of the $H(\hat{\psi})_i$. Censoring by loss to follow-up can be accounted for as in the IPCW estimation. The estimators are then weighted averages of the $H(\hat{\psi})_i$

$$\frac{\sum H(\hat{\psi})_i * I(\overline{\mathbf{Q}}_{w_K i} = 0, Z) / \prod_{k=0}^{K} \lambda(w_k, \hat{\phi})}{\sum I(\overline{\mathbf{Q}}_{w_K i} = 0, Z) / \prod_{k=0}^{K} \lambda(w_k, \hat{\phi})} \tag{4.25}$$

where summation is over completers. Notice that, for patients remaining on randomised treatment, $H(\hat{\psi})_i \equiv Y_i$.

Figures 4.11 and 4.12 plot the results for the Ropinirole and Levodopa arm respectively. Values of a Wald test statistics for θ equal zero, $Z(\psi)$, are plotted against plausible values of ψ; these have been chosen so that the resulting values of H in (4.22) fall within the possible range for the ADL scores. A 95% CI for ψ corresponding to $|Z(\psi)| < 1.96$ includes zero suggesting that, if model (4.24) is correctly specified and (4.23) holds, there is little scope for adjusting for Levodopa supplementation. The confidence interval for ψ is wider in the Levodopa arm compared to the Ropinirole arm possibly reflecting the fact that fewer subjects received supplementation in the former. In both arms, point estimates for ψ are negative implying (from (4.22)) larger estimated rescue-free scores than the ones actually observed; a scenario which is clinically plausible.

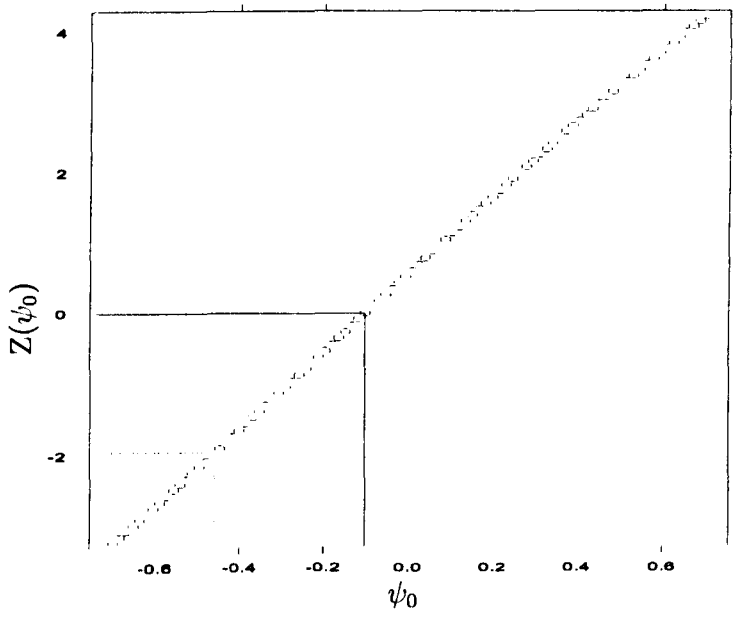

Figure 4.11: Test statistic $Z(\psi_0)$ for $H_0 : \theta = 0$ in (4.24) versus plausible values of ψ_0 in the Ropinirole arm. 95% CI for ψ_0 is shown.

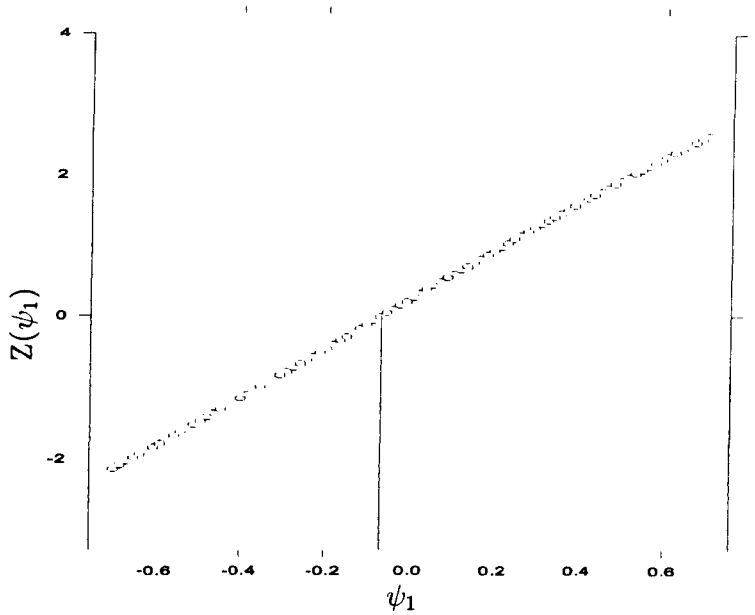

Figure 4.12: Test statistic $Z(\psi_1)$ for $H_0 : \theta = 0$ in (4.24) versus plausible values of ψ_1 in the Levodopa arm. 95% CI for ψ_1 is shown.

In Continuous-time Structural Nested Mean Models (CSNMM) it is assumed that $H(\psi_z)$ and the counterfactual outcome $Y(\overline{\mathbf{R}}_K = \mathbf{0})$ are equal in expectation rather than in distribution. Indicating with C the random variable measuring the number of days after randomisation a patient first received supplemented Levodopa, expression (4.24) is replaced by a Cox model for C

$$\lambda_C(t) = \lambda_0(t) \exp[\alpha_1 A_{t-} + \alpha_2 M_{t-} + \theta H(\psi_z)] \tag{4.26}$$

where A_{t-} and M_{t-} refer to the fitted ADL and Motor scores prior to the time t of start of rescue medication; for a complier, the latter quantities are obtained considering all available measurements prior to the final one.

Using this approach, results similar to those from a SNDM were obtained. The G-estimation algorithm described in the previous Section (where we iterate now between a continuous version of (4.22) and (4.26)) yields a point estimates

(95% percentile bootstrap confidence interval) equal to $\hat{\psi}_0 = 0\,(-0.16, 0.16)$ and $\hat{\psi}_1 = 0\,(-0.49, 0.65)$ for the Ropinirole and Levodopa arm respectively.

4.3.4 Sampling-based sensitivity analysis

The sampling-based sensitivity analysis of Subsection 4.3.2 can be used to explore the uncertainty about the true treatment difference by estimating the treatment difference in a set of imputed rescue-free data sets where imputation is based on a clinically plausible model for the effect of rescue.

As mentioned before, patients received extra open-label doses of Levodopa when their symptoms where found to be poorly controlled by the randomised treatment regime. Based on this, it was assumed that a patient who received Levodopa supplementation would have shown higher ADL scores had he remained on the randomised treatment. In what follows, both the measurement and rescue processes are considered as continuous-time processes and days since randomisation are used as the time scale.

In between two successive visits, a decision was made whether to start supplementation for a patient still on assigned treatment or increase the dose of rescue for patients already on supplemented Levodopa; in many circumstances several increases of supplemented dose would take place in the same time interval. Thus, for the generic patient i in treatment arm Z and generic interval $(t-1, t] = g$ between successive ADL score measurements, the rescue-free ADL score at time t is assumed to be given by

$$Y_{ti}^{*} = Y_{ti} + \sum_{s=1}^{n_g} \psi_Z \text{dose}_s \text{days}_s \tag{4.27}$$

where $Y_{ti}^{*} = Y_{ti}$ for patients still on randomised regime at t, n_g indicates the number of different dose prescriptions in period g and days_s is the number of days on

rescue dose s, $s = 1, \ldots, n_g$. Notice that we have indexed ψ by treatment group in order to allow for a possibly differential effect of the rescue dose in the two arms which seems sensible from a clinical point of view ($Z \in \{1, 0\}$, 1 =Levodopa, 0 =Ropinirole).

We can interpret ψ_Z as the istantaneous effect on Y_t of receiving one unit of supplemented dose of Levodopa considering the rescue process as a right-continuous step function (Robins, 1998).

Varying ψ_Z within a range of plausible positive values we obtain corresponding sets of imputed rescue-free data using (4.27), each of which is then analyzed using the linear mixed model

$$Y_{it}^{*} = \beta_0 + \beta_1 Z + \beta_2 Zt + u_{i0} + u_{i1} t + e_{it}, \tag{4.28}$$

where

$$\begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \sim N(\mathbf{0}, \Omega_u) \perp e_{it} \sim N(0, \sigma^2).$$

Here we consider the effect on the estimated treatment difference at the end of the follow-up period as this was the planned analysis. Figure 4.13 shows the estimated treatment difference at the end of the study from model (4.28) as $\psi_Z, Z = 0, 1$ in (4.27) vary over the chosen domain. Planes representing 95% CI and the null hypothesis of no treatment difference are also shown; the effect of the two treatment on ADL scores was not statistically significant under a ITT analysis ($\psi_0 = \psi_1 = 0$). Under a differential effect of rescue i.e., in Figure 4.13, considering values of ψ which are not on the diagonal $\psi_0 = \psi_1$, a statistically significant interaction term would more easily result assuming that rescue has a larger effect in the Ropinirole arm compared to the Levodopa arm; if the converse was true, treatment effect would almost always remain not significant. More interestingly,

using the methods of Subsection 4.2.2, the area corresponding to the portion of the plane representing the upper bound of the 95% confidence region above the zero plane as a fraction of the base surface was calculated and found it to be $p_b = 0.42$. We could then interpret p_b as the conditional benchmark probability for the estimated treatment difference at the endpoint to remain statistically non-significant as we vary ψ within a plausible region, conditional on the chosen model for the rescue effect (4.27). A bootstrap confidence interval for p_b using the percentile method is [0.23,0.85]. Considering the triangles above and below the main diagonal and their interpretation in terms of differential effect of rescue in the two arms we have $p_{b0} = 0.02[0.00, 0.35]$ and $p_{b1} = 0.39[0.23, 0.50]$ which, of course, confirm earlier considerations.

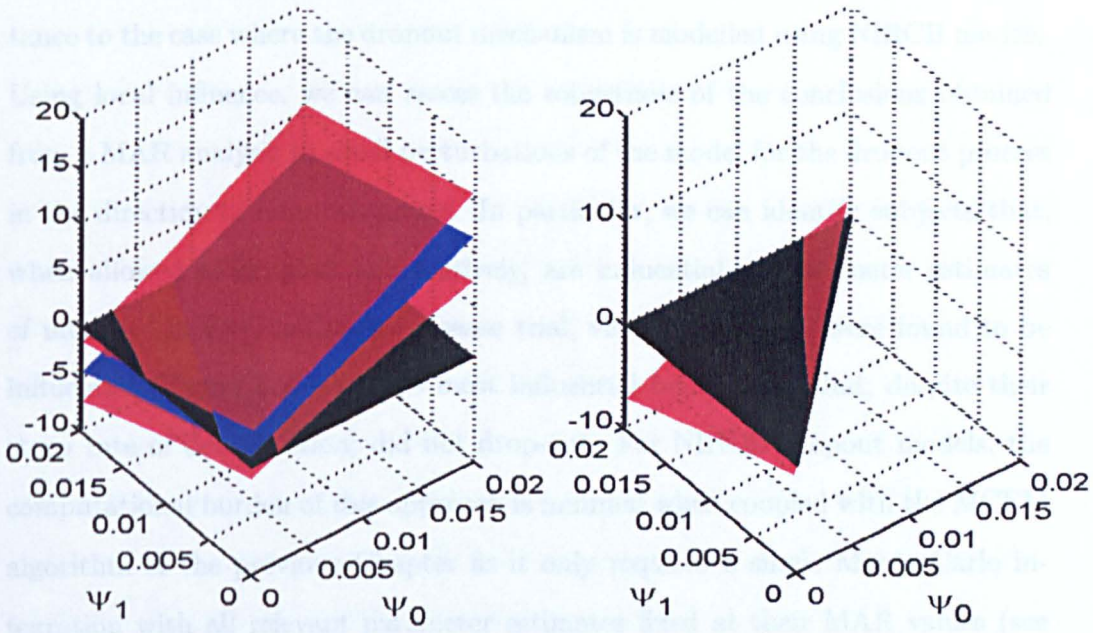

Figure 4.13: Estimated treatment difference at week 244 from model (4.28) when rescue-free observations are imputed using (4.27).

4.4 Discussion

In the presence of patient attrition, results obtained from modelling the dropout and response jointly using, for example, the methods of the previous Chapter, should be subject to careful scrutiny. Conclusions can be sensitive to the distributional and modelling assumptions made which are untestable for the missing data. Various methods can be used to assess the robustness of the conclusions. The simplest consider different models for the dropout mechanism or different distributions for the unseen data or a combination of both; results that do not vary substantially under the different scenarios can then be trusted more confidently. A more systematic sensitivity analysis was presented in the first part of this Chapter, when we adapted the local influence approach based on Cook's distance to the case where the dropout mechanism is modelled using NIRCB models. Using local influence, we can assess the robustness of the conclusions obtained from a MAR analysis to small perturbations of the model for the dropout process in the direction of nonrandomness. In particular, we can identify subjects that, when allowed to dropout informatively, are influential on parameter estimates of interest. In the Parkinson's disease trial, very few subjects were found to be influential (Figure 4.2) with the most influential being those that, despite their steep rate of deterioration, did not drop-out. For NIRCB dropout models, the computational burden of this approach is minimal when coupled with the MCEM algorithm of the previous Chapter as it only requires a single Monte Carlo integration with all relevant parameter estimates fixed at their MAR values (see Appendix B). A drawback of the local influence method is that, by definition, it does not yield a global measure of sensitivity. To this end, as mentioned in Subsection 4.1.1, a possible way to proceed is to calculate the average likelihood

displacement along directions uniformly distributed on the unit hypershere and not just those corresponding to the single subjects. Also, the results obtained do not express the sensitivity of parameter estimates of interest, which is often what may be required, for example, for regulatory purposes.

In the second part of this Chapter, we presented a sampling-based sensitivity analysis for random-coefficient-based dropout models. The method provides a measure of sensitivity for parameters of interest and is, arguably, conceptually simpler than the method of local influence. By varying the sensitivity parameters that determine informativeness in the model for the dropout mechanism over a sensible range, we obtain a 'response surface' for estimates of interest using finite-elements methods. Often, results can be represented graphically, as show in Subsection 4.2.3 for the Parkinson's disease data, and are therefore easy to convey to those without statistical training.

The method was also used in Section 4.3 to quantify the sensitivity of the estimated treatment difference between Ropinirole and Levodopa at the end of the follow-up to non-randomised Levodopa supplementation. The results obtained depend heavily on the models chosen for the dropout and rescue mechanisms the appropriateness of which are again untestable. However, the same holds for alternative approaches such as Robin's IPCW estimators discussed in Subsection 4.3.2. In general, the results yielded by any of these methods will depend on subjective hypotheses about the dropout or rescue processes. Their validity therefore is enhanced by a sensible choice of models for the dropout and rescue mechanisms which should be based on all available information.

In the next Chapter, we consider a different approach to quantify the uncertainty about estimates of coefficients of interest. The idea will be to replace traditional point estimates with intervals or regions of estimates which are compatible with the observed data while making no assumptions about the mechanism driv-

ing the missing data process (Verzilli and Carpenter, 2002b; Molenberghs et al., 2001; Vansteelandt and Goetghebeur, 2001).

Chapter 5

Bounding parameter estimates from incomplete data

In this Chapter we present methods for calculating intervals of ignorance and uncertainty (Molenberghs et al., 2001; Vansteelandt and Goetghebeur, 2001) for parameter estimates from incomplete longitudinal ordinal and continuous data (Verzilli and Carpenter, 2002b). As mentioned in the introduction, the general idea behind this approach is to replace traditional point estimates and confidence intervals for parameters of interest with intervals of ignorance and uncertainty, respectively. Intervals of ignorance account for the lack of knowledge about point estimates caused by the incompleteness of the data collected. Intervals of uncertainty extend familiar statistical imprecision due to finite sampling to the intervals of ignorance i.e. consider their sampling distribution. Intervals of ignorance and uncertainty become regions when considering two or more parameters simultaneously. Our proposed algorithms for constructing such intervals will be presented together with results from simulation studies. Finally, the methods will be applied to the dental pain and Parkinson's disease trials described in Subsections 2.1.2 and 2.1.1.

5.1 Bounds on parameter estimates with incomplete discrete data

In Chapter 4 we argued that attempts to correct for possible bias in point estimates caused by data incompleteness, which rely on modelling the response and missing data process jointly, will inevitably yield different results depending on the model used for the missing data process and the distributional assumptions made about missing data. This suggests the appropriate way to proceed is via sensitivity analysis, varying the former, the latter or both (Kenward, 1998) or using other methods like those described in Sections 4.1 and 4.2 for the random-coefficient-based dropout model of Chapter 3.

Molenberghs et al. (2001) and Kenward et al. (2001) propose a more systematic approach to sensitivity analysis for incomplete categorical data. As discussed also in Molenberghs et al. (1999), the main point is that different nonignorable models for the missing data mechanism could fit the observed part of the data equally well and still give very different predictions of the unobserved data. Thus any conclusions on their validity cannot be based on the observed data alone: subjective assumptions about the plausibility of the different models used have to be made.

Using data from the Slovenian plebiscite survey described in Rubin et al. (1995), Molenberghs et al. (2001) show how the range of possible estimates for a parameter of interest obtained by fitting different *ad hoc* models for the missing data mechanism can be recovered by maximizing an overspecified (with respect to the observed part of the data) likelihood, using what they call a sensitivity parameter approach. Overparametrization is dealt with by fixing certain sensitivity parameters conditional on which all remaining parameters are identifiable; by changing the values that these sensitivity parameters can take, different estimates for the coefficients of interest are obtained. The range of these estimates

identify the intervals or regions of ignorance. The union of the corresponding set of $100(1-\alpha)\%$ confidence intervals yields the intervals of uncertainty although the coverage properties of the intervals thus obtained are still being investigated (Vansteelandt et al., 2002).

The idea of replacing traditional point estimates with intervals is particularly suited to situations where the outcome variable is categorical. For example, one could obtain optimistic-pessimistic bounds for parameter estimates by enumerating all possible outcomes for the missing data and examining the corresponding set of all possible point estimates. In passing note that, in general, the intervals of ignorance given by the approach of Molenberghs et al. (2001) will not correspond to the optimistic-pessimistic bounds; in what follows, we will use the term interval of ignorance to refer to optimistic-passimistic intervals for parameter estimates i.e. the interval identified by the minimum and maximum estimates that are compatible with the observed data (as in Vansteelandt and Goetghebeur (2001), see below).

Horowitz and Manski (2000) obtain sharp (i.e., the tightest possible) optimistic-pessimistic bounds on parameters (for example, the probability of treatment success for individuals assigned at random to that treatment) in a nonparametric analysis of randomised experiments. Closed-form unbiased sample estimates of these bounds are presented and related inferential issues discussed. In particular, they propose constructing confidence intervals for the upper and lower bounds by using the bootstrap method. Bootstrap samples are obtained by sampling with replacement from the observed part of the data; this yields a corresponding number of bootstrap estimates of the upper and lower bounds from which their distribution conditional on the data can be estimated. These confidence intervals are conceptually equivalent to the intervals of uncertainty in Molenberghs et al. (2001).

Bounds on treatment effects in experimental studies with noncompliance are also obtained by Balke and Pearl (1997). They too use a nonparametric approach coupled with linear programming techniques. The bounds they obtain are sharp as in Horowitz and Manski (2000) but no measure of uncertainty is given in this case. Previous work in this area also include methods discussed in Manski (1989).

We mentioned above that optimistic-pessimistic bound for parameters could be obtained by brute enumeration. However, enumeration of all possible data completions becomes unfeasible even with a moderate number of missing data. Vansteelandt and Goetghebeur (2001) propose what they call the Imputing towards Directional Extremes (IDE) algorithm to obtain optimistic-pessimistic bounds on parameter estimates in generalized linear models which they again refer to as intervals or regions of ignorance. Intervals of uncertainty are obtained using the same approach of Molenberghs et al. (2001). This is to identify the lower and upper limits of uncertainty as the lower and upper $100(1-\alpha)\%$ limits of the lowest and highest parameter estimates in the interval of ignorance.

In this Chapter we extend the latter approach to marginal models. We describe how standard procedures used to fit Generalized Estimating Equations (GEE) models can be adapted to rapidly obtain intervals of ignorance and uncertainty for parameters of interest. In particular we use the adaptation of the GEE approach to ordinal repeated measurements described in Kenward et al. (1994) (see also Miller et al. (1993)). The proposed approach will be applied to data from the dental pain trial of Subsection 2.1.2. We saw earlier that, in this trial, a large percentage of patients (especially in the placebo and low dose groups) dropped out early in the follow-up. Interim missingness, however, was limited.

Using a logistic model for the probability of dropping out, the last reported pain relief was found to be strongly related to the probability of completing the study ($p < 0.001$) with patients not experiencing any improvement most likely to

terminate the study early. Thus, the mechanism driving drop-out is likely to be nonignorable.

Our aim is to quantify the uncertainty about the true treatment effects at the different dose levels by ranges of possible estimates corresponding to possible outcomes for the missing data while making no assumptions about the distribution of the missing data.

Later, we will also conduct a sensitivity analysis looking at how these bounds vary under different scenarios for the missing data i.e. as missing data are allowed to vary within prespecified ranges. We start by describing the model for the response variable and the algorithm used to obtain the bounds.

5.1.1 *A modified Fisher scoring algorithm*

In what follows, we refer to the extension of GEE to the analysis of longitudinal ordinal data described in Kenward et al. (1994). As noted in Subsection 2.2.2, in the presence of incomplete data, this approach requires the missing data mechanism to be completely at random for results to be unbiased. On the other hand, if there are no missing data, parameter estimates will be similar to those from marginal maximum likelihood methods. Since, in essence, our method considers ranges of possible outcomes in place of missing data, it can be seen as an application of the GEE approach to a series of pseudo-complete data sets.

For subject i let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$ denote the response vector of measurements taken at T distinct visits where Y_{it} takes values in $\{1, \ldots K\}$ and $t = 1, \ldots, T$. Indicating with $\mathbf{x}_{it}$ a $(p \times 1)$ vector of possibly time-specific covariates for subject i, a marginal proportional odds model for Y_{it} can be written as

$$\text{logit}\{\Pr(Y_{it} \leq k)\} = \alpha_k + \mathbf{x}'_{it}\boldsymbol{\beta}, \quad k = 1, \ldots, K-1. \tag{5.1}$$

Consider now a set $K-1$ binary variables Z_{itk} such that $Z_{itk} = 1$ if $Y_{it} \leq k$ and

$Z_{itk} = 0$ if $Y_{it} > k$ for $k = 1, \ldots, K-1$. Then (5.1) is equivalent to performing a logistic regression for each of the $K-1$ binary variables Z_k that is

$$\text{logit}\{\Pr(Z_{itk} = 1)\} = \alpha_k + \mathbf{x}'_{it}\boldsymbol{\beta}, \quad k = 1, \ldots, K-1. \tag{5.2}$$

This formulation results in an expanded data set with $K-1$ binary responses at each time point to which the GEE modelling approach can be extended as follows.

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i = g(\boldsymbol{\eta}_i) = g(\mathbf{x}'_i\boldsymbol{\theta})$ where $\mathbf{x}_i$ is now a column vector of dimension $(K-1+p) = q$ and includes contrasts for the cutpoint parameters α in (5.2); for n subjects the system of estimating equations can be written as

$$\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i \mathbf{W}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\mu}_i) = 0 \tag{5.3}$$

with

$$\mathbf{D}_i = \text{diag}\left\{\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right\} \text{ and } \mathbf{W}_i = \mathbf{V}_i^{1/2} \mathbf{R}_i \mathbf{V}_i^{1/2}$$

where $\mathbf{V}_i = \text{diag}\{\text{var}(z_{itk})\}$ is of dimension $T(K-1) \times T(K-1)$ as is $\mathbf{R}_i$, the working correlation matrix. Kenward *et al.*'s definition of the Z_{itk} results in a particular form for $\mathbf{R}_i$. For example, the assumption of independence between observations on the same person corresponds to a block diagonal matrix $\mathbf{R}_i$ with time-specific blocks $\mathbf{R}$ of dimension $(K-1) \times (K-1)$ expressing the correlation between the binary variables Z_k which can be written as functions of the cutpoint parameters $\boldsymbol{\alpha}$ (details in Kenward et al. (1994)). Notice that the blocks that make up $\mathbf{R}$ do not vary with time or between subjects as a result of the proportionality assumption in (5.1). On convergence, standard errors of parameter estimates can be calculated using a robust (sandwich) estimator in which the off diagonal, between repeated measurement correlations in $\mathbf{R}_i$ are estimated using Pearson's

residuals.

In general (5.3) has to be solved iteratively using, for example, a Fisher scoring algorithm (Liang and Zeger, 1986; Ziegler et al., 1998). Given current estimates of $\boldsymbol{\theta}$ at iteration h, new estimates at $h+1$ are obtained as

$$\hat{\boldsymbol{\theta}}^{(h+1)} = \hat{\boldsymbol{\theta}}^{(h)} - \left[\sum_{i=1}^{n} \mathbf{X}_i^T \widehat{\mathbf{D}}_i^{(h)} \widehat{\mathbf{W}}_i^{(h)-1} \widehat{\mathbf{D}}_i^{(h)} \mathbf{X}_i\right]^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \widehat{\mathbf{D}}_i^{(h)} \widehat{\mathbf{W}}_i^{(h)-1} (\mathbf{Z}_i - \hat{\boldsymbol{\mu}}_i^{(h)}). \tag{5.4}$$

With missing data, estimation of intervals of ignorance and uncertainty for all relevant parameters by enumerating all possible pseudo-complete datasets involves fitting this model an unfeasibly high number of times corresponding to all possible combinations of the Z_{itk} for the missing measurements. Indeed, this will give all possible estimates when in fact interest will mainly be in the minimum or maximum value of, for instance, the treatment effect corresponding to a particular dose level.

To this end, a modified version of the iterative algorithm (5.4), which exploits its linearity in the missing data given current estimates of $\mathbf{D}$ and $\mathbf{W}$, can be used to find the minimum and maximum values of parameter estimates over all possible sample completions. Notice that our algorithm adapts that of §3.2 in Vansteelandt and Goetghebeur (2001) to the case where the model for the response variable is given by expression (5.2).

Suppose, without loss of generality, that we are interested in finding intervals of ignorance and uncertainty for the j-th entry of the vector $\boldsymbol{\theta}$, $\theta_j = \mathbf{e}_j'\boldsymbol{\theta}$ where $\mathbf{e}_j$ is a q vector of all zeros except for a 1 at the j-th entry. Our proposed algorithm is as follows (Verzilli and Carpenter, 2002b):

1. obtain initial estimate of $\boldsymbol{\theta}$ by fitting (5.2) to all available cases;
2. at each step of the scoring algorithm, for each subject i with data missing at visit t, for the minimum (maximum) of the interval of ignorance choose

the $(K-1) \times 1$ vector $\mathbf{Z}_{it}$ such that

$$\mathbf{e}_j' \left[\sum_{i=1}^{n} \mathbf{X}_i^T \widehat{\mathbf{D}}_i \widehat{\mathbf{W}}_i^{-1} \widehat{\mathbf{D}}_i \mathbf{X}_i \right]^{-1} \mathbf{X}_i^T \widehat{\mathbf{D}}_i \widehat{\mathbf{W}}_i^{-1} (\mathbf{u}_t \otimes \mathbf{Z}_{it})$$

in (5.4) is maximum (minimum) where in the expression above $\mathbf{u}_t$ is a $T \times 1$ vector of all zeros apart from entry t which is equal to one, $t \in \{1, \ldots, T\}$ and $\otimes$ denotes the Kroneker product;

3. obtain new estimate of $\boldsymbol{\theta}$ and thus of $\mathbf{D}_i$ and $\mathbf{W}_i$ corresponding to the minimum (maximum) in step 2, and
4. iterate between step 2 and 3 until convergence.

In Appendix C we show that the modified algorithm with this nested maximization (minimization) converges to the maximum (minimum) estimate of θ_j over all possible sample completions, for all $j = 1, \ldots, q$. Upon convergence, the estimate of the interval of ignorance is obtained as the interval spanning the minimum to the maximum. The corresponding interval of uncertainty can be approximated using robust standard errors as $[\text{minimum} \pm 1.96\text{SE}_{min}] \bigcup [\text{maximum} \pm 1.96\text{SE}_{max}]$ (Vansteelandt and Goetghebeur, 2001).

5.1.2 *A simulation study*

We conducted a simulation study to confirm the correctness of the results given by the proposed algorithm against exact results obtained by enumerating all possible estimates corresponding to possible outcomes for the missing data. We considered a categorical response variable with $K = 3$ categories measured at five time points on 60 subjects in two treatment groups using the proportional odds

Scenarios	Interval of ignorance	Interval of uncertainty	CPU times (sec) Enum.	MFS†
Extreme	(−0.730,−0.315)	[−1.295,0.214]	3138.070	4.300
A	(−0.697,−0.378)	[−1.261,0.146]	182.690	4.900
B	(−0.536,−0.430)	[−1.055,0.098]	13.350	4.240
C	(−0.708,−0.390)	[−1.283,0.136]	347.590	4.460

† Modified Fisher scoring algorithm of Subsection 5.1.1

Table 5.1: Intervals of ignorance and uncertainty for β_1 in (5.5) using the modified Fisher scoring algorithm of Subsection 5.1.1; these were equal to those obtained by enumeration. Extreme scenario: no constraint on the values missing observations can take; Scenario A: intermittent missing observations (or first missing observation for dropouts) cannot vary by more than one score from last observed measurement; Scenario B: missing observations same or lower than last observed value; Scenario C: missing observations same or higher than last observed value.

model similar to (5.1),

$$\text{logit}\{\Pr(Y_{it} \leq k)\} = \alpha_k + \beta_1 \text{treat}_i + \beta_2 \text{visit}_t + \beta_3 \text{x}_i, \quad k = 1, 2. \tag{5.5}$$

Here $\alpha_1 = 2, \alpha_2 = 2, \beta_1 = -1, \beta_2 = -2.5, \beta_3 = 0.2$, $\text{treat}_i \in \{0, 1\}$, $t = 1, \ldots, 5$ and $\text{x}_i \sim N(20, 7)$ a generic subject-specific covariate. Notice that with m missing observations there will be K^m possible complete data sets corresponding to all possible combinations of values for the missing measurements; for simplicity we consider $m = 7$ missing data randomly chosen.

Table 5.1 reports the results for β_1. In particular, results corresponding to four possible scenarios for the missing data are shown. In many practical situations, it may be possible to restrict the range of values that missing outcomes are allowed to take based, for instance, on clinically plausible assumptions. Our extreme scenario puts no restriction on the values missing observations can take;

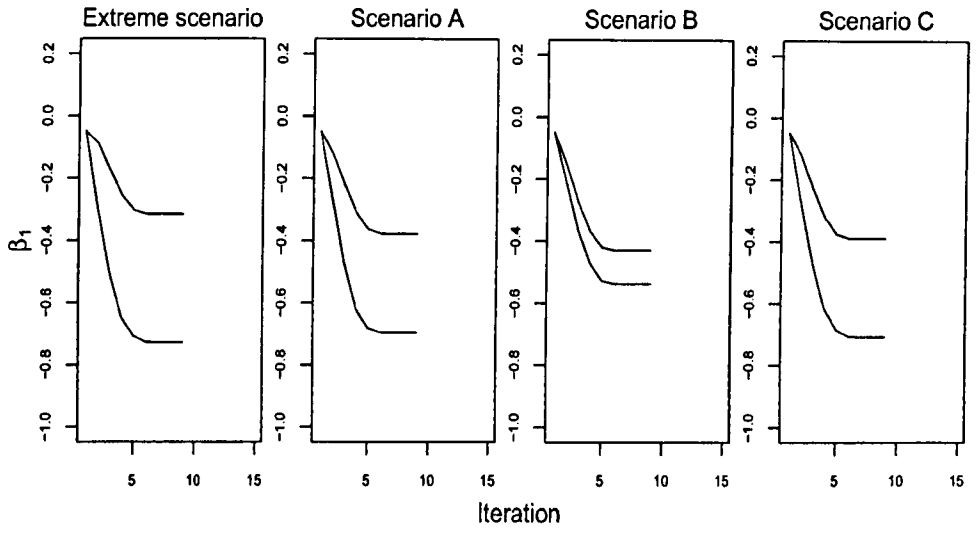

Figure 5.1: Sequences of parameter estimates from the iterative procedure described in Subsection 5.1.1.

under this scenario, we ensure coverage of all possible completed data sets. Thus pessimistic-optimistic bounds for the parameter of interest β_1 are obtained. In scenario A, intermittent missing observations (or the first missing value in case of dropouts) cannot vary by more than one category from the last observed measurement. Scenarios B and C constrain missing observations to be the same or lower or the same or higher than the last observed value respectively. The same four scenarios will be considered in the application to the dental pain data in the next Section. In all cases, the proposed algorithm yields the same results as those given by the 'exact' enumeration method, the latter in the extreme scenario scanning over $3^7 = 2187$ possible data completions. Predictably, the largest intervals of ignorance and uncertainty are obtained under the extreme scenario; scenario B results in the narrowest intervals although this need not always be the case.

Figure 5.1 plots sequences of parameter estimates from our iterative procedure. In most cases, convergence takes place in less than 10 iterations with CPU times that are both significantly lower than the corresponding times for the

enumeration method and unaffected by the particular scenario considered. All analyses have been conducted using R version 1.5.1 Ihaka and Gentleman (1996) on a Sun Ultrasparc II workstation.

5.2 Application to the dental pain trial

In the application to the dental pain data, the important covariates are patient's treatment allocation (differences between the five increasing dose levels of the experimental drug and the placebo), height, weight, time and squared time since randomisation (up to and including 8 hours):

$$\text{logit}\{\Pr(Z_{itk}=1)\} = \alpha_k + \beta_{\text{treat}_i} + \beta_6 \text{ht}_i + \beta_7 \text{wt}_i + \beta_8 \text{hour}_t + \beta_9 \text{hour}_t^2. \quad (5.6)$$

Here $i = 1, \ldots, 313$ patients, $k = 1, \ldots, K-1 = 4$, $(K = 5)$, $\text{treat}_i \in \{1, \ldots, 5\}$ and $\text{hour}_t \in \{0, 0.25, 0.5, \ldots, 8\}$. We tested for treatment-by-time interaction which was found to be not statistically significant (p=0.23). For the parameters of interest β_{treat_i}, point estimates and 95% confidence intervals obtained from an analysis of available cases are shown in the second column of Table 5.2. Compared to patients receiving placebo, patients randomised to the experimental drug show a statistically significant reduction in perceived pain at all dose levels except the lowest. As discussed earlier however, these results are likely to be biased as the dropout mechanism is likely to be nonignorable thus the estimating equations have non-zero expectation.

Intervals of ignorance and uncertainty obtained using the methods of Subsection 5.1.1 are shown in the last two columns under the four different scenarios described in Subsection 5.1.2. The unconstrained, extreme scenario leads to the widest intervals of ignorance and uncertainty which do not support the efficacy of the experimental drug at any dose level. This scenario however, appears to be

Scenarios	Est. (95%CI)	Intervals of ignorance	Intervals of uncertainty
Available cases			
Test dose 1	−0.44 (−0.97,0.09)		
Test dose 2	−0.89 (−1.41,−0.36)		
Test dose 3	−0.94 (−1.49,−0.38)		
Test dose 4	−0.99 (−1.54,−0.44)		
Test dose 5	−0.98 (−1.52,−0.45)		
Extreme			
Test dose 1		(−4.01,2.75)	[−4.56,3.17]
Test dose 2		(−4.12,1.68)	[−4.67,2.04]
Test dose 3		(−4.06,1.35)	[−4.64,1.74]
Test dose 4		(−4.22,1.52)	[−4.79,1.88]
Test dose 5		(−4.09,1.24)	[−4.64,1.61]
Scenario A			
Test dose 1		(−3.74,2.39)	[−4.27,2.80]
Test dose 2		(−3.92,1.29)	[−4.47,1.64]
Test dose 3		(−3.88,0.96)	[−4.44,1.33]
Test dose 4		(−4.02,1.15)	[−4.58,1.49]
Test dose 5		(−3.92,0.85)	[−4.46,1.20]
Scenario B			
Test dose 1		(−1.37,−0.36)	[−1.92,0.24]
Test dose 2		(−2.16,−1.09)	[−2.70,−0.52]
Test dose 3		(−2.29,−1.36)	[−2.85,−0.76]
Test dose 4		(−2.19,−1.22)	[−2.75,−0.62]
Test dose 5		(−2.29,−1.44)	[−2.87,−0.84]
Scenario C			
Test dose 1		(−3.40,2.49)	[−3.96,2.87]
Test dose 2		(−3.50,1.46)	[−4.06,1.82]
Test dose 3		(−3.45,1.27)	[−4.03,1.65]
Test dose 4		(−3.60,1.39)	[−4.18,1.76]
Test dose 5		(−3.48,1.18)	[−4.04,1.55]

Table 5.2: Point estimates and intervals of ignorance and uncertainty for treatment effects (Placebo-Active groups) estimated with data from the first 12 visits (up to 8 hours since randomization) under four different scenarios for the missing observations. Extreme scenario: no constraint on the values missing observations can take; Scenario A: intermittent missing observations (or first missing observation for dropouts) cannot vary by more than one unit from last observed measurement; Scenario B: missing observations same or lower than last observed value (worsening pain); Scenario C: missing observations same or higher than last observed value (improvement in pain). Test doses 1 to 5 correspond to parameters β_1 to β_5 in (5.6).

rather unrealistic in our case as it implies that patients who dropped out, presumably because they were not experiencing sufficient pain relief, could have instead reported complete relief had they remained in the study; this is even more unrealistic considering that the study is a single-dose trial.

Under scenario A where intermittent missing observations (or first missing observation in case of dropouts) cannot vary by more than one category from the last observed measurement, narrower intervals of ignorance and uncertainty are obtained. They do however still include zero at all dose levels.

Scenario B which implies that patients who missed a visit or dropped out would have scored the same value as their last observed measurement or worse is arguably the most plausible for this study. Under this assumption, there is evidence for the efficacy of the experimental drug compared to placebo at all dose levels except the lowest. Interestingly, in this case the intervals of ignorance show an increased treatment effect for patients receiving dose level 2 or above compared to the point estimates obtained from an analysis of the observed data; this is because there are many more patients in the placebo group with 1 as their last observed measurement (which will then be carried forward up to visit 12) compared to patients in dose groups 2 to 5. Finally, were we to assume scenario C, that non-response was related to an increase in pain relief (and a patient would re-enter the study when his pain relief decreased) there would again be no evidence to support the superiority of the experimental drug.

The results of using model (5.6) under these four scenarios are displayed in Figure 5.2 where, in each case, time points up to and including the values on the x-axes have been considered (thus results in Table 5.2 correspond to the rightmost intervals). As more time points are included in the analysis, the number of missing scores increases and the intervals of ignorance get wider. Again, except under scenario B, there is a large uncertainty as to the real efficacy of the experimental drug.

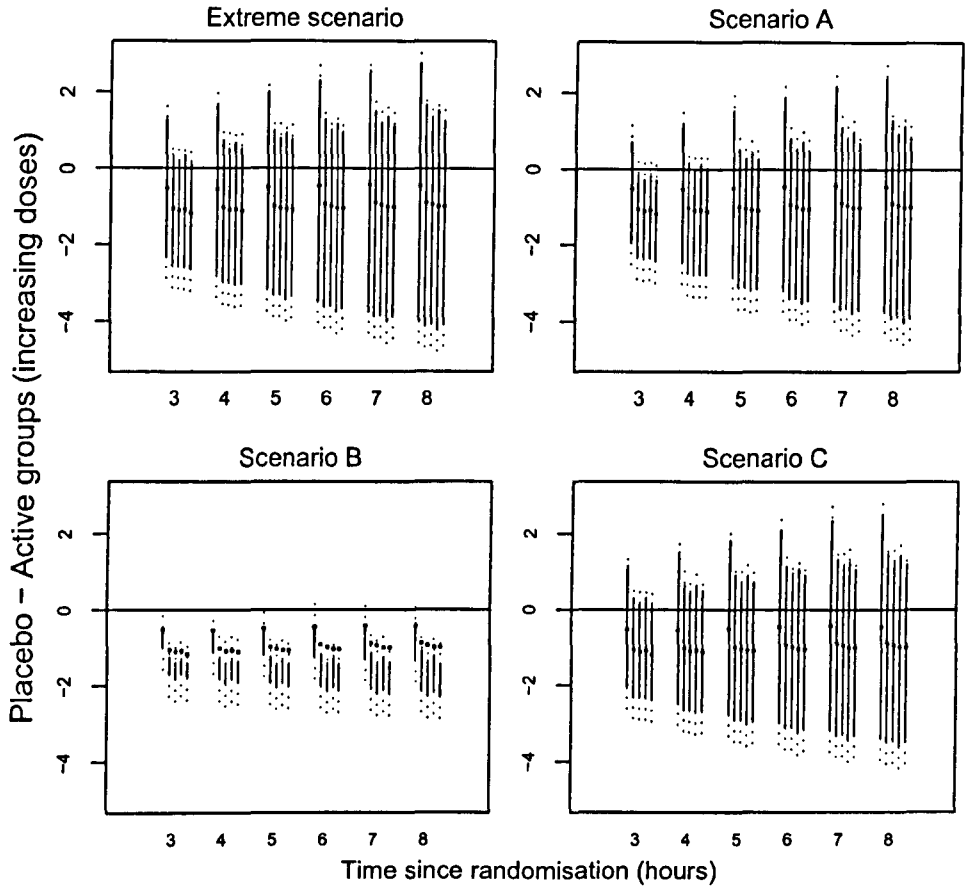

Figure 5.2: Intervals of ignorance (solid lines) and uncertainty (dotted lines) for differences between placebo and active groups under various scenarios for the missing data (at each time point, increasing dose levels are shown from left to right). Squares represent point estimates from the fit of model (5.6) to all available cases.

In conclusion, we have shown how standard iterative procedures used to fit marginal models for longitudinal categorical data can be modified to yield intervals of ignorance and uncertainty for parameters in the presence of missing data. These are given by the minimum and maximum values of parameter estimates and corresponding $100(1-\alpha)\%$ confidence intervals as missing measurements are allowed to vary within specified ranges. Predictably, the choice of this range has a considerable impact on the final results.

However, in most practical situations, one should be able to restrict this range to plausible values, thus obtaining sensible intervals of ignorance and uncertainty. Furthermore, given the flexibility and ease of implementation of this method, one can check how results vary under different scenarios. In this way, the robustness of the conclusions from a complete or all available case analysis can be readily assessed. In the example we saw that the treatment effect is likely to remain significant at all dose levels except the lowest under the plausible assumption that patients who missed a visit or were lost to follow-up would have shown the same or worse pain relief had they remained on treatment (Scenario B). This is because we started from highly significant treatment effects; had the latter not been statistically significant, the corresponding intervals of uncertainty would have presumably included zero even under this scenario.

An interesting feature of this approach and one that adds to its flexibility is that uncertainty about any coefficient and model can be readily assessed. For the dental pain trial for instance, we also considered uncertainty about the interaction terms between treatment arms and time. The latter were not statistically significant when considering the available cases only and the corresponding intervals of ignorance and uncertainty included zero under all four scenarios (results not shown).

The method can be extended to allow for estimation of regions of ignorance and uncertainty for two or more parameters simultaneously using, for instance, the dimension reduction approach as in Vansteelandt and Goetghebeur (2001). Further, the approach is still valid in case of independent observations or if we replace (5.4) with iterative weighted least squares (or a multivariate extension of it) .

In the next Section we develop methods for constructing intervals of ignorance and uncertainty for parameters in random effects models with incomplete con-

tinuous data. A modified Iterative Generalized Least Squares (IGLS) algorithm will be presented and used to assess the uncertainty about treatment difference between Ropinirole and Levodopa in the Parkinson's disease trial described in Subsection 2.1.1.

5.3 Bounds on parameter estimates with incomplete continuous data

We mentioned earlier that the idea of replacing point estimates for parameters of interest with optimistic-pessimistic bounds to account for the uncertainty induced by data incompleteness is particularly appealing with categorical outcomes, as in this case the set of possible estimates corresponding to possible data completions is closed. With a continuous outcome there is of course no such closed set as the number of possible data completions and corresponding point estimates is infinite. A possible approach would be to use the methods of the previous Section on a categorized version of the original continuous outcome. The problem with this approach is the inevitable loss of information associated with such transformations.

An alternative way of proceeding, and one that we adopt here, consists in bounding the values that missing continuous outcomes can take and then applying an approach similar to that of the previous Section. Although this may seem a crude approximation, in many circumstances the analyst can define sensible bounds for the values that missing data can take. Consider for example the Parkinson's disease trial. There, measurements consisted of discrete ADL scores ranging from 1 to 52; these values therefore provide bounds for the missing scores. Thus the distributional assumption made about the missing data concerns their domain.

After presenting our modified IGLS algorithm, we conduct a sensitivity analysis similar to the one of the previous Section, by considering how the intervals of ignorance and uncertainty for the coefficients of interest change as we vary the

range of values that missing data are allowed to take.

5.3.1 A modified IGLS algorithm

First, we review the Iterative Generalized Least Squares (IGLS) method for fitting linear mixed effect models (Goldstein, 1986, 1995). A modified IGLS algorithm will then be described and used to construct intervals of ignorance and uncertainty for parameter estimates.

Using the notation introduced in Subsection 2.1.1, for the measurement taken on subject i at visit j, a linear mixed model can be written as

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b} + e_{ij}, \tag{5.7}$$

where $\mathbf{X}_{ij}$ is the j-th row of the $n_i \times p$ design matrix $\mathbf{X}_i$ of explanatory variables for the fixed effects part of the model, n_i the number of observations taken on subject i, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect parameters, $\mathbf{Z}_{ij}$ is the j-th row of the $n_i \times q$ design matrix $\mathbf{Z}_i$ for the random part of the model, $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{D})$ and $e_{ij} \sim N(0, \sigma^2)$ for all $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$.

If the variance-covariance parameters in $\mathbf{D}$ are known, indicating with $\mathbf{X}$ the $(\sum_i^m n_i = N) \times p$ matrix obtained by stacking the $\mathbf{X}_i$ on top of each other and with $\mathbf{Y} = (y_{11}, y_{12}, \ldots, y_{n_m m})$ the $N \times 1$ vector of responses, the Generalized Least Squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} \tag{5.8}$$

where V is block diagonal with generic block corresponding to subject i given by $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \oplus_i\sigma^2$, $\oplus_i$ indicating the $n_i \times n_i$ identity matrix. In general, the elements of $\mathbf{D}$ have to be estimated; the IGLS algorithm then iterates between estimating σ and the distinct parameters in $\mathbf{D}$ and updating the estimate of $\boldsymbol{\beta}$ using (5.8). Appendix D describes the IGLS algorithm in more detail.

The focus will generally be on assessing the uncertainty about the fixed effects parameters in (5.7). With incomplete data, in order to obtain intervals (or regions) of ignorance and uncertainty, we exploit the fact that, given current estimates of the elements in $\mathbf{D}$, expression (5.8) is a linear function of the response variable. This leads to a modified IGLS algorithm where, as in the algorithm described in the previous Section, at each step a constrained minimisation (maximisation) is used to find the minimum (maximum) update of the fixed coefficient of interest with constrains given by the assumed ranges for the missing data.

If interest lies in measuring the uncertainty for two or more parameters simultaneously, a dimension reduction approach can be used as described in Vansteelandt and Goetghebeur (2001) which will produce regions of ignorance (e.g. a convex set for two parameters or convex hull for three parameters). The idea is to consider the minimum and maximum of linear combinations of the parameters along a chosen direction in the unit hypersphere. By varying these directions we outline the convex hull of ignorance for the parameters considered; this collapses to an interval of ignorance when considering a direction corresponding to a particular coefficient.

Thus, indicating with $\mathbf{w}$ a direction in the unit hypersphere S^p, the modified IGLS algorithm is implemented as follows:

- step 1: obtain initial estimates of all relevant parameters from the fit of (5.7) to the available data;
- step 2: find the minimum and maximum of

$$\mathbf{w}^T\hat{\boldsymbol{\beta}} = \mathbf{w}^T(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{Y}$$

where the entries in $\mathbf{Y}$ corresponding to missing data are constrained to take values in the chosen ranges, as discussed before. Notice that, because of the

linearity in $\mathbf{Y}$ of the expression above, finding the minimum and maximum will reduce to imputing the lower or upper bounds of the domains for the missing data, the choice depending on the sign of the corresponding elements in $\mathbf{w}^T(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}$;

- step 3: obtain estimates of the vector of fixed coefficients $\boldsymbol{\beta}$ corresponding to the minimum and maximum, $\hat{\boldsymbol{\beta}}_{min}$ and $\hat{\boldsymbol{\beta}}_{max}$ respectively;
- step 4: obtain new estimates of $\mathbf{V}$ corresponding to $\hat{\boldsymbol{\beta}}_{min}$ and $\hat{\boldsymbol{\beta}}_{max}$, $\widehat{\mathbf{V}}_{min}$ and $\widehat{\mathbf{V}}_{max}$ respectively;
- step 5: find the minimum of

$$\mathbf{w}^T\hat{\boldsymbol{\beta}} = \mathbf{w}^T(\mathbf{X}^T\widehat{\mathbf{V}}_{min}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}_{min}^{-1}\mathbf{Y}$$

and maximum of

$$\mathbf{w}^T\hat{\boldsymbol{\beta}} = \mathbf{w}^T(\mathbf{X}^T\widehat{\mathbf{V}}_{max}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}_{max}^{-1}\mathbf{Y};$$

- Step 6: iterate between steps 3 to 5 until convergence.

The procedure is then repeated a sufficient number of times corresponding to different directions $\mathbf{w}$ to map out the convex hull of ignorance. Thus, if considering for example the region of ignorance for three parameters simultaneusly, the algorithm will yield values for the three parameters in $\mathbb{R}^3$ that correspond to the minimum and maximum of (5.8) along each of the chosen directions. The convex hull corresponding to this set of points gives a graphical representation of the ignorance about the three estimates induced by data incompleteness.

The directions along which minima and maxima are calculated can be obtained by sampling at random points on the surface of the unit hypersphere S^p. Here, we shall restrict ourselves to the 3-dimensional case. To select a random point on the surface of the unit sphere, one could pick spherical coordinates θ and

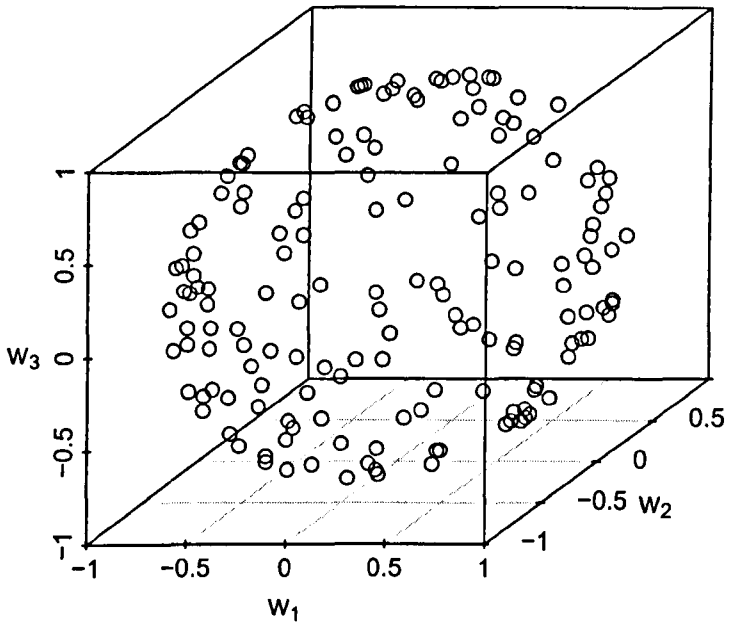

Figure 5.3: 500 uniformly distributed points on the surface of the unit sphere identifying uniformly distributed directions in $\mathbb{R}^3$.

ϕ from uniform distributions $\theta \in [0, 2\pi)$ and $\phi \in [0, 2\pi)$. However, this approach would lead to oversampling near the poles. Instead we use the method proposed by Marsaglia (1972) which produces a set of points uniformly distributed on the unit sphere. Choose pairs (x_1, x_2) from independent uniform distributions on $(-1, 1)$ and reject pairs for which $x_1^2 + x_2^2 \geq 1$. From the remaining pairs, obtain the coordinates of the points on the surface as

$$w_1 = 2x_2\sqrt{1 - x_1^2 - x_2^2}$$
$$w_2 = 2x_2\sqrt{1 - x_1^2 - x_2^2}$$

$$w_3 = 1 - 2(x_1^2 + x_2^2).$$

An example is given in Figure 5.3 where the 500 points plotted identify directions $\mathbf{w}$ to be used in the modified IGLS algorithm.

As mentioned before, when considering the single direction corresponding to the jth parameter $\mathbf{w} = \mathbf{e}_j = (0, 0, \ldots, 1, \ldots, 0)^T$ given by a vector of all zeros apart from the jth element which is equal to one, the modified IGLS algorithm will yield the interval of ignorance for $\hat{\beta}_j$, the jth entry of the vector $\hat{\boldsymbol{\beta}}$. In this case $100(1-\alpha)\%$ intervals of uncertainty can be derived as in the previous Section, i.e. as the interval joining the lower and upper limits of the classical $100(1-\alpha)\%$ confidence intervals for $\hat{\beta}_{j_{min}}$ and $\hat{\beta}_{j_{max}}$ respectively. A generalisation of the concept of interval of uncertainty to two or more dimensions is also possible but will not be considered here.

5.3.2 Simulation study

To illustrate our approach, we present results from a small simulation study that considers the 3-dimensional case, i.e. we will construct the convex hull of ignorance for three parameters. Repeated measurements were simulated for 200 subjects split into two equally sized groups, at 5 time points using the linear mixed model

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j + \beta_2 \text{group}_i + \beta_3 \text{group}_i t_j + e_{ij} \qquad (5.9)$$

where β_3 is a slope-by-group fixed interaction term, $\mathbf{b} \sim N_2(\mathbf{0}, \mathbf{D})$, $e_{ij} \sim (0, \sigma^2)$ for all $i = 1, \ldots, 200$, $j = 1, \ldots, 5$ and

$$\mathbf{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}.$$

True values for the parameters were $\beta_0 = 1.0$, $\beta_1 = 1.5$, $\beta_2 = 3.0$, $\beta_3 = 1.5$, $\sigma^2 = 1.5$ and

$$\mathbf{D} = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}.$$

About 50% of the observations within each group were randomly deleted to produce a pseudo-incomplete data set. In order to apply the methods described above, ranges for the missing data were defined. As discussed in the previous section, this choice will often involve restricting ranges for the unseen data, perhaps drawing on expert opinion to do so. However, by repeating the analysis under different scenarios for the missing data, the extent of the uncertainty about paramater estimates due to data incompleteness can be readily assessed. In particular, in this simulation study, a sensitivity analysis was conducted with missing data allowed to take values in symmetric intervals about the predicted values obtained by fitting (5.9) to the 'observed' data. The widths of these intervals were defined as fractions of the estimated variance of the measurement error, $\hat{\sigma}^2$. Results for two scenarios will be presented. Let, $\hat{y}_{ij}$ indicate the predicted value for a missing observation from (5.9),

$$\hat{y}_{ij} = (\hat{\beta}_0 + \hat{b}_{0i}) + (\hat{\beta}_1 + + \hat{b}_{1i})t_j + \hat{\beta}_2 \text{group}_i + \hat{\beta}_3 \text{group}_i t_j. \qquad (5.10)$$

Then, two sets of domains for the missing data were considered: $\hat{y}_{ij} \pm 0.5\hat{\sigma}^2$ (Scenario A) and $\hat{y}_{ij} \pm \hat{\sigma}^2$ (Scenario B).

The convex hull of ignorance for estimates of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ under Scenario A is shown in Figure 5.4. This was obtained considering the minimum and maximum of the set of estimating equations (5.8) corresponding to model (5.9) along each on 300 uniformly distributed directions in the unit sphere. Notice that the convex hull of ignorance is an approximation of the true surface of ignorance

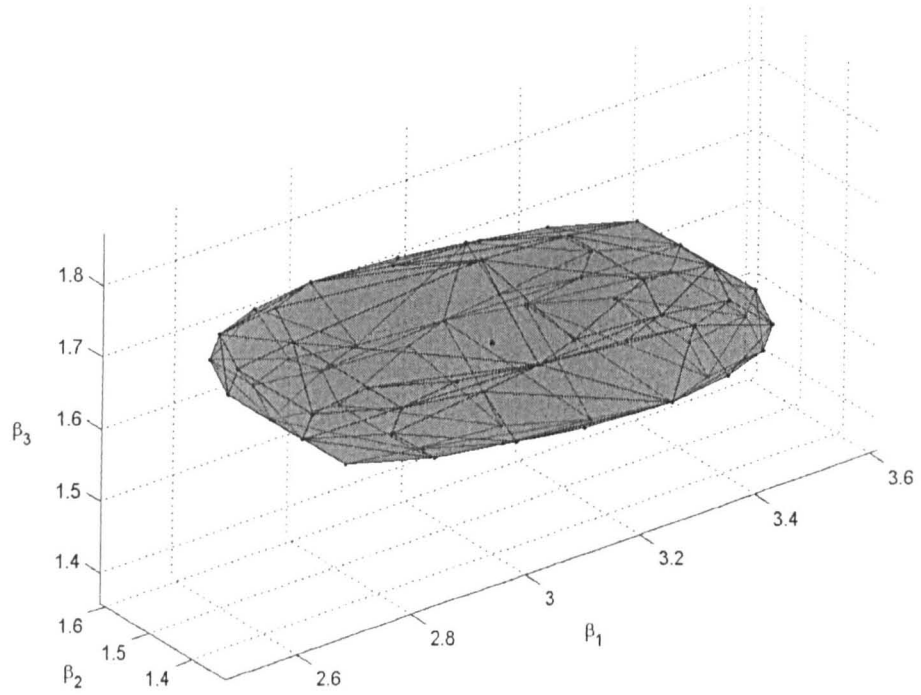

Figure 5.4: Region of ignorance for parameters β_1, β_2 and β_3 in (5.9) when each missing measurement is allowed to take value in the interval defined as $\hat{y}_{ij} \pm 0.5\hat{\sigma}^2$ (Scenario A). $\hat{y}_{ij}$ and $\hat{\sigma}^2$ are the predicted values for the mean of the missing observations and the estimated variance of the error term, respectively, from the fit of (5.9) to the available cases.

in three dimensions corresponding to the three parameters considered. In fact, by definition, the convex hull of a data set in n-dimensional space is the smallest convex region that contains the data set, the latter being in our case the minima and maxima obtained from the modified IGLS algorithm.

Note, the approximation can be improved by considering more directions on the unit sphere.

From inspection of Figure 5.4, it appears that, for the particular simulated data set considered, there is larger ignorance about β_2, the overall group difference as opposed to the overall slope and slope-by-group interaction, β_1 and β_3 respec-

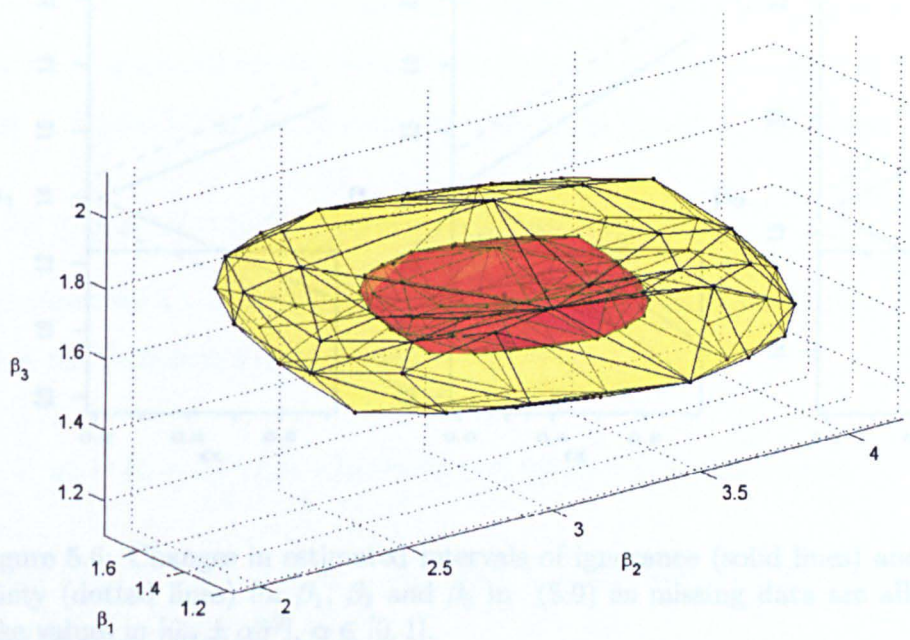

Figure 5.5: Convex hull of the region of ignorance for parameters β_1, β_2 and β_3 in (5.9) when each missing measurement is allowed to take values in the interval defined as $\hat{y}_{ij} \pm 0.5\hat{\sigma}^2$ (inner hull, Scenario A) and $\hat{y}_{ij} \pm \hat{\sigma}^2$ (outer hull, Scenario B). $\hat{y}_{ij}$ and $\hat{\sigma}^2$ are the predicted value for the mean of the missing observations and the estimated variance of the measurement error term, respectively, from the fit of (5.9) to the available cases.

tively. In Figure 5.5, the convex hull under scenario A in plotted within the convex hull corresponding to scenario B. The substantially larger ignorance about all three parameters under the latter scenario is a direct consequence of the wider domains for missing data.

A univariate analysis was then considered looking at each of the three parameters separately, that is, considering the intervals of ignorance and uncertainty corresponding to directions $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$. In the univariate case, a sensitivity analysis considers how these intervals change as the domains for the missing data are gradually widened from the predicted values $\hat{y}_{ij}$ to the intervals $\hat{y}_{ij} \pm \hat{\sigma}^2$ where, as before, the predicted values are from the fit of (5.9) to the incomplete data set.

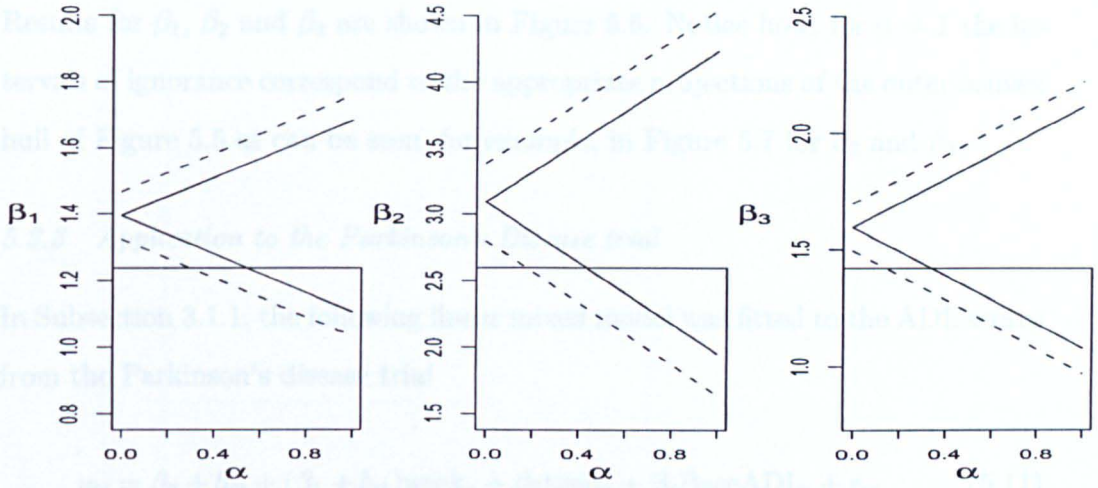

Figure 5.6: Changes in estimated intervals of ignorance (solid lines) and uncertainty (dotted lines) for β_1, β_2 and β_3 in (5.9) as missing data are allowed to take values in $[\hat{y}_{ij} \pm \alpha\hat{\sigma}^2]$, $\alpha \in [0, 1]$.

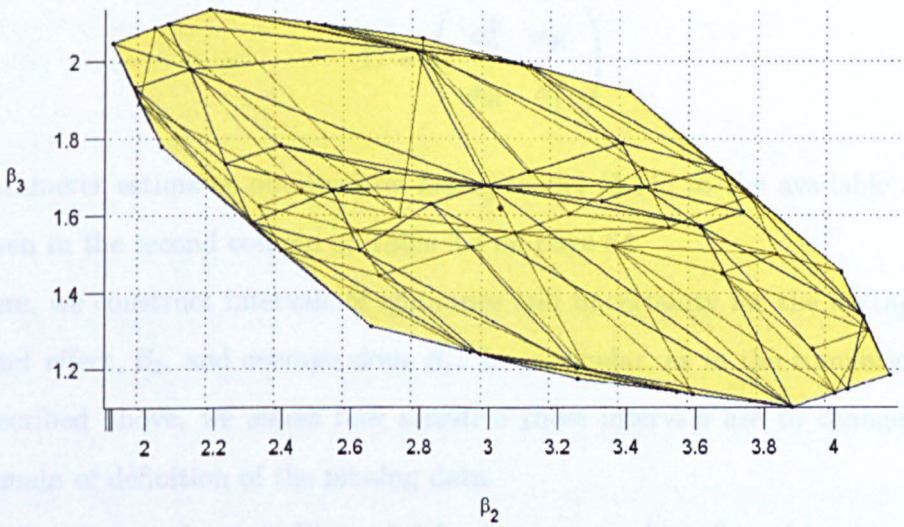

Figure 5.7: Projection of the outer convex hull in Figure 5.5 on the (β_3, β_2) plane; further projections on the axes of β_2 and β_3 correspond to the intervals of ignorance for β_2 and β_3 in Figure 5.6 when $\alpha = 1$.

Results for β_1, β_2 and β_3 are shown in Figure 5.6. Notice how, for $\alpha = 1$ the intervals of ignorance correspond to the appropriate projections of the outer convex hull of Figure 5.5 as can be seen, for example, in Figure 5.7 for β_2 and β_3.

5.3.3 Application to the Parkinson's Disease trial

In Subsection 3.1.1, the following linear mixed model was fitted to the ADL scores from the Parkinson's disease trial

$$y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})\text{week}_j + \beta_2\text{treat}_i + \beta_3\text{BaseADL}_i + e_{ij}, \qquad (5.11)$$

where $i = 1, \ldots, 250$, $j = 1, \ldots, n_i$, $\text{treat}_i = \{1(\text{Ropinirole}), 0(\text{Levodopa})\}$, $e_{ij} \sim N(0, \sigma^2)$, $\mathbf{b}_i \sim N_2(0, \mathbf{D})$ and

$$\mathbf{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}.$$

Parameter estimates obtained by fitting model (5.11) to the available data are given in the second column of Table 3.4 on page 64.

Here, we construct intervals of ignorance and uncertainty for the average treatment effect, β_2, and average slope β_1. In particular, as in the simulation study described above, we assess how sensitive these intervals are to changes in the domain of definition of the missing data.

Results are shown in Figure 5.8 for the average slope β_1 and treatment effect β_2. We see that there is much greater uncertainty about the average treatment effect than the average slope. The former is no longer statistically significant when missing data are allowed take value in intervals defined as $\hat{y}_{ij} \pm 0.23\hat{\sigma}^2 = \hat{y}_{ij} \pm 0.23 \times 5.4$ (recall that $\hat{y}_{ij}$ refers to the predicted value for a missing ADL score as obtained from the fit of (5.11) to the available cases).

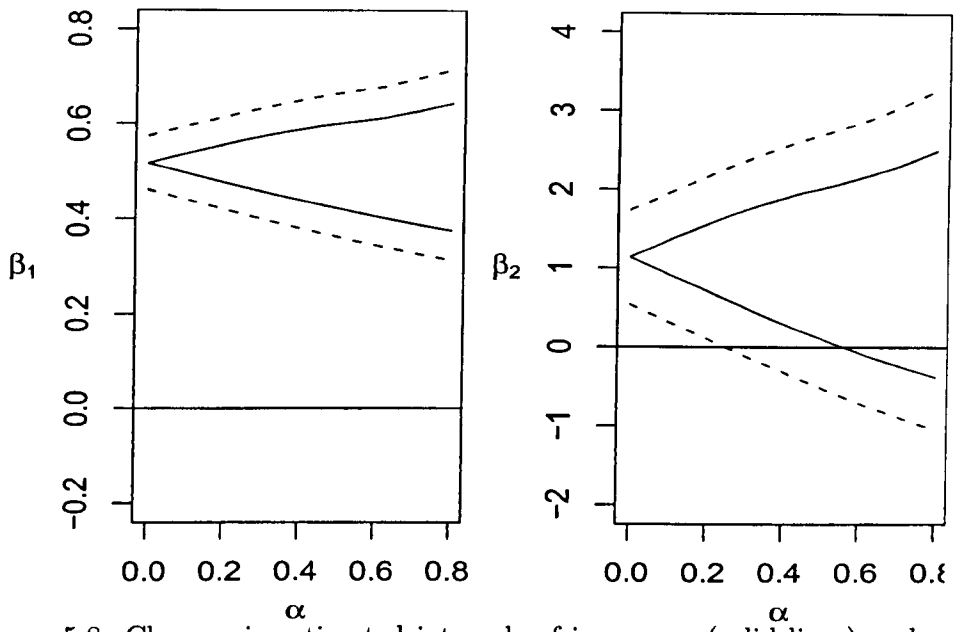

Figure 5.8: Changes in estimated intervals of ignorance (solid lines) and uncertainty (dotted lines) for the average slope β_1 and treatment effect β_2 in (5.11) as missing data are allowed to take values in $[\hat{y}_{ij} \pm \alpha\hat{\sigma}^2]$, $\alpha \in [0, 0.8]$. $\hat{y}_{ij}$ are the predicted values for the mean of the missing observations from fitting model (5.11) to the available cases.

Finally, we conducted a sensitivity analysis similar to that performed in the previous Section with the dental pain trial. Recalling that higher ADL scores indicate a deterioration of the patient's ability to perform daily activities, it seems sensible to assume that patients who dropped out of the study would have shown higher values. Therefore, we allow the missing observations after a patient's dropout to be the same or greater (by no more than 4 or 8 scores under scenarios A and B, respectively) than their last observed measurement. The chosen ranges allow for a significant increase in ADL scores (see Figures 3.1 and 3.2 on pages 48 and 49).

For β_1 (average slope) and β_2 (average treatment effect) in model (5.11), the intervals of ignorance and uncertainty thus obtained are given in Table 5.3. Interestingly, under these clinically sensible scenarios, the average treatment difference is likely to remain (statistically) significantly in favor of Levodopa.

Parameter	Scenario	Interval of ignorance	Interval of uncertainty
β_1 (Avg. slope)	A	(0.435,0.572)	[0.361,0.648]
	B	(0.384,0.652)	[0.305,0.737]
β_2 (Avg. treat. effect)	A	(1.011,1.111)	[0.330,1.845]
	B	(1.066,1.181)	[0.345,1.970]

Table 5.3: Intervals of ignorance and uncertainty for β_1 and β_2 in (5.11). Scenario A: missing observations same or greater (by no more than 4 scores) than last observed measurement; Scenario B: missing observations same or greater (by no more than 8 scores) than last observed measurement.

5.4 Discussion

In this Chapter, we extended the intervals of ignorance and uncertainty of Vansteelandt and Goetghebeur (2001) to longitudinal discrete and continuous data settings. With discrete data, when no restrictions are placed on the values that missing data can take, the intervals of ignorance correspond to best-worse case estimates. With continuous data, bounds on the values that missing data can take have to be specified in order to proceed. In many cases, this is not as big an obstacle as it may appear as these bounds can be sensibly defined, as with the ADL scores from the Parkinson's disease trial. In both the discrete and continuous data cases, sensitivity analysis can be readily conducted by varying the domains assumed for the missing data and looking at the impact that this has on the intervals of ignorance and uncertainty.

By contrast, the sampling-based sensitivity analysis described in the previous Chapter also yielded (loosely termed) intervals of 'ignorance' for the coefficient of interest, conditional on the model assumed for the dropout mechanism (as in Molenberghs et al. (2001)). However, a drawback of the sampling-based sensi-

tivity approach is that is is often difficult to justify the domain(s) chosen for the sensitivity parameter(s) in the model for the dropout mechanism. In terms of the sensitivity of the results, a very different picture may emerge using a different domain for a sensitivity parameter or a different model for the dropout process or both. This is different to the methods described in this Chapter where for example, under the extreme scenario of Subsection 5.1.2, no assumptions are made about the missing data. The interval of ignorance for a coefficient then replaces the usual point estimate and identifies the range of possible estimates that are compatible with the observed data. Admittedly, these intervals may at times be very wide and almost of no practical value. However, we agree with Manski (1989) in saying that, in these cases, the wide intervals of ignorance just reflect the fact that no definitive conclusions are possible without making further assumptions (about, for example, the mechanism driving the missing data process). Furthermore, by using these methods, we are able to explore 'what-if' scenarios regarding the missing data in a direct way as we saw in the application to both the Parkinson's disease and dental pain trials, making results extremely easy to convey.

This approach is particularly attractive when dealing with incomplete discrete data especially ordered categorical data as one can then assess the effect of restricting the domain for the missing data to one or more categories and quantify the impact that this has on the intervals of ignorance and uncertainty. For example, one can then look for a possible domain that makes a statistically significant result from an analysis from available data no longer significant or vice versa while assessing its likelihood from a clinical perspective (as we did in Section 5.2 for the dental pain trial).

The coverage properties of the intervals of uncertainty are currently being investigated (Vansteelandt et al., 2002) and certainly further research in this area is required. In particular, intervals of uncertainty obtained by adding $(1-\alpha)100\%$

confidence limits to the estimated ignorance limits are referred to by Vansteelandt et al. (2002) as strong intervals; they show that their coverage levels lie between $(1-\alpha)$ and $(1-\alpha/2)$. As mentioned earlier in this Chapter, an alternative approach for the construction of intervals of uncertainty is the bootstrap method. As usual with the bootstrap, its flexibility has to be weighted against the computational burden which, however, should in general still be acceptable for the algorithms of Subsections 5.1.1 and 5.3.1. Another issue that needs to be explored further is related to the fact that, the ignorance intervals refer to estimates obtained under all possible data completions; for some pseudo-complete samples however, the fit of the chosen model for the response variable may be grossly inadequate. Indeed, if it were possible to restrict the analysis to those pseudo-complete data sets for which the model fit is acceptable, this would certainly reduce the width of the intervals of ignorance. A possible way of addressing this problem is to consider nested models for the response variable and the use the statistical significance levels of the intervals of uncertainty as a criteria for model reduction (as we did for the treatment-by-time interaction term in the dental pain trial).

In the next Chapter, we build on the idea of intervals of ignorance and uncertainty. Rather than quantifying the lack of knowledge caused by missing data just in terms of ignorance limits, we argue that it is more sensible to consider the proportion of possible estimates of a parameter of interest that are greater or smaller than some threshold value chosen by the analyst, over all estimates arising under all possible ways of completing the sample.

Chapter 6

Further methods for understanding the uncertainty about parameter estimates due to data incompleteness

In the previous Chapter, the uncertainty about coefficients of interest caused by data incompleteness was quantified in terms of the possible estimates which, under various assumptions about the missing data, are compatible with the observed data. In particular, intervals or regions of ignorance were derived that correspond to optimistic-pessimistic bounds on parameter estimates when, for example, the outcome variable is discrete and no constraints are put on the values that missing data can take (the extreme scenario of Subsection 5.1.2).
As a measure of ignorance however, these intervals are quite sensitive to extreme results arising under particular data completions. For this reason, with discrete data, a more appropriate measure is the proportion of possible estimates that are greater (or smaller) than certain thresholds specified by the analyst, possibly weighted by the likelihood of such estimates under certain models.

In the first part of this Chapter we present a modified Fisher scoring algorithm with a nested saddlepoint approximation that allows rapid calculation of the proportion of parameter estimates above or below a threshold over all values

arising from all possible sample completions (Verzilli and Carpenter, 2002c). The general idea underlying the method is as follows: first, a data completion yielding an estimate as close as possible to the chosen threshold is found and then, the proportion of data completions giving rise to estimates above the threshold is calculated. The latter step is carried out using a saddlepoint approximation to the distribution function of the missing part of the relevant sufficient statistics.

If sample completions are weighted using probabilities corresponding to plausible scenarios for the missing data, the methods presented yield weighted proportions instead. Thus, ignorance about point estimates can be investigated further by means of a sensitivity analysis, varying these weights under different clinically plausible scenarios for the missing data. For example, one could set the weights equal to the predicted probabilities for the missing data, where the predicted probabilities are obtained from the fit of the response model to the observed data thus assuming random missingness in the sense of Rubin (1976): ignorance about a point estimate would then be reported in terms of the (weighted) proportion of estimates that are greater than the chosen threshold across all enumerable solutions, were the missing at random assumption to hold true.

We can further get confidence intervals for these proportions, which account for familiar sampling variability, by bootstrapping the observed part of the data and re-running the algorithm on each bootstrapped sample. The proposed approach bears close similarities with multiple imputation techniques; however, it has the distinct advantage of being computationally efficient, providing a solution close to that arising from complete imputation in a computational time that is linear in the number of missing observations.

Bootstrap confidence intervals for the proportion of estimates above a threshold are conceptually similar to the intervals of uncertainty of the previous Chapter: they attempt to address at the same time both our lack of knowledge due

to the missing data and traditional sampling variability. In the second part of this Chapter, we take a different approach to this end. We extend methods used for exact conditional inference in generalized linear models to allow for the extra uncertainty caused by missing data with a particular focus on incomplete binary data. This time, we consider the one-to-one map from the set of possible values of the missing part of the sufficient statistic corresponding to a parameter of interest to the set of possible p-values (approximated using the double saddlepoint method of Davison (1988)). Importance sampling of the missing part of the sufficient statistic is then used to obtain a Monte Carlo approximation to the 'average' p-value over all possible sample completions.

The Chapter is organized in 7 sections. We start by reviewing saddlepoint techniques to approximate the density and distribution function of the sum of random variables, as they play a key role in what follows. In Section 2 and 3 we describe the methods used for computing the proportion of estimates above a threshold when data are incomplete in the context of generalized linear models; for binary and Poisson data, results from simulation studies are presented in Section 4. The application of the proposed approach to the dental pain trial is discussed in Section 5 where, for the purposes of illustrating our methods, we will consider only one time point (8 hours since randomisation) and use a dichotomised version of the original ordinal outcome. Section 6 illustrates the method used to approximate the 'expected' p-value for a particular coefficient in the context of incomplete binary data based on importance sampling Monte Carlo integration. We end this Chapter with a discussion in Section 7.

6.1 Saddlepoint approximation of the density and distribution function of the sum of random variables

Consider n i.i.d. random variables $Y_1, \ldots, Y_n$ with density, moment and cumulant generating functions $f(y)$, $M(\alpha) = \int_{-\infty}^{+\infty} e^{\alpha y} f(y)\, dy$ and $K(\alpha) = \ln M(\alpha)$, respectively. Then, the saddlepoint approximation to the density at t, $f(t)$ is

$$\left[\frac{1}{2\pi n K''(\alpha_0)}\right]^{1/2} \exp\{nK(\alpha_0) - \alpha_0 t\} \tag{6.1}$$

where α_0 is the saddlepoint, a root of

$$nK'(\alpha) - t = 0$$

and K' and K'' are the first and second derivatives of the cumulant generating function.

A saddlepoint approximation of tail area is obtained as

$$P(S \geq s) \approx 1 - \Phi(A) - \phi(A)\left(\frac{1}{A} - \frac{1}{B}\right) \tag{6.2}$$

where in the latter expression

$$A = \operatorname{sgn}(\alpha_0)\{2[\alpha_0 s - nK(\alpha)]\}^{1/2}$$
$$B = \alpha[nK''(\alpha_0)]^{1/2}.$$

In the case of lattice distributions (when approximating sums of discrete random variables), throughout the expressions above, the saddlepoint α_0 now solves

$$nK'(\alpha) = s - \frac{1}{2}$$

and

$$B = 2\sinh(\alpha_0/2)[nK''(\alpha_0)]^{1/2}$$

is used in (6.2) (Skovgaard, 1987).

In Appendix E, we give further details on the saddlepoint method mainly based on the illustration of the technique given in Field and Ronchetti (1990).

6.2 Estimates above a threshold as a measure of ignorance

In this Section, we describe a computationally efficient approach that enables calculation of the proportion of estimates that are greater that an user-specified threshold over all estimates corresponding to all possible sample completions. The approach is valid for discrete outcome variables with distributions belonging to the exponential family modelled using canonical link functions. In particular, for binary and Poisson data, results from simulation studies will be presented. Note that we will only consider the case of cross-sectional data here; thus, for example, in the application to the dental pain trial, the analysis will consider data collected at 8 hours after start (which, incidentally, was the primary endpoint stated in the study protocol).

Indicate with $\mathbf{Y} = (Y_1, \ldots, Y_n)$ a vector of length n with components having distribution belonging to the exponential family i.e.

$$f(y_i; \theta_i, \phi) = \exp\{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\} \tag{6.3}$$

for known functions $a(.)$, $b(.)$ and $c(.)$. Using standard GLM notation, the link function $g(.)$ relates the expected values of $\mathbf{Y}$, $\boldsymbol{\mu}$, to some covariates of interest $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$, that is

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \tag{6.4}$$

for $i = 1, \ldots, n$.

Denote by $\mathbf{y}$ the vector of *intended* observations. With missing measurements, partition $\mathbf{y}$ into the observed and missing part $\mathbf{y} = (\mathbf{y}^{obs}, \mathbf{y}^{miss})$ where $\mathbf{y}^{miss}$ is a vector of size m, the number of missing data, and $\mathbf{y}^{obs}$ is of size $(n - m)$. Also, assuming that the discrete outcome can only take C values, indicate with $\mathcal{M}$ the set of all C^m possible values of the vector $\mathbf{y}^{miss}$.

Consider now the finite-dimensional set $\mathcal{B}$ of estimates of the vector of parameters β corresponding to all possible sample completions that is $\mathcal{B} = \{\hat{\beta}^k, k = 1, \ldots, C^m\}$ and denote by $|\mathcal{B}|$ the number of elements in this set. Let $\mathbf{B}_j$ indicate the subset of $\mathcal{B}$ in which the j-th element of the parameter vector β is greater than some threshold value specified by the analyst. The aim is then to compute the proportion $|\mathbf{B}_j|/|\mathcal{B}|$ i.e. the proportion of possible estimates of β_j that are greater than the chosen critical value.

This is achieved using a modified version of standard iterative algorithms; in particular, we use a nested saddlepoint approximation to the distribution of the part of the relevant sufficient statistics that refer to the missing observations where the latter have probabilities specified by the analyst. Notice that, when all outcomes for the missing observations are assumed to be equally likely (corresponding to an uncertain scenario where all completions are assumed equally likely), $|\mathbf{B}_j|/|\mathcal{B}|$ is also the probability of obtaining estimates of β_j that are greater than the threshold, which we approximate using the saddlepoint method. In other words, our modified Fisher scoring algorithm gives probabilities of observing estimates above the threshold, which correspond to proportions $|\mathbf{B}_j|/|\mathcal{B}|$ only when assigning an equal probability mass to each sample completion. Thus, in this case, results from our algorithm can be checked against the exact proportions obtained from

enumerating all possible estimates.

Later, we will also conduct a sensitivity analysis by varying the probabilities attached to missing observations. In particular, in the application to the dental pain data these weights will range from the predicted probabilities obtained from the MAR model fitted to the available cases to more clinically plausible scenarios where the chance of an improvement for the missing observations are gradually decreased from the MAR prediction. In this case, the methods described here will yield probabilities corresponding to weighted proportions with weights given by the (different) probabilities associated with each sample completion.

6.3 Computing the proportion of estimates above a threshold

As a preliminary step to calculating the proportion $|\mathbf{B}_j|/|\mathcal{B}|$, bounds on parameter estimates are obtained using the method described in Subsection 5.1.1 for a binary outcome modelled using logistic regression. Let $\hat{\eta}_i^{(k)}$ and $\hat{\mu}_i^{(k)}$ be estimates of the linear predictor and expected value respectively at iteration k and form adjusted dependent variables as

$$z_i^{(k)} = \hat{\eta}_i^{(k)} + (y_i - \hat{\mu}_i^{(k)})\left(\frac{d\eta}{d\mu}\right)_i^{(k)}.$$

A Fisher scoring algorithm updates the current value of the generic parameter β_j to

$$\hat{\beta}_j^{(k+1)} = \hat{\beta}_j^{(k)} + \left\{\left(\mathbf{X}^T\widehat{\mathbf{W}}^{(k)}\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{W}}^{(k)}\left(\mathbf{z}^{(k)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(k)}\right)\right\}_j \tag{6.5}$$

where $\mathbf{W} = \text{diag}\left\{\widehat{V}_i^{-1}\left(\frac{d\mu}{d\eta}\right)_i^2\right\}$ and V denotes the variance function.
Without loss of generality, in the case of canonical link functions, we can

write (6.5) as

$$\hat{\beta}_j^{(k+1)} = \hat{\beta}_j^{(k)} + O_{obs}^{(k)} + c_1 y_1^{miss} + \cdots + c_m y_m^{miss} = \hat{\beta}_j^{(k)} + O_{obs}^{(k)} + S^{(k)} \qquad (6.6)$$

where $O_{obs}^{(k)}$ is the contribution of the observed data to the update of β_j and $\mathbf{c} = (c_1, \ldots, c_m)^T$ is an m-dimensional vector of constants that depends on current parameter estimates.

The range of possible estimates of β_j can be determined by exploiting the linearity of (6.6) in the missing data as seen in Subsection 5.1.1. At each step, for the maximum we find

$$\max_{\mathbf{y}^{miss} \in \mathcal{M}} S^{(k)}$$

which can be easily determined as it will only depend on the sign of the elements of the vector $\mathbf{c}$. Similarly, for the minimum of β_j the minimum of $S^{(k)}$ is found.

As mentioned at the beginning of this Chapter and shown in the next Section, the range of values thus obtained can appear symmetric about the chosen threshold when in fact the proportions of possible estimates greater or smaller than the threshold are very different. Hence, ignorance about parameter estimates can be explored further by looking at the proportion of estimates either side of the threshold. We describe next how this is achieved using a nested saddlepoint approximation.

6.3.1 The EAT algorithm for computing the proportion of parameter Estimates Above a Threshold

Calculation of the proportion of parameter estimates that are greater or smaller than a pre-specified value will only make sense if the interval calculated in the previous Section includes the threshold itself.

Assume without loss of generality a threshold value of zero. Then, starting with

the value of $\mathbf{y}^{miss}$ which yielded the upper bracket of the interval, the idea is to move towards the smallest possible positive estimate and determine the proportion of estimates that are greater than this value. This proportion could be calculated at each step of the algorithm to monitor its convergence as we do here, or as a one one-off calculation once convergence to the smallest positive estimate of β_j is seen.

Thus, the EAT algorithm is performed in 4 steps as follows

Step 1 (initialization): consider the quantities in (6.6) corresponding to the maximum estimate of the parameter of interest (the upper limit of the interval of ignorance).

Step 2: given the vectors $\mathbf{y}_{miss}$ and $\mathbf{c}$ which depend on current parameter estimates, find a solution to the constrained satisfaction problem (CSP)

$$\max_{\mathbf{y}^{miss} \in \mathcal{M}} S^{(k)}$$
$$T^{(k)} = -\beta_j^{(k)} - O_{obs}^{(k)} < S^{(k)} < 0 \tag{6.7}$$

i.e. find the value of $\mathbf{y}^{miss}$ and therefore of $S^{(k)}$ that gives a new estimate of β_j closest to $T^{(k)}$.

Simple backtracking or random sampling methods can be employed to solve this CSP (Tsang, 1994). Let us write $S^{(k)} = c_1 y_1^{miss} + c_2 y_2^{miss} + \cdots + c_m y_m^{miss} = Z_1 + Z_2 + \cdots + Z_m$ where $Z_l \in \mathcal{D} = \{D_1, D_2, \ldots, D_C\}$, $l = 1, \ldots, m$ and the elements in $\mathcal{D}$ depend on the C values taken by the particular discrete outcome considered. For binary data, for example, we have

$$Z_l = \begin{cases} c_l & \text{with probability } p_l \\ 0 & \text{with probability } 1 - p_l \end{cases} \tag{6.8}$$

where the probabilities p_l are specified by the analyst.

Backtracking (in its simplest version) consists then in successively initializing the variables Z_l, $l = 1, \ldots, m$, possibly sorted in decreasing order based on their domains, while condition (6.7) is satisfied and, in case of violation, backtracking to the last valid initialization, say $l = 3$, and proceed by choosing a different value from $\mathcal{D}$ for Z_4. Although this method yields the exact solution, for large m, it tends to be computationally inefficient.

A more efficient approach, and one that we adopt here, is based on random sampling the values of Z. Namely, we sample a large number of vectors $\mathbf{y}^{miss} \in \mathcal{M}$ (by sampling with replacement each element of $\mathbf{y}^{miss}$ a corresponding number of times as they are assumed independent) and then retain the particular vector giving the best approximation to (6.7). Although this method gives an approximate solution to the CSP, by calculating and plotting the proportion of estimates above the threshold at each step (Step 3 described next), one can monitor convergence to a stable solution. Formal tools for monitoring convergence could also be used as described in Gilks et al. (1996).

Step 3: in order to calculate the proportion $|\mathbf{B}_j|/|\mathcal{B}|$ (or weighted proportion when sample completions are not assumed equally likely), we need to calculate the distribution function of the sum $S^{(k)}$. For large values of m we would be tempted to approximate the distribution of $S^{(k)}$ by a normal distribution. However, the elements of $\mathbf{c}$ can be large or very small which violates the Lindeberg condition for the central limit theorem (see Feller (1966), Vol 2, page 491). Also, in the sensitivity analysis, we shall vary the probabilities assumed for different sample completions and these could at times be very small. Finally, $T^{(k)}$ could lie in the extreme tail of the distribution of $S^{(k)}$. We therefore use the saddlepoint approximation to the distribution function of $S^{(k)}$ using (6.2) with $s = T^{(k)}$.

Step 4: Update the estimate of β_j using (6.6).

Step 5: Iterate between steps 2-4 until convergence.

6.4 Simulation studies

We report the results of simulation studies conducted to compare the results of the EAT algorithm (using the nested saddlepoint approximation) with the exact results from complete enumeration. Recall that it is valid to compare probabilities produced by the EAT algorithm and the exact proportions obtained from enumerating all solutions only when assuming an equal probability mass for each sample completion or, equivalently, for each element of $\mathcal{B}$.
Here, we present results from logistic and Poisson regression.

6.4.1 Logistic regression

Data have been simulated from the logistic model

$$\text{logit}\{\Pr(y_i = 1)\} = \beta_0 + \beta_1 x_i \tag{6.9}$$

with x_i a generic normally distributed covariate. We simulated data under these conditions

(i) β_0=1, $\beta_1 = 0.01$, $n = 50$, $m = 12$

(ii) β_0=1, $\beta_1 = 0.005$, $n = 200$, $m = 50$

(iii) β_0=1, $\beta_1 = 0.05$, $n = 200$, $m = 50$

(iv) β_0=1, $\beta_1 = 0.09$, $n = 200$, $m = 50$

resulting in 2^m possible sample completions in each case. Missing data were generated noninformatively. Notice that, from (6.8), we have in this case

$$\sum_{l=1}^{m} K_l(s_0) = \sum_{l=1}^{m} \log\{e^{s_0 c_l} p_l + (1 - p_l)\}$$

	True value	MAR estimates (SE)	Range over $\mathcal{M}$	Proportion of positive estimates: Exact (enumer.)	Proportion of positive estimates: EAT algorithm with SP approx.
(i)	0.010	-0.0001(0.065)	[-0.082,0.113]	0.886	0.887
(ii)	0.005	-0.009(0.022)	[-0.072,0.049]	N/A	0.120
(iii)	0.050	0.031(0.022)	[-0.041,0.082]	N/A	0.955
(iv)	0.090	0.109(0.027)	[-0.002,0.149]	N/A	0.999

Table 6.1: Minima, maxima and proportion of positive estimates for β_1 in (6.9) considering all possible sample completions for the simulated data sets: (i) 50 observations and 12 missing data; (ii) to (iv) 200 observations and 50 missing data. All p_l in (6.8) are fixed at 0.5. Enumeration of all possible estimates is not feasible for (ii) to (iv).

$$\sum_{l=1}^{m} K_l'(s_0) = \sum_{l=1}^{m} \frac{p_l c_l}{\{p_l + (1-p_l)e^{-s_0 c_l}\}}$$
$$\sum_{l=1}^{m} K_l''(s_0) = \sum_{l=1}^{m} \frac{p_l c_l^2 (1-p_l)e^{-s_0 c_l}}{\{p_l + (1-p_l)e^{-s_0 c_l}\}^2}$$

where s_0 is the saddlepoint solution of the equation $\sum_{l=1}^{m} K' = T^{(k)}$.

The results for β_1 are reported in Table 6.1. In all cases, the probabilities in (6.8) for the missing observations are fixed and equal to 0.5.

From Table 6.1, it can be seen that the Fisher scoring algorithm with nested saddlepoint approximation gives results which agree closely with the exact solutions.

It is interesting to see that, though the interval given by the minimum and maximum estimates can appear to be, for some simulated data, rather symmetric about the threshold considered, the actual proportion of possible estimates above and below the threshold can be quite different.

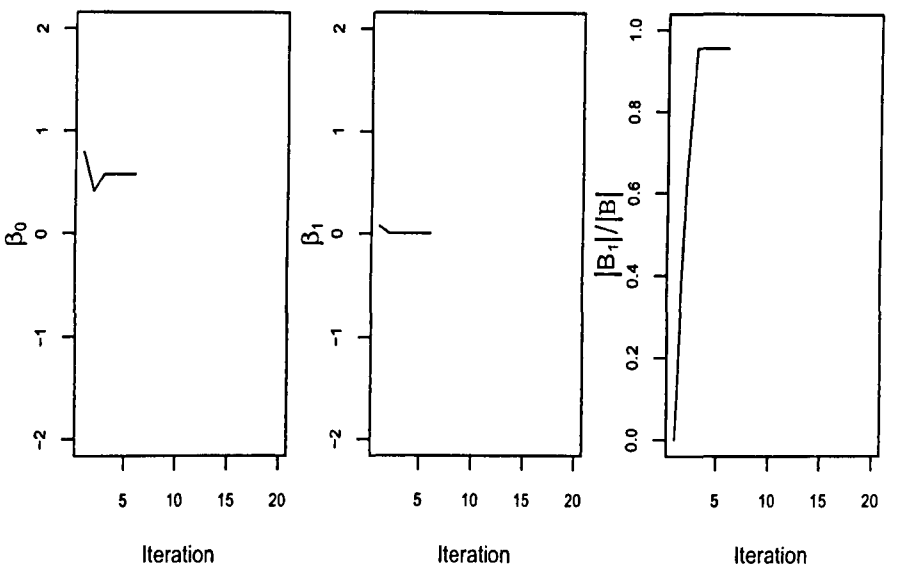

Figure 6.1: Sequences of parameter estimates form the EAT algorithm with saddlepoint approximation for simulated data (iii). The last plot refers to the proportion of positive estimates of β_1 in (6.9) over all possible data completions in $\mathcal{M}$.

Looking then at the results from the larger simulated data sets (ii)-(iv), a clear pattern emerges: as the true value of β_1 increases leading to statistically more significant MAR point estimates, the range of possible estimates over $\mathcal{M}$ includes more positive values and the proportion of positive estimates calculated using the EAT algorithm increases.

Figure 6.1 shows sequences of parameter estimates for model (6.9) using the EAT algorithm with nested saddplepoint approximation, for the second simulated data set in Table 6.1. The converging sequence for the proportion of positive estimates for β_1 is also plotted; in general, convergence takes place in less than ten iterations. CPU times to calculate $|\mathbf{B}_j|/|\mathcal{B}|$ for the smaller simulated data set (i) were 53.96 and 0.14 seconds when enumerating all possible solutions or using the saddlepoint approximation, respectively, using R version 1.5.1 (Ihaka and Gentleman, 1996) on a Sun Ultrasparc II workstation.

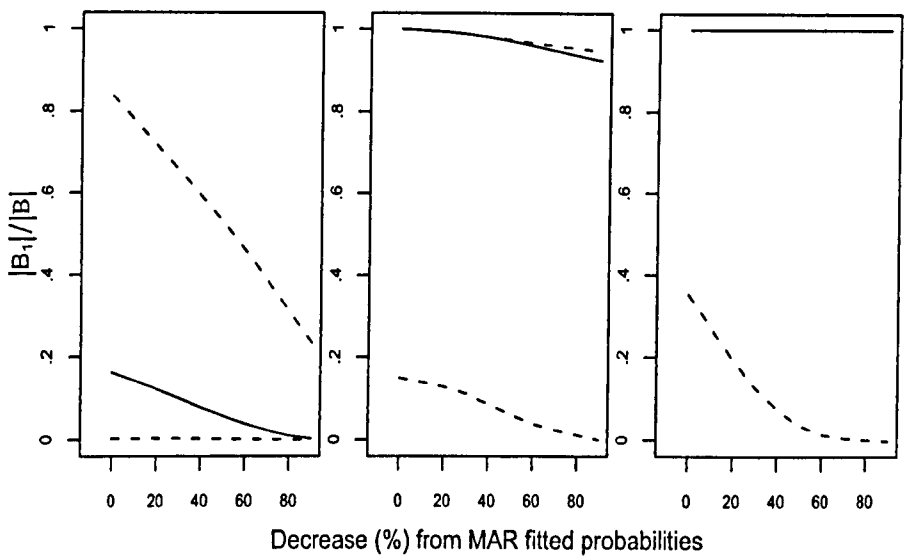

Figure 6.2: Weighted proportions (solid lines) of estimates of β_1 in (6.9) that are greater than zero over all possible sample completions when the values of p_l in (6.8) vary in the range shown on the x-axis. Dotted lines represent 95% bootstrap confidence intervals. Simulated data (ii) to (iv) are shown from left to right.

Next we consider a sensitivity analysis using data sets (ii) to (iv) by varying the values of p_l in (6.8). In particular, we start by assigning to missing observations the probabilities predicted from fitting model (6.9) to the observed data and then consider scenarios where these probabilities are gradually reduced. Results are shown in Figure 6.2 where the solid lines represent the weighted proportions corresponding to $\mathbf{B}_j$ and the dotted lines indicate 95% bootstrap confidence limits.

The latter have been calculated using the bias corrected and accelerated method described, for example, in Carpenter and Bithell (2000); in particular, we sample with replacement from the original simulated data sets samples of size $(n-m)$ and, for each of the bootstrapped sample, calculate the required weighted proportion using the EAT algorithm where the p_l in (6.8) for the missing observa-

tions are obtained from the fit of the MAR model on each bootstrapped sample. From Figure 6.2 and Table 6.1 it can be seen that, as the statistical significance of the MAR point estimates increases, the uncertainty about the sign of point estimates corresponding to possible completions decreases with bootstrap CI for situation (iv) that do not come close to zero unless a very sharp decrease is allowed from the MAR predicted probabilities for the missing scores.

6.4.2 Poisson regression

The same methods were applied to simulated Poisson count data. In this case however, it is necessary to specify an upper limit for the value that the response variable can take. In practical situations, this should not constitute a problem as a sensible upper bracket could be defined based, for example, on the empirical distribution function of the data at hand.

We simulated 300 count data from the following model

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, 300 \tag{6.10}$$

with true values $\beta_0 = 0.1$, $\beta_1 = 0.3$ and $x_i \sim N(30, 1)$.

$m = 8$ missing data were generated noninformatively and, for simplicity, allowed to take values in $\{1, 2, 3\}$ giving rise to $3^8 = 6561$ possible ways of completing the sample. In this case we have $Z_l \in \mathcal{D} = \{c_l, 2c_l, 3c_l\}$ and, placing equal mass on each possible outcome i.e. assuming each sample completion is equally probable, $P(Z_l) \equiv 1/3$, $l = 1, \ldots, 8$. The cumulant generating function and its first and second derivative can be readily derived and have expression similar to those in the previous Subsection.

Results for parameter β_1 in (6.10) are given in Table 6.2. Various threshold values

	Proportion of positive estimates	
Threshold	*Exact*	*EAT algorithm*
	(enumer.)	*with SP approx.*
0.305	0.0014	0.0013
0.290	0.2336	0.2316
0.270	0.9529	0.9522
0.260	0.9998	0.9998

Table 6.2: Proportion of positive estimates for β_1 in (6.10), for various thresholds, over all possible sample completions for the simulated data.

have been considered within the range of possible estimates which was $[0.259, 0.308]$. In all cases, the modified Fisher scoring algorithm with nested saddlepoint approximation yielded results that were very similar to those obtained by enumerating all possible solutions, even when considering extreme thresholds. Also, CPU times did not exceed 10 seconds using the EAT algorithm compared to about 300 seconds required by the exact solution.

6.5 Application to the dental pain trial

We return now to the dental pain trial described in Subsection 2.1.2. As mentioned before, here we use a dichotomized version of the original ordinal outcome grouping the first three and last two categories. Also, in what follows we ignore patients in either the positive control or highest dose level group and focus on the investigators' primary endpoint which was pain relief at 8 hours.

Assume that the dichotomized scores at 8 hours since randomisation, $\mathbf{Y} = (Y_1, \ldots, Y_n)$, follow a Bernoulli distribution; the following model relates the *complete* intended vector of observations $\mathbf{y}$ to some covariates of interest including treatment arm and patient's weight and age

Parameter	MAR est. (SE)	Range over $\mathcal{M}$	Proportion of positive estimates EAT algorithm with saddlepoint approx.	Bootstrap 95%CI
β_1	-1.09(1.18)	[-5.76,4.43]	0.637	[0.267,0.893]
β_2	-0.56(1.19)	[-4.51,4.75]	0.918	[0.562,0.987]
β_3	0.30(1.23)	[-4.10,5.14]	0.995	[0.948,0.999]
β_4	0.20(1.23)	[-4.34,5.49]	0.990	[0.890,0.999]

Table 6.3: Minima, maxima and proportion of positive estimates for the treatment contrasts in (6.11) using the methods described in Section 6.3. All p_l in (6.8) are fixed at 0.5.

$$\text{logit}\{\Pr(y_i = 1)\} = \beta_0 + \beta_g I(\text{group}_i = g) + \beta_6 \text{weight}_i + \beta_7 \text{age}_i \qquad (6.11)$$

where $g = 1, \ldots, 4$ indexes treatment contrasts with placebo and $i = 1, \ldots, n$. The focus here is on the treatment contrasts β_g, $g = 1, \ldots, 4$ between the active dose groups and placebo at 8 hours. Parameter estimates and associated standard errors for model (6.11) fitted to available cases are reported in the second column of Table 6.3. Thus, using a dichotomized outcome, we found no statistically significant beneficial effect of the test drug at any dose level compared to the placebo. However, these results are affected by the large number of missing observations, especially in the placebo and lower dose level groups; it is therefore interesting to assess the extent of the uncertainty induced by the large number of missing data using the methods described in the previous Section.

The first step is to obtain maxima and minima for all contrasts over all possible data completions. Results are shown in the third column of Table 6.3. The very wide intervals reflect the high uncertainty due to the many missing data and this is particularly true for the treatment difference between placebo and the lowest dose group (third column of Table 6.3). Nevertheless, when considering

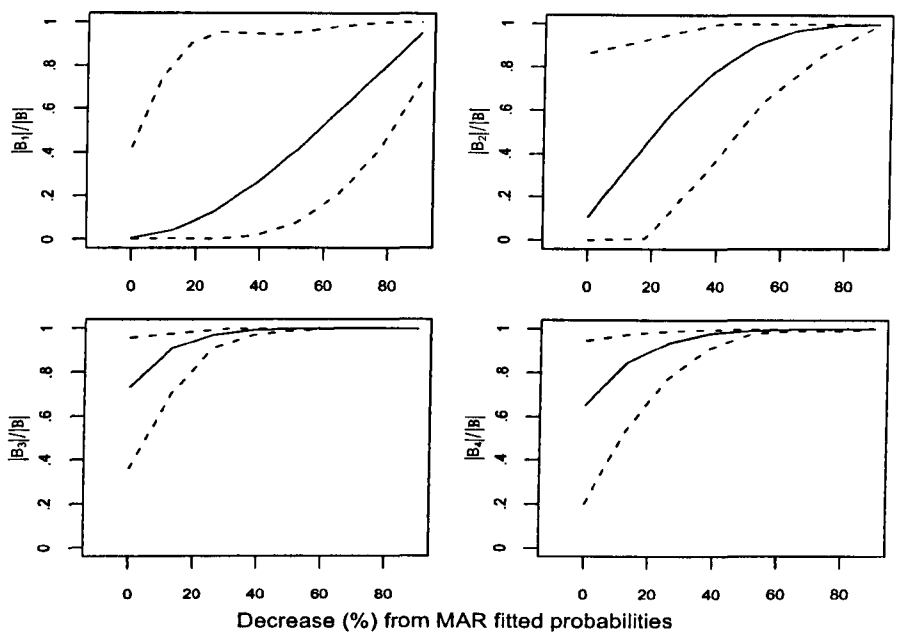

Figure 6.3: Weighted proportions (solid lines) of estimates of β_g in (6.11), $g = 1, \ldots, 4$, that are greater than zero over all possible sample completions when the values of p_l in (6.8) vary in the range shown on the x-axis. Dotted lines represent 95% bootstrap confidence intervals.

the actual proportions of treatment contrasts that remain positive, we are still able to draw useful conclusions.

Under an uncertain scenario where all p_l for the missing observations in (6.8) are fixed at 0.5 (thus all sample completions are equally probable), the proportions of treatment contrasts which remain positive over all possible estimates are all greater than 0.9 apart from the lowest dose group. None of the 95% bootstrap CI for the proportions covers zero and the wider CI for β_1 reflects the higher uncertainty due to the larger number of missing data in dose group 1.

A slightly different picture emerges if we assume that all p_l are equal to the corresponding predicted quantities obtained from the MAR model fitted to the

available cases. In particular we have performed a sensitivity analysis using the same approach described in Section 6.4 for the simulated binary data. Results are plotted in Figure 6.3.

It can be seen that, were we to use the predicted probabilities from the MAR model for the missing observations, this would result in a statistically significant (weighted) proportion of positive treatment contrasts (over all enumerable estimates) only in the higher dose groups corresponding to Test drug 3 and 4 which, incidentally, have positive MAR point estimates. However, all proportions become significantly greater than zero when considering lower values for these probabilities. It is in fact sensible to assume that for this particular data set, if the dropout mechanism was truly nonignorable, these probabilities would be lower than the corresponding ones from a MAR model. This is because the probability of an improvement in pain score would be reduced if a patient missed the scheduled observation.

In conclusion, there appears to be little uncertainty about the efficacy of the test drug at dose level 3 and 4. On the other hand, less definitive conclusions can be drawn about the true efficacy of the test drug at dose level 1 and 2. This is because of the larger number of missing data in these groups together with negative values for the corresponding MAR point estimates of the constrasts with placebo.

6.6 Importance sampling of missing sufficient statistics and approximate conditional inference

In this Section, we quantify the uncertainty about parameter estimates caused by the presence of missing data using a different approach based on exact conditional inference.

There is an extensive literature on methods for exact conditional inference in generalized linear models. With conditional testing, the statistical significance of a parameter of interest is assessed considering tail probabilities of the conditional distribution of the corresponding sufficient statistic given the value of the sufficient statistics associated with nuisance parameters. Davison (1988) describes an approach based on double saddlepoint approximations to the (univariate) conditional distribution of the sufficient statistic for a coefficient of interest. Extensions to the case where hypothesis testing involves more than one parameter have been considered by Kolassa and Tanner (1999); they couple the double saddlepoint approximation with noniterative Monte Carlo methods to obtain a sample from the joint conditional distribution of the sufficient statistics for the parameters of interest. In previous work, the same authors (1994) considered a Gibbs sampling approach to generating a Markov chain converging to the desired conditional multivariate density. Other methods which do not use saddlepoint approximations but rather develop efficient algorithms for enumerating (or sampling) from the reference set of all possible permutations (giving rise to the observed values of the sufficient statistics for the nuisance parameters) have been considered by Tritchler (1984), Pagano and Tritchler (1983), Hirji et al. (1988) and Mehta et al. (2000). A review of the various approaches in the context of exact conditional logistic regression is given in Corcoran et al. (2001).

6.6.1 Formulation

Consider the vector of independent binary outcomes $\mathbf{Y} = (Y_1, \ldots, Y_n)$ following a Bernoulli distribution and assume a logistic regression relates the log-odds of response to a set of p, possibly subject-specific covariates

$$\text{logit}\{\Pr(y_i = 1)\} = \sum_{j=0}^{p} x_{ij}\beta_j = \mathbf{x}_i'\boldsymbol{\beta} \tag{6.12}$$

with β_0 the intercept term. We write the $(p+1)$-dimensional vector of sufficient statistics for $\boldsymbol{\beta}$ as

$$\mathbf{T} = (T_0, T_1, T_2, \ldots, T_p)' = \mathbf{X}'\mathbf{Y}$$

where $\mathbf{X}$ is the $n \times (p+1)$ matrix with i-th row $\mathbf{x}_i$ and the corresponding vector of observed values of $\mathbf{T}$ is $\mathbf{t} = (t_0, t_1, t_2, \ldots, t_p)$. Suppose now that we are interested in making inferences about β_1 and treat the remaining parameters as nuisance parameters. Conditional inference then requires finding the conditional probability mass function $P(T_1 = t_1 | T_0 = t_0, T_2 = t_2, \ldots, T_p = t_p) = P(T_1 = t_1 | \mathbf{T}_{-1} = \mathbf{t}_{-1})$ under the null hypothesis of interest (say $H_0 : \beta_1 = 0$). In particular an exact level of significance is given by the tail probability

$$P = \sum_{t_1^* > t_1} P(T_1 = t_1^* | \mathbf{T}_{-1} = \mathbf{t}_{-1}). \tag{6.13}$$

Consider now the joint cumulant generating function of the vector $\mathbf{T}$

$$\ln\Big\{E\big[\exp(\mathbf{T}'\mathbf{u})\big]\Big\} = \sum_{i=1}^{n} \ln E\Big[\exp(Y_i\mathbf{x}_i'\mathbf{u})\Big] = \sum_{i=1}^{n} K_i(\mathbf{u}). \tag{6.14}$$

A saddlepoint approximation to the joint density of $\mathbf{T}$ at $\mathbf{t}$ is given by a multivariate version of (6.1)

$$(2\pi)^{-((p+1)/2)}|I(\mathbf{u}^*)|^{-1/2}\exp\left[\sum_{i=1}^{n} K_i(\mathbf{u}^*) - \mathbf{t}'\mathbf{u}^*\right] \tag{6.15}$$

where the saddlepoint $\mathbf{u}^*$ is now the root of the system of equations

$$\sum_{i=1}^{n} \partial K_i(\mathbf{u}^*)/\partial \mathbf{u}^* = \mathbf{t}$$

and $I(\mathbf{u}^*)$ is the $(p+1) \times (p+1)$ Hessian matrix $\sum_{i=1}^{n} \partial^2 K_i(\mathbf{u})\partial \mathbf{u}\partial \mathbf{u}'$ evaluated at $\mathbf{u}^*$.

Since the density in (6.13) can be written as the ratio of the marginal densities of $\mathbf{T}$ and $\mathbf{T}_{-1}$, the double saddlepoint approximation to the conditional density is given by the ratio of the corresponding joint saddlepoint approximations where the latter for $\mathbf{T}_{-1}$ replaces u_1 with zero throughout.

Tail probabilities are then obtained using the Skovgaard (1987) formula adapted to the double saddlepoint case

$$P_{SP} = 1 - \Phi(A) - \phi(A)(1/A - 1/B) \tag{6.16}$$

where in the previous expression

$$A = \text{sign}(u_1^*)\left\{2\left[\sum_{i=1}^{n} K_i(\mathbf{u}_{-1}^*) - \sum_{i=1}^{n} K_i(\mathbf{u}^*) + \mathbf{t}'\mathbf{u}^* - \mathbf{t}'_{-1}\mathbf{u}_{-1}^*\right]\right\}$$

and

$$B = u_1^*\left[|I(\mathbf{u}^*)|/|I(\mathbf{u}_{-1}^*)|\right]^{1/2}$$

with Φ and ϕ the standard normal distribution and density function respectively.

When some of the intended measurements are missing, we partition the vector $\mathbf{Y}$ into the observed and missing part as $(\mathbf{Y}^{obs}, \mathbf{Y}^{miss})$ where $\mathbf{Y}^{miss}$ is of size m, the number of missing data. Similarly, we partition the vector of sufficient statistics as $\mathbf{T} = \mathbf{T}^{obs} + \mathbf{T}^{miss} = (T_0^{obs} + T_0^{miss}, T_1^{obs} + T_1^{miss}, T_2^{obs} + T_2^{miss}, \ldots, T_p^{obs} +$

T_p^{miss}). Now, the expected value of (6.13) over the distribution of the missing data is arguably a sensible approach to summarizing the extra uncertainty due to unplanned missingness. In particular, let us assume that possible outcomes for the missing measurements are equally likely i.e. $\Pr(Y_j^{miss} = 0) = \Pr(Y_j^{miss} = 1) = 0.5$ for all $j = 1, \ldots, m$. We can write the expected value thus defined as

$$E(P) = \sum_{t^{miss}} \sum_{\substack{t_1^{*obs}+t_1^{*miss} \geq \\ t_1^{obs}+t_1^{miss}}} P(T_1^{obs} + T_1^{miss} = t_1^{*obs} + t_1^{*miss} | \mathbf{T}_{-1}^{obs} + \mathbf{T}_{-1}^{miss} =$$

$$= \mathbf{t}_{-1}^{obs} + \mathbf{t}_{-1}^{miss}, \mathbf{T}^{miss} = \mathbf{t}^{miss}) \times P(\mathbf{T}^{miss} = \mathbf{t}^{miss}) \quad (6.17)$$

where the outer summation is over the set of possible outcomes of the missing components of the sufficient statistics.

We propose a computationally efficient approach to evaluating a Monte-Carlo approximation of (6.17) using importance sampling. Candidate values of $\mathbf{t}^{miss}$ are sampled from a multivariate normal approximation to the density function of the vector $\mathbf{T}^{miss}$ with mean $\mu_{T^{miss}}$ and, covariance matrix $\Sigma_{T^{miss}}$ the latter having generic entry

$$\mathrm{Cov}(T_s^{miss}, T_t^{miss}) = \sum_{j=1}^{m} x_{js} x_{jt} \mathrm{Var}(Y_j^{miss})$$

for all $s, t \in \{1, \ldots, p\}$ with $\mathrm{Var}(Y_j^{miss}) = 0.25$, $j = 1, \ldots, m$.
For each sampled vector $\mathbf{t}^{miss(k)}$, weights are calculated as

$$w^{(k)} = \frac{(2\pi)^{-((p+1)/2)} |I(\mathbf{u}^*)|^{-1/2} \exp\left[\sum_{j=1}^{m} K_j(\mathbf{u}^*) - \mathbf{t}^{miss(k)} \mathbf{u}^*\right]}{(2\pi)^{-(p+1)/2} |\Sigma_{T^{miss}}|^{-1/2} \exp\left[-1/2 (\mathbf{t}^{miss(k)} - \mu_{T^{miss}})' \Sigma_{T^{miss}}{}^{-1} (\mathbf{t}^{miss(k)} - \mu_{T^{miss}})\right]} \quad (6.18)$$

the ratio of the saddlepoint approximation (6.15) of the target density evaluated at $\mathbf{t}^{miss(k)}$ to the importance density, where $\mathbf{u}^*$ now solves $\sum_{j=1}^{m} \partial K_j(\mathbf{u}^*)/\partial \mathbf{u} = \mathbf{t}^{miss(k)}$ and, in our case $K_j(\mathbf{u}^*) = \ln\left[\exp(\mathbf{x}_j' \mathbf{u}^*) 0.5 + 0.5\right]$. In order for the proposal

density function to have the same support as the target density, sample values are rounded to the nearest integer and values outside the possible ranges for the missing sufficient statistics discarded.

As well as calculating the weights (6.18), pseudo-complete sufficient statistics $t^{(k)} = (\mathbf{t}^{obs} + \mathbf{t}^{miss(k)})$ are used in (6.16) to obtain the corresponding value of the tail probability. Also, a continuity correction to (6.16) replaces $u_1 = u_1 - .5$ throughout, and B with

$$B = 2\sinh(u_1^*)\left[|I(\mathbf{u}^*)|/|I(\mathbf{u}_{-1}{}^*)|\right]^{1/2}.$$

An importance sampling Monte Carlo evaluation of (6.17) with nested double saddlepoint approximation is then obtained as

$$\overline{P}_{ISMC} = \frac{\sum_{k=1}^{K}\left[\Phi(z^{(k)}) + \phi(z^{(k)})(1/z^{(k)} - 1/\zeta^{(k)})\right]w^{(k)}}{\sum_{k=1}^{K} w^{(k)}} = \frac{\sum_{k=1}^{K} w^{(k)} P_{SP}^{(k)}}{\sum_{k=1}^{k} w^{(k)}}. \tag{6.19}$$

6.6.2 A simulation study

To evaluate the proposed method, we simulated data from the following logistic regression model with two parameters and a single binary covariate flagging, for instance, treatment allocation in a placebo controlled trial

$$\text{logit}\{\Pr(y_i = 1)\} = \beta_0 + \beta_1 x_{i1}. \tag{6.20}$$

A total sample size of $n = 85$ was considered with varying number of missing observations m_1 and m_2 in the two groups generated noninformatively, with $m_1 + m_2 = 15$ in all cases. In particular, Table 6.4 shows results from the following artificial data sets

(i) $n_1 = 35$, $n_2 = 35$, $m_1 = 13$, $m_2 = 2$

(ii) $n_1 = 35$, $n_2 = 35$, $m_1 = 7$, $m_2 = 8$

(iii) $n_1 = 35$, $n_2 = 35$, $m_1 = 2$, $m_2 = 13$.

For these small data sets, exact results can be obtained. Namely, for each of the 2^{15} possible sample completions, tail probabilities under the null hypothesis of no treatment effect are calculated using Fisher's exact test and their average value computed ($\overline{P}_E$). We then compare the exact results with a crude Monte Carlo approximation to (6.17), $\overline{P}_{MC}$. This is obtained by sampling data completions at random from all possible 2^{15} values that $\mathbf{Y}^{miss}$ can take and the computing the tail probabilities for each pseudo-complete data using the double saddlepoint approximation (6.16). Finally, results from the proposed approach which uses a Monte Carlo approximation with importance sampling of the missing part of the sufficient statistics are also shown ($\overline{P}_{ISMC}$).

In all cases, Monte Carlo with importance sampling gives a better approximation compared to plain Monte Carlo.

For the artificial data considered here, the estimated value of β_1 using data from the available cases only was equal to -1.47 corresponding to an odds ratio of 0.23 (one sided p-value 0.0004). Since we have assumed that outcomes for the missing observations are equally likely, we see that, as the number of missing observations decreases in group 2 and increases in group 1 (from scenario (iii) to scenario (i)), the statistical significance associated with β_1 decreases as the odds ratio moves towards the null hypothesis.

The opposite is true when considering scenarios (i) to (iii): with more missing data in group 2, the expected p-value for β_1 gets closer to the p-value obtained from the analysis on available cases and moves away from the null.

Monte Carlo approximation errors were assessed using arguments similar to those in Booth and Butler (1999). Indicating with $\widetilde{P}$ the (true) expected value

of (6.16) calculated over all possible 2^m data completions, an approximation to $\mathrm{Var}(\overline{P}_{ISMC})$ based on K pseudo-complete datasets is given by

$$\hat{\sigma}^2_{ISMC} = \frac{1}{\bar{w}^2}\frac{1}{K}\sum_{k=1}^{K}(w^{(k)}P^{(k)}_{SP} - w^{(k)}\widetilde{P})^2 \tag{6.21}$$

and the corresponding approximation to $\mathrm{Var}(\overline{P}_{MC})$ is

$$\hat{\sigma}^2_{MC} = \frac{\frac{1}{K}\sum_{k=1}^{K}(P^{(k)}_{SP} - \widetilde{P})^2}{K}. \tag{6.22}$$

In each case we then calculated the number of sampled values K such that

$$|z_{\alpha/2}|\hat{\sigma}/\sqrt{K} \leq \varepsilon \tag{6.23}$$

where $\varepsilon = 0.001$ and z_α is the $(1-\alpha)$ quantile of the standard normal distribution.

Results are shown in Table 6.4. For the MC and importance sampling MC methods, the values shown refer to averages over 50 separate runs, where the number of sampled values in each run is determined by the value of K satisfying (6.23). In particular, notice how the average value of K is much smaller for importance sampling compared to plain Monte Carlo.

6.6.3 Application to the dental pain trial

In the application to the dental pain trial, the log-odds of improvement in pain score is related to subject-specific covariates using the model

$$\mathrm{logit}\{\Pr(y_i = 1)\} = \beta_0 + \beta_g I(\mathrm{group}_i = g) + \beta_6 \mathrm{weight}_i + \beta_7 \mathrm{age}_i \tag{6.24}$$

Scenario		*Enumeration*	*MC*	*ISMC*
	$\overline{P}$	0.0166	0.0151	0.0168
(i)	CPU	135	169	71
	K	—	2914	528
	$\overline{P}$	0.0127	0.0095	0.0125
(ii)	CPU	125	132	56
	K	—	2320	504
	$\overline{P}$	0.0078	0.0028	0.0078
(iii)	CPU	110	55	19
	K	—	1016	202

Table 6.4: Exact (enumeration) and Monte Carlo and importance sampling Monte Carlo approximations of expected one-sided p-values for β_1 in (6.20). In the latter two cases, all values reported refer to average values over 50 separate runs, the number of sampled values in each run determined by K satisfying (6.22) and (6.21). CPU times (averages per run) are in seconds.

where $g = 1, \ldots, 4$ indexes treatment contrasts with placebo and $i = 1, \ldots, 261$, the number of intended measurements.

As shown in the previous Section, when fitting (6.24) to the observed data, none of the treatment differences between the active dose groups and placebo is found to be statistically significant (first two columns of Table 6.5). From a practitioner's point of view however, it would be interesting to assess the robustness of these results under different scenarios for the missing data.

Let us start by considering the case where outcomes for the missing observations are assumed equally likely, i.e. we consider the extreme scenario of the previous Sections first.

Arm	*MAR analysis*		$\overline{P}_{ISMC}$	
	Estimate	P-value	Extreme	MAR fitted prob.
Dose 1	-1.093	0.353	0.700	0.324
Dose 2	-0.561	0.638	0.145	0.178
Dose 3	0.308	0.802	0.018	0.018
Dose 4	0.205	0.868	0.037	0.030

Table 6.5: Estimates of β_g, $g = 1, \ldots, 4$ in (6.24) considering all available cases (MAR analysis) and corresponding estimates of the expected one-sided p-values over all possible sample completions using (6.19).

Results for treatment contrasts β_g, $g = 1, \ldots, 4$ under this scenario are shown in the third column of Table 6.5; in particular, in each case values refer to the importance sampling Monte Carlo approximations of the upper tails of the distribution of the relevant sufficient statistics for each treatment contrast conditioning on the others.

A clear pattern emerges as we move form the lowest dose group to the highest: because of the larger number of missing data in the placebo arm and dose groups 1 and 2, it is most unlikely that results from pseudo-complete datasets would actually show a significant effect of the testing drug at these dose levels.

Different conclusions apply to the higher dose levels for which the estimated expected upper tail probabilities show a statistically significant improvement in pain relief at the .05 level. In other words, considering the possible values of the missing part of the relevant sufficient statistics, the estimated odds ratios tend to move away from the null hypothesis of no treatment effect and towards the alternative of an increased efficacy of the test drug at these dose levels. This is less evident for test dose two while for test dose one the results strengthen the

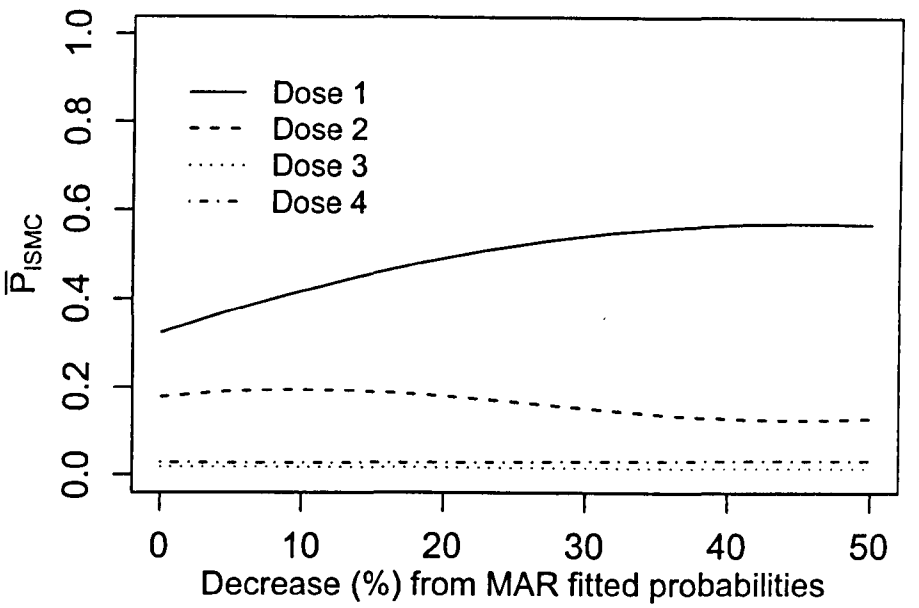

Figure 6.4: Estimates of the expected p-values corresponding to treatment contrasts in (6.24) over all possible possible sample completions assuming that the probabilities of pain relief among missing subjects vary in the range shown on the x-axis.

null hypothesis of no treatment effect. Therefore, even under a most uncertain scenario for the missing scores, results appear to support the efficacy of the testing drug at dose levels 3 and 4 after accounting for the extra uncertainty due to the large number of missing data.

As with the methods of the previous sections, a feature of the proposed approach is that sensitivity analyses can be readily performed by varying the probabilities assumed for the missing observations. We may again consider a sensitivity analysis where we assign to missing observations the predicted probabilities from the fit of (6.24) to the available cases and then gradually decrease these values. Recall in fact that, here, the missing data mechanism is likely to be nonignorable and is therefore sensible to assume a smaller probability of improvement among noncompleters compared to the value predicted from the analysis of available cases. For each treatment contrast $\beta_g, g = 1, \ldots, 4$ results are show in Figure 6.4.

The leftmost values are also given in the fourth column of Table 6.5. It is important to notice that, in this case, as we are implicitly imputing missing data using the fitted probabilities from a MAR model, the imbalance in the number of missing data in the different arms leads to increased confidence in the efficacy of the experimental treatment. This is more evident for dose groups 3 and 4. Furthermore, in order to take into account the statistical uncertainty about the MAR fitted probabilities, we could construct bootstrap confidence intervals for the expected p-values in the spirit of the methods of the previous Subsection. These confidence intervals are likely to contain the MAR p-values.

6.7 Discussion

When data are missing on a discrete outcome related to some covariate of interest using, for example, logistic regression, the set of parameter estimates corresponding to all possible sample completions is finite. However, in many cases, just looking at the range of possible estimates does not give a satisfactory measure of ignorance in the sense of Vansteelandt and Goetghebeur (2001); for example, if we are interested in a particular threshold value, the interval may appear symmetric about the critical value when in fact the proportion of possible estimates (over all possible sample completions) either side of it may be quite different.

Assuming that all possible sample completions are equally probable, the methods described in Sections 2 and 3 allow an accurate and fast computation of the proportion of parameter estimates above or below the chosen threshold. Weighted proportions are obtained instead if we assume that sample completions are not all equally plausible and use, for instance, the predicted probabilities from a model fitted to the observed data for the values in (6.8) corresponding to

the unseen observations. Useful conclusions can then be drawn even when the number of missing observations is very large as in the dental pain trial.
Also, a sensitivity analysis can be performed by varying the fitted probabilities for the unseen measurements and allowing for different scenarios. In the dental pain trial for instance, it was sensible to consider scenarios where the probabilities of improvement for the missing observations are lower than their predicted MAR counterpart. One can then assess the robustness of the MAR results obtained from fitting the model to the available cases only and their sensitivity under clinically plausible scenarios. Results will in any case depend on the number of missing observations (possibly within each treatment arm as in our case), and their statistical significance as we have seen in the simulation studies of Section 6.4. Further, familiar sampling variability can be taken into account by constructing bootstrap confidence intervals for these proportions.

Our approach could be extended to longitudinal discrete data modelled using, for example, the marginal model of Section 5.1. The nested saddlepoint approximation would only require calculation of the cumulant distribution function of the discrete outcome as we did with Poisson regression in Section 6.4. Bootstrap confidence intervals for the proportions could then be constructed taking into account the longitudinal nature of the data, although interpretation of the results would be less straightforward than in the independent data setting.
Finally, from the practitioners' point of view, the proposed approach is particularly attractive as it provides an intuitive estimate of the extent of the uncertainty about a particular coefficient of interest caused by missing data, making minimal or no assumptions about the mechanism driving the missing data process.

Similar considerations apply to the methods of Section 6. There, we considered the set of levels of significance over all possible pseudo-complete data and the resulting distribution of p-values which could be summarized, for example, considering the mean or quantiles (Delucchi, 1994). However, enumeration of

all possible permutations is infeasible with only a moderate number of missing data. In the context of incomplete binary data, we have shown how the 'expected' exact one-sided p-value can be approximated coupling methods used for conditional inference in generalized linear models based on double-saddlepoint techniques with importance sampling Monte Carlo integration. In particular, samples of the missing part of the sufficient statistics are obtained from an approximating multivariate normal distribution and then used to calculate an importance sampling Monte Carlo approximation to the exact tail probability of the parameter of interest with weights given by the ratio of the saddlepoint approximation to their joint distribution and the proposal multivariate normal density. Results obtained are similar to those from plain Monte Carlo integration but require a smaller number of sampled values and attain a better accuracy (see Table 6.4). In this case too, sensitivity analyses can be readily performed considering different scenarios for the missing data.

The advantage of this method over those of the previous Chapter and Section 6.1 is that it deals directly with the uncertainty of usual levels of significance caused by missing data overcoming subtle interpretation issues related to the intervals of uncertainty and the bootstrap confidence intervals for proportions of estimates above a threshold. Of course, this approach is particularly suited in cases where conditional inference itself is recommended, namely, in the presence of sparseness or separability in the data (Albert and Anderson, 1984; Santner and Duffy, 1986) and when focus is on few parameters in a model containing a large number of parameters with respect to the number of observations (for instance when adjusting for various possible confounders in epidemiological studies). The proposed approach is valid with any generalized linear models as long as canonical link functions are used; in all these cases in fact, both expressions (6.17) and (6.18) are still valid. Extension to longitudinal data models in general and

random effect models in particular are not straightforward as conditional inference itself is more questionable in this setting. For instance, one cannot treat the variance components in a random effect model as truly nuisance parameters since in most cases they are as relevant as the parameters in the fixed part of the model.

Chapter 7

Conclusions

In this thesis we have presented various approaches to the analysis of incomplete data, focusing in particular on incomplete longitudinal data. Broadly speaking, four non-exclusive methods have been described. These are: the Monte Carlo EM algorithm to fit non-ignorable random-coefficient-based dropout models; sensitivity analyses based on local influence and sampling-based methods; intervals of ignorance and uncertainty for the parameters of marginal models for categorical data, and the method based on calculation of the proportion of possible estimates above (or below) a critical threshold. Common to all approaches is the need to make assumptions about the mechanism driving missingness in order to proceed. These are stronger for the methods of Chapters 3 and 4 compared to the methods described in later chapters. The underlying truth is that part of the intended set of observations is missing, and in all cases we are implicitly or explicitly imputing the missing data. Therefore efforts should be made to obtain as complete a data set as possible; failing to do this, it would be useful to obtain as much information as possible about the missing data mechanism, for example by following some of the dropouts in a longitudinal study after they have exited that study. This knowledge can then be used to assess the need for modelling the dropout mechanism and possibly to define an appropriate model for the latter.

From a computational point of view, the methods presented may appear rather ad hoc; however, as discussed in the introduction, this is almost unavoidable, since different missing data mechanisms need tailored approaches. Nevertheless, the methods based on intervals of ignorance and uncertainty described in Chapter 5 are at the same time easy to implement and very flexible, as they can be applied to continuous and discrete outcomes, leading to conclusions that can have a very practical interpretation, for example if the statistical significance of a treatment effect is preserved under what we have referred to as an extreme scenario for the missing data. Furthermore, various authors have recently presented convincing evidence in favour of the Monte Carlo EM algorithm for fitting generalized linear mixed models (Booth et al., 2001; Booth and Hobert, 1999; Ibrahim et al., 2001; Palmgren and Ripatti, 2002); thus, if this were to become the mainstream approach for fitting GLMM then our MCEM for NIRCB models could find wider applications. An important alternative approach that can be used to fit outcome-based or random-coefficient-based missing data models is MCMC based Bayesian methods (see Carpenter et al. (2002) for outcome-based models). However, a Bayesian approach does not exempt us from the need to make untestable assumptions as, for example, prior distributions for the missing data have to be made; on the other hand, sensitivity analyses can be readily performed by varying these priors, using, for instance, heavy-tailed distributions.

7.1 Further areas of research

The MCEM of Chapter 3 can be used in more complex settings, such as multilevel models with more than two levels of clustering. In fact, it is not difficult to envisage situations where missingness may be related to the deviations, at any level of clustering, from the population average, just as in Parkinson's disease it was related to a patient's deviation from the overall rate of change. There-

fore an important area of research is the assessment of the performance of the method presented in these situations, both in computational terms and in terms of the interpretability of the results. It would also be interesting to compare the performance of the MCEM algorithm to the Stochastic EM or other methods based, for instance, on Stochastic Approximation (Booth et al., 2001). Another issue that should be investigated is the robustness of the results from random-coefficient-based dropout models to the assumption of normality of the random effects. Finally, it is straightforward to extend this method to categorical outcome data for which, in fact, the MCEM algorithm was originally proposed. Indeed, we could have used this method to account for missing data in the dental pain trial; however, because of the very short follow-up period, it would not have been sensible to relate dropout to some underlying rate of change in pain. Notice that all these possible developments carry through to the sensitivity analysis methods of Chapter 4, the validity of which, in these new settings, could also be studied.

The properties of the intervals of ignorance and uncertainty of Chapter 5 are another interesting area of research, especially with regard to the definition of coverage that should apply to the latter, as mentioned at the end of that Chapter. In general, the method is particularly suited to situations like the dental pain trial, where the outcome variable is ordered categorical, and sensitivity analyses like those of Subsection 5.3.3 lead to useful conclusions. Nevertheless, further applications of this approach to the continuous outcome setting would show its potential as a practical analysis tool. There are also parallels with fuzzy regression where one defines *fuzzy membership functions* for the outcome variable and obtains fuzzy (interval) estimates of the relevant parameter (Kacprzyk and Fedrizzi, 1992). Thus, the two approaches could be compared, at least in the case of i.i.d. continuous outcomes, where fuzzy membership functions would apply only to the missing data. Furthermore, possible extensions of fuzzy regression to

longitudinal data are another area of research.

The method based on calculation of the proportion of estimates exceeding a threshold over all possible estimates, as a measure of the uncertainty about traditional point estimates caused by missing data, could also be investigated further. In particular, an interesting alternative approach would be to consider Bayesian methods. One could then look at how the posterior distribution (tail probabilities) of a coefficient of interest are affected by the choice of domains and density functions for the missing observations, as we did in Section 6.5.

Appendix A

Calculation of standard errors for MCEM estimates

In the conditional parametric bootstrap approximation of Louis' formula (Diebolt and Ip in Gilks et al. (1996)) we draw a sample of size m from (3.7) using the methods of Section 3.2 with all relevant parameters fixed at the converging values of the MCEM algorithm and then approximate $E\left[-\ddot{\ell}_{compl}\right]$ and $\operatorname{cov}\left[\dot{\ell}_{compl}\right]$ via Monte Carlo integration. In particular, here $m = 500$.
Expressions for $\dot{\ell}_{compl}$:

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i}^{250} \sum_{j}^{n_i} \frac{1}{\sigma^2} \mathrm{A}_{ij} \qquad \frac{\partial \ell}{\partial \beta_1} = \sum_{i}^{250} \sum_{j}^{n_i} \frac{1}{\sigma^2} \mathrm{treat}_i \, \mathrm{A}_{ij}$$

$$\frac{\partial \ell}{\partial \beta_2} = \sum_{i}^{250} \sum_{j}^{n_i} \frac{1}{\sigma^2} \mathrm{week}_j \, \mathrm{A}_{ij} \qquad \frac{\partial \ell}{\partial \beta_3} = \sum_{i}^{250} \sum_{j}^{n_i} \frac{1}{\sigma^2} \mathrm{BaseADL}_i \, \mathrm{A}_{ij}$$

$$\frac{\partial \ell}{\partial \sigma^2} = \sum_{i}^{250} \sum_{j}^{n_i} -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \mathrm{A}_{ij}^2$$

$$\frac{\partial \ell}{\partial \alpha_0} = \sum_{i}^{250} \sum_{j}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} [R_{ij} \, \mathrm{B}_{il} + (1 - R_{ij}) \, \mathrm{C}_{il}]$$

$$\frac{\partial \ell}{\partial \alpha_1} = \sum_{i}^{250} \sum_{j}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} [R_{ij} \, \mathrm{treat}_i \, \mathrm{B}_{il} + (1 - R_{ij}) \, \mathrm{treat}_i \, \mathrm{C}_{il}]$$

$$\frac{\partial \ell}{\partial \alpha_2} = \sum_{i}^{250} \sum_{j}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} \left[R_{ij} \, b_{i1}^{(k)} \, \mathrm{B}_{il} + (1 - R_{ij}) \, b_{i1}^{(k)} \, \mathrm{C}_{il}\right]$$

$$\frac{\partial \ell}{\partial \alpha_3} = \sum_{i}^{250} \sum_{j}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} [R_{ij}\,\mathrm{BaseADL}_i\,\mathrm{B}_{il} + (1-R_{ij})\,\mathrm{BaseADL}_i\,\mathrm{C}_{il}]$$

$$\frac{\partial \ell}{\partial \gamma_l} = \sum_{i}^{250} \sum_{j}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} [R_{ij}\,\mathrm{I}(\mathrm{week}_j{=}\mathrm{l})\,\mathrm{B}_{il} + (1-R_{ij})\,\mathrm{I}(\mathrm{week}_j{=}\mathrm{l})\,\mathrm{C}_{il}]$$

$$\frac{\partial \ell}{\partial \sigma_1^2} = \sum_{i}^{250} -\frac{\sigma_1^2(\sigma_2^2)^2 - (\sigma_{1,2})^2\sigma_2^2 - b_{i0}^{(k)^2}(\sigma_2^2)^2 + 2b_{i0}^{(k)}b_{i1}^{(k)}\sigma_{1,2}\sigma_2^2 - b_{i1}^{(k)^2}(\sigma_{1,2})^2}{2(\sigma_1^2\sigma_2^2 - (\sigma_{1,2})^2)^2}$$

$$\frac{\partial \ell}{\partial \sigma_2^2} = \sum_{i}^{250} -\frac{(\sigma_1^2)^2\sigma_2^2 - \sigma_1^2(\sigma_{1,2})^2 + 2b_{i0}^{(k)}b_{i1}^{(k)}\sigma_{1,2}\sigma_1^2 - (b_{i1}^{(k)})^2(\sigma_1^2)^2 - (b_{i0}^{(k)})^2\sigma_{1,2}{}^2}{2(\sigma_1^2\sigma_2^2 - (\sigma_{1,2})^2)^2}$$

$$\frac{\partial \ell}{\partial \sigma_{1,2}} = \sum_{i}^{250} -\frac{-2\sigma_1^2\sigma_2^2\sigma_{1,2} + 2(\sigma_{1,2})^3 + 2(b_{i0}^{(k)})^2\sigma_{1,2}\sigma_2^2 - 2b_{i0}^{(k)}b_{i1}^{(k)}(\sigma_{1,2})^2 + 2(b_{i1}^{(k)})^2\sigma_1^2\sigma_{1,2} - 2b_{i0}^{(k)}b_{i1}^{(k)}\sigma_1^2\sigma_2^2}{2(\sigma_1^2\sigma_2^2 - (\sigma_{1,2})^2)^2}$$

where $\sigma_1^2 = \sigma_{int}^2$, $\sigma_2^2 = \sigma_{slope}^2$, $\sigma_{12} = \sigma_{int,slope}$ and

$$\mathrm{A}_{ij} = (y_{ij} - \beta_0 - \beta_1\mathrm{treat}_i - \beta_2\mathrm{week}_j - \beta_3\mathrm{BaseADL}_i)$$

$$\eta_{il} = \alpha_0 + \alpha_1\mathrm{treat}_i + \alpha_2 b_{i1}^{(k)} + \alpha_3\mathrm{BaseADL}_i + \gamma_l$$

$$\mathrm{B}_{il} = \frac{\exp(-\exp(\eta_{il}))\exp(\eta_{il})}{1-\exp(-\exp(\eta_{il}))}$$

$$\mathrm{C}_{il} = \frac{-\exp(-\exp(\eta_{il}))\exp(\eta_{il})}{\exp(-\exp(\eta_{il}))}$$

The elements of the Hessian matrix $\ddot{\ell}_{compl}$ are easily derived considering the second derivatives of the expressions above.

Appendix B

Elements of the matrix Δ of Section 4.1

Considering the 21×1 vector of parameters

$$\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \alpha_0, \alpha_1, \alpha_2, \alpha_3, \gamma_l, \sigma_i^2, \sigma_s^2, \sigma_{i,s}),$$

the matrix Δ corresponds to the 21×250 submatrix $\frac{\partial^2 \ell(\theta|\mathbf{w})}{\partial\theta\partial\mathbf{w}}$ of the Hessian $\ddot{\ell}_{obs}(\theta|\mathbf{w})$; the latter has been estimated using the parametric bootstrap approximation of the Louis formula where the score vector $\dot{\ell}_{compl}(\theta|\mathbf{w})$ in (3.8) now contains the additional elements

$$\frac{\partial \ell}{\partial w_i} = \sum_{i}^{250} \sum_{j}^{(n_i+1)^{d_i}(n_i)^{1-d_i}} [R_{ij}\, b_{i1} \mathrm{B}_{il} + (1 - R_{ij})\, b_{i1} \mathrm{C}_{il}]$$

(see Appendix A for expressions for B and C).

Appendix C

Convergence of the modified Fisher scoring algorithm of Subsection 5.1.1

Consider the generic p-dimensional vector of parameters $\boldsymbol{\theta}$ and a likelihood based model fitting. In the presence of missing data and with categorical outcomes, define the map $M : \mathcal{M} \rightarrow \mathcal{B}$ from the set of possible completions of the data $\mathcal{M}$ to the corresponding set of possible parameter estimates $\mathcal{B}$. We show that, for $j \in 1, \ldots, p$, the algorithm described in Subsection 5.1.1 converges to $\bar{\boldsymbol{\theta}}^j \in \mathcal{B}$ where $\bar{\boldsymbol{\theta}}^j$ indicates the vector of parameter estimates with the minimum value for the j^{th} element of $\boldsymbol{\theta}$. Convergence to the maximum can be shown using similar arguments and will be omitted.

We first prove that, globally, the algorithm will convergence to $\bar{\boldsymbol{\theta}}^j$. Given a starting point $\boldsymbol{\theta}_{(k)}$, a step of the Fisher scoring produces a successor point $\boldsymbol{\theta}_{(k+1)}$ as

$$\boldsymbol{\theta}_{(k+1)} = \boldsymbol{\theta}_{(k)} - \{I(\boldsymbol{\theta}_{(k)})\}^{-1} \nabla f(\boldsymbol{\theta}_{(k)}). \tag{C.1}$$

where f is the log-likelihood function and $I = E\{H\}$ is the expectation of the Hessian matrix H. Expression (C.1) is a point-to-point map, A, generating a sequence $\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \boldsymbol{\theta}_{(3)}, \ldots$ where $\boldsymbol{\theta}_{(k+1)} = A(\boldsymbol{\theta}_{(k)})$ for each k. In particular for

the algorithm in Subsection 5.1.1 it is

$$A(\boldsymbol{\theta}_{(k)}) = \boldsymbol{\theta}_{(k+1)} = \boldsymbol{\theta}_{(k)} - \{I(\boldsymbol{\theta}_{(k)})\}^{-1} \max_{\mathbf{y}^{miss} \in \mathcal{M}} \{\nabla f(\boldsymbol{\theta}_{(k)})\}^j \qquad \text{(C.2)}$$

where the superscript j on the RHS indicates that the maximum of the j-th entry of the gradient is to be found over all possible completions in $\mathcal{M}$.

We want to show that the sequence of points $\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \boldsymbol{\theta}_{(3)}, \ldots$ with $\boldsymbol{\theta}_{(1)} > \bar{\boldsymbol{\theta}}^j$ converges to $\bar{\boldsymbol{\theta}}^j$ where convergence takes place if, given $\epsilon > 0$, $\boldsymbol{\theta}_{(k+1)} - \boldsymbol{\theta}_{(k)} < \epsilon$ or, equivalently, if

$$\max_{\mathbf{y}^{miss} \in \mathcal{M}} \{\nabla f(\boldsymbol{\theta}_{(k)})\}^j < \epsilon.$$

From (C.1) and known results on the concavity of the log-likelihood (Pratt, 1981), the descent direction of the scoring algorithm is necessarily downhill for minimizing f as I is positive (semi)definite. This implies that the succession of points generated by A must converge to $\bar{\boldsymbol{\theta}}^j$. In fact, if by contradiction we had $\boldsymbol{\theta}_{(k)} \rightarrow \boldsymbol{\theta}^j > \bar{\boldsymbol{\theta}}^j$ then

$$\exists \epsilon > 0 : \max_{\mathbf{y}^{miss} \in \mathcal{M}} \{\nabla f(\boldsymbol{\theta})\}^j > \epsilon$$

and the stopping rule would not be met. Also

$$\max_{\mathbf{y}^{miss} \in \mathcal{M}} \{\nabla f(\boldsymbol{\theta}_{(k+1)})\}^j < \max_{\mathbf{y}^{miss} \in \mathcal{M}} \{\nabla f(\boldsymbol{\theta}_{(k)})\}^j$$

for all k.

Close to the solution then, usual local convergence properties of the Fisher scoring algorithm apply as described, for example, in Osborne (1992) and Bazaraa et al. (1993).

Appendix D

The IGLS algorithm for fitting linear mixed models

For the measurement taken on subject i at visit j, a linear mixed model can be written as

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b} + e_{ij}$$

where $\mathbf{X}_{ij}$ is the j-th row of the $n_i \times p$ design matrix $\mathbf{X}_i$ of explanatory variables for the fixed effects part of the model, n_i the number of observations taken on subject i, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect parameters, $\mathbf{Z}_{ij}$ is the j-th row of the $n_i \times q$ design matrix $\mathbf{Z}_i$ for the random part of the model, $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{D})$ and $e_{ij} \sim N(0, \sigma^2)$ for all $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$.

If the variance-covariance parameters in $\mathbf{D}$ are known, indicating with $\mathbf{X}$ the $n_i N \times p$ matrix obtained by stacking the $\mathbf{X}_i$ on top of each other and with $\mathbf{Y} = (y_{11}, y_{12}, \ldots, y_{n_N N})$ the $n_i N \times 1$ vector of responses, the Generalized Least Squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} \tag{D.1}$$

where V is block diagonal with generic block corresponding to subject i given by $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \oplus_j\sigma^2$, $\oplus_j$ indicating the $n_i \times n_i$ identity matrix. In general, the elements of $\mathbf{D}$ have to be estimated.

- Step 1: Obtain initial OLS estimates of the fixed coefficients $\hat{\beta}$.
- Step 2: Form the vector of raw residuals $\tilde{\mathbf{Y}} = \{\tilde{y}_{ij}\} = \{y_{ij} - \mathbf{X}_{ij}\hat{\beta}\}$ and the corresponding cross-product matrix $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$. This can be rearranged as a vector by stacking the columns on top of each other to obtain $\mathbf{Y}^{**} = vec(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T)$, say. The latter are now considered as observed responses in a linear model with the distinct entries of $\mathbf{D}$, $\boldsymbol{\theta}$, as unknown parameters to be estimated. For example, in a mixed model with only a random intercept term and two subjects we have

$$\mathbf{Y}^{**} = vec(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T) = \begin{pmatrix} y_{11}^2 \\ y_{12}y_{11} \\ \vdots \\ y_{22}^2 \end{pmatrix} = \sigma_{b_0}^2 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix} + R = \mathbf{Z}^*\boldsymbol{\theta} + R.$$

Also, $E(\mathbf{Y}^{**}) = \mathbf{Z}^*\boldsymbol{\theta}$ and therefore the GLS estimator of the vector $\boldsymbol{\theta}$ is

$$\hat{\theta} = (\mathbf{Z}^{*T}\mathbf{V}^{*-1}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*T}\mathbf{V}^{*-1}\mathbf{Y}^{**}$$

where in the expression above $\mathbf{V}^* = \mathbf{V} \otimes \mathbf{V}$, the symbol $\otimes$ indicating the kronecker product of two matrices.

The generic element of $\mathbf{Z}^{*T}\mathbf{V}^{*-1}\mathbf{Z}^*$ or, similarly, $\mathbf{Z}^{*T}\mathbf{V}^{*-1}\mathbf{Y}^{**}$ can be written as

$$L = vec(\mathbf{A})^T(\mathbf{V}^{-1} \otimes \mathbf{V}^{-1})vec(\mathbf{B})$$

For example, if $\mathbf{A}$ refers to a between-subjects covariance term we have

$$\mathbf{A} = \mathbf{x}_l\mathbf{x}_m^T + \mathbf{x}_m\mathbf{x}_l^T,$$

if it refers to a between-subjects variance term

$$\mathbf{A} = \mathbf{x}_l \mathbf{x}_l^T$$

and for the error term we have the identity matrix

$$\mathbf{A} = \oplus_j 1$$

where in all the expressions above the vectors $\mathbf{x}_l$ and $\mathbf{x}_m$ are the explanatory variables whose random coefficients define the relevant covariance and variance parameters (Goldstein and Rasbash, 1992).

A matrix algebra result gives $L = tr\{\mathbf{A}^T\mathbf{V}^{-1}\mathbf{B}\mathbf{V}^{-1}\}$ (Searle et al., 1982) and since $\mathbf{V}$ is block-diagonal we can evaluate L block by block.

The inverse of the generic block $\mathbf{V}_i$ is

$$\mathbf{V}_i^{-1} = \mathbf{\Sigma}_e^{-1}(\mathbf{I}_{n_i} - \mathbf{Z}_i\mathbf{D}\mathbf{G}_i^{-1}\mathbf{Z}_i\mathbf{\Sigma}_e^{-1})$$

where, for a white noise $\mathbf{\Sigma}_e = \mathbf{I}_{n_i}\sigma^2$ and $\mathbf{G}_i = (\mathbf{I}_q + \mathbf{Z}_i^T\mathbf{\Sigma}_e^{-1}\mathbf{Z}\mathbf{D}) = (\mathbf{I}_q + \mathbf{Z}_{i\Sigma}\mathbf{Z}_{iD})$. Therefore, for the generic block in L and a covariance term we can write

$$\mathbf{A}\mathbf{V}_i^{-1} = (\mathbf{x}_l\mathbf{x}_m^T)\mathbf{\Sigma}_e^{-1}(\mathbf{I}_{n_i} - \mathbf{Z}_{iD}\mathbf{G}_i\mathbf{Z}_{i\Sigma}),$$

for a variance term we have

$$\mathbf{A}\mathbf{V}_i^{-1} = \mathbf{x}_l\mathbf{x}_l^T\mathbf{\Sigma}_e^{-1}(\mathbf{I}_{n_i} - \mathbf{Z}_{iD}\mathbf{G}_i\mathbf{Z}_{i\Sigma})$$

and for the (level-1) error term

$$\mathbf{AV}_i^{-1} = \boldsymbol{\Sigma}_e^{-1}(\mathbf{I}_{n_i} - \mathbf{Z}_{iD}\mathbf{G}_i\mathbf{Z}_{i\Sigma}).$$

- Step 3: obtain new estimates of the fixed coefficients using (D.1).
- Step 4: iterate until convergence.

Upon convergence, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated as $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$. The IGLS estimates thus obtained are equivalent to maximum likelihood (ML) estimates. The algorithm can be modified to obtain restricted iterative generalised least squares estimates (RIGLS) which are equivalent to Restricted Maximum Likelihood estimates (REML) (Goldstein, 1989).

This is done by replacing $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$ by $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T + \mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$ throughout.

Appendix E

Saddlepoint approximations of densities and distribution functions

The key idea behind the saddlepoint approximation of the density of the sum of n random variables is the fact that this can be expressed, by means of Fourier transform, as an integral on the complex plane where the integrand has the form $\exp(n\omega(z))$. Thus, for n large, the major contribution to this integral comes from a neighborhood of the saddlepoint z_0 which is also a zero of $\omega'(z)$. Using the method of steepest descent, one can get a complete expansion of the density with terms in powers of n^{-1}.

Here, we follow the illustration of the technique given in Field and Ronchetti (1990). Other references include the original paper by Daniels (1954) and those by Reid (1988), Goutis and Casella (1999), Huzurbazar (1999) and Mazumdar and Gaver (1984).

E.1 The method of steepest descent to compute asymptotic expansions of integrals

Suppose our aim is to obtain an asymptotic expansion of the integral

$$\int_{\mathcal{P}} e^{v\omega(z)}\xi(z)\,dz \tag{E.1}$$

with v large and positive and ω an analytic functions of z in a domain of the complex plane containing the path of integration $\mathcal{P}$. The way we proceed is as follows (Field and Ronchetti, 1990). We first deform the path of integration so that (1) the new path passes through a zero of the derivative $\omega'(z)$ of ω and (2), on this new path, the imaginary part of ω, $\Im\omega(z)$, is constant.

Write $z = x+iy$, $z_0 = x_0+iy_0$, $\omega(z) = u(x,y)+iv(x,y)$. Then, as $\omega'(z_0) = 0$, the Cauchy-Riemann differential equations and the two conditions above imply that the new path passes through a saddlepoint of V, the surface $(x,y) \mapsto u(x,y)$, and coincides with the path of steepest ascent or descent on V as we move away from the saddlepoint (x_0, y_0). Next, considering that on the new path $\Im\omega(z)$ is constant and equal to $\Im\omega(z_0)$, we have

$$\omega(z) = u(x,y) + iv(x_0,y_0) = u(x_0,y_0) + iv(x_0,y_0) - (u(x_0,y_0) - u(x,y)) = \omega(z_0) - \gamma(x,y). \quad \text{(E.2)}$$

In the previous expression γ is a real function that either increases to $+\infty$ or decreases to $-\infty$. We choose the direction where γ increases to $+\infty$ as otherwise integral (E.1) diverges; this is the path of steepest descent from the saddlepoint and, on this path, (E.1) becomes

$$\int_{\mathcal{P}} e^{v\omega(z)}\xi(z)\,dz = e^{v\omega(z_0)} \int_0^\infty e^{-v\gamma}\xi(z)\frac{dz}{d\gamma}\,d\gamma. \quad \text{(E.3)}$$

Therefore, instead of approximating $\omega(z)$ in the exponent where the error would grow very rapidly, we need only approximate $\frac{dz}{d\gamma}$. This we do with a series expansion near the saddlepoint z_0.

E.2 Saddlepoint approximation of the density function of the sum of random variables

Consider n i.i.d. random variables $X_1, \ldots, X_n$ with density, moment and cumulant generating functions $f(x)$, $M(\alpha) = \int_{-\infty}^{+\infty} e^{\alpha x} f(x)\, dx$ and $K(\alpha) = \ln M(\alpha)$, respectively. Then, by Fourier inversion, the density of the sum $S = \sum_{i=1}^{n} X_i$ can be written as

$$f_n(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} M^n(ir) e^{-irt}\, dr. \tag{E.4}$$

With the change of variable $z = ir$ the previous expression becomes

$$\frac{1}{2\pi i} \int_{-i\infty}^{+i\infty} M^n(z) e^{-tz}\, dz = \frac{1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} \exp\{n[K(z) - zt/n]\}\, dz \tag{E.5}$$

the latter equality following from the fact that it is possible to shift the path of integration by considering any straight line parallel to the imaginary axis (Kolassa, 1994).

Expression (E.5) is in the same form as (E.1) with $v = n\omega(z) = n(K(z) - zt/n)$, $\xi(z) = \frac{1}{2\pi i}$, t fixed and $\mathcal{P}$ any line parallel to the imaginary axis.
The method of steepest descent then requires a change in the path of integration such that this goes through a zero of $\omega'(z)$

$$\omega'(z) = nK'(z) - t = 0.$$

Thus, the saddlepoint z_0 is the solution of the equation $nK'(z_0) = t$.
Daniels (1954) shows that the saddlepoint z_0 is real, so $\omega(z_0)$ is real and since on the new path $\Im\omega(z)$ is constant, this implies $\Im\omega(z) \equiv \Im\omega(z_0) = 0$. Thus, to apply the method of steepest descent, we choose a path of integration consisting of a straight line parallel to the imaginary axis and that goes through the saddlepoint

$z_0 = \alpha_0 \in \mathbb{R}$.

On the new path then, consider a small circle around the saddlepoint α_0 of radius ε. It can be shown that the contribution to the integral outside this circle can be ignored.

Inside this circle, by definition $\omega(z)$ is real and write γ as in (E.2), that is

$$\gamma = \omega(\alpha_0) - \omega(z) = K(\alpha_0) - \alpha_0 t/n - [K(z) - zt/n]. \tag{E.6}$$

We now expand the latter expression in a series around α_0

$$\begin{aligned}\gamma = &-(z-\alpha_0)[K'(\alpha_0) - t/n] - (z-\alpha_0)^2 K''(\alpha_0)/2 \\ &-(z-\alpha_0)^3 K'''(\alpha_0)/6 - (z-\alpha_0)^4 K^{(iv)}(\alpha_0)/24 - \ldots\end{aligned} \tag{E.7}$$

Then, writing $\gamma = \frac{\delta^2}{2}$ (since γ increases monotonically) and with the changes of variable

$$\zeta = (z-\alpha_0)[K''(\alpha_0)]^{1/2} \tag{E.8}$$

and

$$\begin{aligned}\lambda_3(\alpha_0) &= K'''(\alpha_0)/[K''(\alpha_0)]^{3/2} \\ \lambda_4(\alpha_0) &= K^{(iv)}(\alpha_0)/[K''(\alpha_0)]^2\end{aligned}$$

we can write (E.7) as

$$-\delta^2/2 = \zeta^2/2 + \lambda_3(\alpha_0)\zeta^3/6 + \lambda_4(\alpha_0)\zeta^4/24 + \ldots \tag{E.9}$$

as the first term on the RHS of (E.7) vanishes at the saddlepoint. Expression (E.9)

can be inverted to express ζ as a series of δ as

$$\zeta = i\delta + \lambda_3(\alpha_0)\delta^2/6 + \{\lambda_4(\alpha_0)/24 - (5/72)\lambda_3^2\}i\delta^3 \ldots \tag{E.10}$$

Therefore, inside a circle of radius ε around the saddlepoint α_0 and on the new path (of steepest descent), since from (E.8) follows $dz = d\zeta/[K''(\alpha_0)]^{1/2}$, our integral can be written as

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{\mathcal{P}_0} \exp\{n[K(z) - zt/n]\}\, dz = \\
=&\frac{1}{2\pi i}\exp\{nK(\alpha_0) - \alpha_0 t\}\int_{\mathcal{P}_0} e^{-n\gamma}\, dz \\
=&\frac{1}{2\pi i}\frac{\exp\{nK(\alpha_0) - \alpha_0 t\}}{[K''(\alpha_0)]^{1/2}}\int_A^B e^{-n\delta^2/2}\frac{d\zeta}{d\delta}\, d\delta
\end{aligned} \tag{E.11}$$

where the limits of integration A, B correspond to the values of δ where the circle intersects the path. From (E.10) we can write

$$\frac{d\zeta}{d\delta} = i + \frac{\lambda_3(\alpha_0)\delta}{3} + i\left[\lambda_4(\alpha_0)/8 - \frac{5}{24}\lambda_3^2(\alpha_0)\right]\delta^2 \ldots \tag{E.12}$$

which, substituted in (E.11), gives

$$\begin{aligned}
&(1/2\pi i)\frac{\exp\{nK(\alpha_0) - \alpha_0 t\}}{[K''(\alpha_0)]^{1/2}} \times \\
&\int_{-A}^{B} e^{-n\delta^2/2}\left\{i + \lambda_3(\alpha_0)\delta/3 + i\left[\lambda_4(\alpha_0)/8 - \frac{5}{24}\lambda_3^2(\alpha_0)\right]\delta^2 + \ldots\right\} d\delta
\end{aligned} \tag{E.13}$$

Now, applying Watson's Lemma

$$(n/2\pi)^{1/2}\int_{-A}^{B} e^{-n\delta^2/2}\psi(\delta)\, d\delta \sim \psi(0) + \frac{1}{2n}\psi''(0) + \cdots + \frac{1}{(2n)^r}\frac{\psi^{2r}(0)}{r!} + \ldots$$

to (E.13) finally gives

$$f_n(t) = \left[\frac{1}{2\pi n K''(\alpha_0)}\right]^{1/2} \exp\{nK(\alpha_0) - \alpha_0 t\} \times \left\{1 + \frac{1}{n}\left[\frac{1}{8}\lambda_4(\alpha_0) - \frac{5}{24}\lambda_3^2(\alpha_0)\right] + \ldots\right\}$$

where the leading term

$$\left[\frac{1}{2\pi n K''(\alpha_0)}\right]^{1/2} \exp\{nK(\alpha_0) - \alpha_0 t\} \tag{E.14}$$

is the saddlepoint approximation.

E.3 Saddlepoint approximation of tail areas

From (E.4), the density function of the sum S of n i.i.d. random variables can be written as

$$f_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{nK(ir)-irx}\, dr.$$

Thus, a tail area corresponding to a value s is given by

$$\begin{aligned} P(S \geq s) &= \int_s^{\infty} \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{nK(ir)-irx}\, dr dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{nK(ir)-irs}\, dr/ir \end{aligned}$$

$$= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{nK(z)-zs}\, dz/z. \tag{E.15}$$

Denote with α the saddlepoint that solves $nK'(z) = s$ and make a change of variable from z to t as follows

$$nK(z) - zs = \frac{t^2}{2} - \gamma t \tag{E.16}$$

such that $t = \gamma$ implies $z = \alpha$. It follows that

$$-\gamma^2 = 2(nK(\alpha) - \alpha s).$$

The tail area (E.15) can therefore be written as

$$P(S \geq s) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{(t^2/2-\gamma t)} G_0(t)\, dt/t \tag{E.17}$$

with $G_0(t) = t/z\frac{dz}{dt}$. Next, we approximate $G_0(t)$ linearly as

$$a_0 + a_1 t$$

with $a_0 = G_0(0)$ and $a_1 = (G_0(0) - G_0(\gamma))/\gamma$. $G_0(0)$ and $G_0(\gamma)$ can be readily evaluated and are equal 1 and $1/(\gamma + \frac{1}{\alpha(K''(\alpha))^{1/2}})$ respectively. Therefore we can write (E.17) as

$$\begin{aligned} P(S \geq s) &\approx \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{1}{t}\, e^{t^2/2-\gamma t}\, dt \\ &\quad + \frac{1}{2\pi i}\left(-\frac{1}{\gamma} + \frac{1}{\alpha(K''(\alpha))^{1/2}} \right) \int_{-i\infty}^{i\infty} e^{t^2/2-\gamma t}\, dt \\ &\approx 1 - \Phi(A) - \phi(A)\left(\frac{1}{A} - \frac{1}{B}\right) \end{aligned} \tag{E.18}$$

where the latter expression follows from the inversion formula for distributions (page 483, Feller, 1966 Vol 2) and

$$A = \text{sgn}(\alpha)\{2[\alpha s - nK(\alpha)]\}^{1/2}$$
$$B = \alpha[nK''(\alpha)]^{1/2}.$$

In the case of lattice distributions (when approximating sums of discrete random variables), throughout the expressions above, the saddlepoint α now solves

$$nK'(\alpha) = s - \frac{1}{2}$$

and

$$B = 2\text{sinh}(\alpha/2)[nK''(\alpha)]^{1/2}$$

is used in (E.18) (Skovgaard, 1987).

References

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1990). *Statistical modelling in GLIM.* New York: Oxford University Press.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10.

Attay, A. M. G. (1999). Fitting selection models to longitudinal data with dropout using the stochastic EM algorithm. *PhD thesis. University of Kent, Canterbury, UK.*

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92:1171–1176.

Bazaraa, M., Sherali, H., and Shetty, C. (1993). *Nonlinear Programming: Theory and Algorithms.* New York: John Wiley and Sons.

Bock, H. H. (1983). *Classification and related methods of data analysis.* Amsterdam: North Holland.

Booth, J. G. and Butler, R. W. (1999). An importance sampling algorithm, for exact conditional tests in log-linear models. *Biometrika*, 86:321–332.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed

model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B*, 61:265–285.

Booth, J. G., Hobert, J. P., and Jank, W. (2001). A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling*, 1:333–349.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19:1141–1164.

Carpenter, J., Pocock, S., and Lamm, C.-J. (2002). Coping with missing data in clinical trials: a model based approach applied to asthma trials. *Statistics in Medicine*, 21:1043–1066.

Celeux, G., Chauveau, D., and Diebolt, J. (1995). On stochastic versions of the EM algorithm. *Technical report, Institute national de recherche en informatique et en automatique, Rhone-Alpes, France.*

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

Corcoran, C., Mehta, C., Patel, N., and Senchaudhuri, P. (2001). Computational toots for exact conditional logistic regression. *Statistics in Medicine*, 20:2723–2739.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25:631–650.

Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *Journal of the Royal Statistical Society B*, 50:445–461.

DeGruttula, V. and Tu, X. M. (1994). Modelling prograssion of CD4-Lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014.

Delucchi, K. L. (1994). Methods for the analysis of binary outcome results in the presence of missing data. *Journal of Consulting and Clinical Psychology*, 62:569–575.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38.

Dielbolt, J. and Celeux, G. (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Communication in Statistics and Stochastic Models*, 9:599–613.

Diggle, P. (1989). Testing for random dropouts in repeated measurement data. *Biometrics*, 45:1255–1258.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43:49–93.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford: University press.

Fang, K. T. and Wang, Y. (1994). *Number-theoretic methods in statistics.* London: Chapman and Hall.

Feller, W. (1966). *An introduction to probability theory and its applications.* New York: Wiley.

Field, C. and Ronchetti, E. (1990). *Small sample asymptotics.* Hayward, CA: Institute of Mathematical Statistics.

Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the American Statistical Association*, 57:691–704.

Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, 51:151–168.

Follmann, D. and Wu, M. (1999). Use of summary measures to adjust for informative missingness in repeated measures data with random effects. *Biometrics*, 55:75–84.

Frison, L. and Pocock, S. J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, 11:1685–1704.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* London: Chapman and Hall.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73:43–56.

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika*, 76:622–623.

Goldstein, H. (1995). *Multilevel statistical models.* London: Edward Arnold.

Goldstein, H. (1999). Repeated measures models with informatively missing responses. *Technical report. Institute of Education, London.*

Goldstein, H. and Rasbash, J. (1992). Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. *Computational Statistics & Data Analysis*, 13:63–71.

Goutis, C. and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, 53:216–224.

Hirji, K. F., Mehta, C. R., and Patel, N. R. (1988). Computing distribution for exact logistic regression. *Journal of the American Statistical Association*, 82:1110–1117.

Hogan, J. W. and Laird, N. M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16:259–272.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95:77–88.

Huzurbazar, S. (1999). Practical saddlepoint approximations. *The American Statistician*, 53:225–232.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88:551–564.

Ibrahim, J. G. and Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, 52:1071–1078.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., and Dartigues, J.-F. (1997).

A 5-year longitudinal study of the mini-mental state examination in normal aging. *American Journal of Epidemiology*, 145:498–506.

Kacprzyk, J. and Fedrizzi, M. (1992). *Fuzzy Regression Analysis.* Heidelberg: Physica-Verlag.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine*, 17:2723–2732.

Kenward, M. G., Goetghebeur, E. J. T., and Molenberghs, G. (2001). Sensitivity analysis for incomplete categorical data. *Statistical Modelling*, 1:31–48.

Kenward, M. G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and generalised estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, 50:945–953.

Kenward, M. G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8:51–83.

Kolassa, J. E. (1994). *Series approximation methods in statistics.* New York: Springer-Verlag.

Kolassa, J. E. and Tanner, M. A. (1999). Approximate Monte Carlo conditional inference in exponential families. *Biometrics*, 55:246–251.

Laird, B. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:936–974.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7:305–315.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Little, R. J. A. (1993). Pattern mixture models for multivariate incomplete data. *Journal of the Royal Statistical Society B*, 88:125–134.

Little, R. J. A. (1995). Modelling the dropout mechanism in repeated measures studies. *Journal of the American Statistical Association*, 90:1112–1121.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley and Sons.

Manski, C. F. (1989). Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80.

Marsaglia, G. (1972). Choosing a point from the surface of a sphere. *Annals of Mathematical Statistics*, 43:645–646.

Mazumdar, M. and Gaver, D. P. (1984). On the computation of power-generating system reliability indexes. *Technometrics*, 26:173–185.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). A comparison of three methods of selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions.* New York: John Wiley and Sons.

Mehta, C. R., Patel, N. R., and Senchaudhuri, P. (2000). Monte Carlo methods for conditional logistic regression. *Journal of the American Statistical Association*, 95:99–108.

Miller, M. E., Davis, C. S., and Landis, J. R. (1993). The analysis of longitudinal polytomous data: Generalized Estimating Equations and connections to weighted least squares. *Biometrics*, 49:1033–1044.

Molenberghs, G., Goetghebeur, E., Lipsitz, S. R., and Kenward, M. G. (1999). Non-random missingness in categorical data: strengths and limitations. *The American Statistician*, 53:110–118.

Molenberghs, G., Kenward, M. G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics.*

Molenberghs, G., Kenward, M. G., and Lesaffre (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84:33–44.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, 52:153–161.

Murray, G. D. and Findlay, J. G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine*, 7:941–946.

Osborne, M. R. (1992). Fisher's method of scoring. *International Statistical Review*, 60:99–117.

Pagano, M. and Tritchler, D. (1983). Permutation distributions in polynomial time. *Journal of the American Statistical Association*, 78:435–440.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92:1320–1329.

Palmgren, J. and Ripatti, S. (2002). Fitting exponential family mixed models. *Statistical modelling*, 1:23–38.

Pan, J.-X. and Thompson, R. (1998). Quasi-Monte Carlo EM algorithm for MLEs in generalized linear mixed models. *Compstat 1998: Proceedings on the 13th Symposium in Computational Statistics*, pages 419–424.

Poon, W.-Y. and Poon, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society B*, 61:51–61.

Pratt, J. W. (1981). Concavity of the Log Likelihood. *Journal of the American Statistical Association*, 76:103–106.

Quintana, F. A., Liu, J. S., and del Pino, G. E. (1999). Monte Carlo EM with importance reweighting and its applications in random effects models. *Computational Statistics & Data Analysis*, 29:429–444.

Rascol, O., Brooks, D. J., Korczyn, A. D., Deyn, P. P. D., Clarke, C. E., and Lang, A. E. (2000). A five-year study of the incidence of dyskinesia in patients with early parkinson's disease who were treated with ropinirole or levodopa. *New England Journal of Medicine*, 342:1484–1491.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, 3:213–238.

Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79:321–334.

Robins, J. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank tests. *Biometrics*, 56:779–788.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17:269–302.

Robins, J. M. and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association*, 89:737–749.

Romero, V. J. and Bankston, S. D. (1998). Finite-element/progressive-lattice-sampling response surface methodology and application to benchmark probability quantification problems. *Technical report, Sandia National Laboratories, Albuquerque, New Mexico.*

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995). Handling 'Don't know' survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90:822–828.

Saltelli, A., Chan, K., and Scott, E. M. (2000). *Sensitivity Analysis.* New York: John Wiley and Sons.

Santner, T. J. and Duffy, D. E. (1986). A note on A. Albert and J.A. Anderson's condition for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73:755–758.

Schluchter, M. D. (1988). Analysis of incomplete multivariate data using linear models with structured covariance matrices. *Statistics in Medicine*, 7:317–324.

Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, 11:1861–1870.

Schwartz, D. and Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases*, 20:637–648.

Searle, S. R., Casella, G., and McCulloch, C. E. (1982). *Variance components.* New York: John Wiley and Sons.

Shi, J.-Q. and Lee, S.-Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society B*, 62:77–87.

Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *Journal of Applied Probability*, 24:875–887.

TenHave, T. R., Pulkstenis, E., Kunselman, A., and Landis, J. R. (1997). Mixed effects logistic regression models for longitudinal binary response data with informative dropout. Technical report, Dept of Biostatistics, Pennsylvania State University.

Touloumi, G., Pocock, S. J., Babiker, A. G., and Darbyshire, J. H. (1999). Estimation and comparison of rates of change in longitudinal studies wiht informative drop-outs. *Statistics in Medicine*, 18:1215–1233.

Tritchler, D. (1984). An algorithm for exact logistic regression. *Journal of the American Statistical Association*, 79:709–711.

Tsang, E. (1994). *Foundations of Constraint Satisfaction.* Academic Press.

Vansteelandt, S. and Goetghebeur, E. (2001). Analyzing the Sensitivity of Generalized Linear Models to Incomplete Outcomes via the IDE Algorithm. *Journal of Computational and Graphical Statistics*, 10:656–672.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., and Molenberghs, G. (2002). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Submitted to Biometrika.*

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in practice.* New York: Springer-Verlag.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Verzilli, C. J. and Carpenter, J. R. (2002a). A Monte Carlo EM algorithm for random-coefficient-based dropout models. *Journal of Applied Statistics*, 29:1011–1021.

Verzilli, C. J. and Carpenter, J. R. (2002b). Assessing uncertainty about parameter estimates from incomplete repeated ordinal data. *Statistical Modelling*, 2:203–215.

Verzilli, C. J. and Carpenter, J. R. (2002c). Measuring uncertainty in the presence of missing data using saddlepoint approximations: an application to binary data. *Biostatistics, Submitted to.*

Wang-Clow, F., Lange, M., Laird, N. M., and Ware, J. H. (1995). A simulation study of estimators for rates of change in longitudinal studies with attrition. *Statistics in Medicine*, 14:283–297.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the

EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704.

White, I. R., Bamias, C., Hardy, P., Pocock, S., and Warner, J. (2001). Randomized clinical trials with added rescue medication: some approaches to their analysis and interpretation. *Statistics in Medicine*, 20:2995–3008.

White, I. R. and Goetghebeur, E. J. T. (1998). Clinical trials comparing two treatment policies: which aspects of the treatment policies make a difference? *Statistics in Medicine*, 17:319–339.

White, I. R. and Pocock, S. J. (1996). Statistical reporting of clinical trials with individual changes from allocated treatment. *Statistics in Medicine*, 15:249–262.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 45:939–955.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, 44:175–188.

Ziegler, A., Kastner, C., and Blettner, M. (1998). The Generalized Estimating Equations: an annotated bibliography. *Biometrical Journal*, 40:115–139.