Universiteit Hasselt

Faculteit Wetenschappen

Center for Statistics

Analysis and Sensitivity Analysis for Incomplete Longitudinal Data

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

Caroline Beunckens

Promotor:
Prof. Dr. G. Molenberghs

September 7, 2007

Contents

1

Introduction

In longitudinal studies the response of interest is designed to be measured repeatedly over time for each subject. The so-obtained longitudinal data features correlation in two directions, that is, besides the correlation between the subjects enrolled in the study, there is correlation within each subject due to the collection of repeated measurements over time.

A key characteristic of correlated data is the type of outcome. For the analysis of Gaussian longitudinal data, the linear mixed model is widely accepted as the unifying framework for a variety of correlated settings, including longitudinal data (Verbeke and Molenberghs, 2000). The model contains both subject-specific and autoregressive effects at the same time. Further, this general hierarchical model marginalizes in a straightforward way to a multivariate normal model with directly interpretable mean and covariance parameters, owing to the unique property of the normal distribution that both the marginal, and in fact also the conditional, distribution of a multivariate normal is again normal. This does not hold for the non-Gaussian case, since no natural analog to the multivariate normal distribution is available. Therefore, depending on which of the three model families is chosen, that is, the marginal, random-effects, or conditional model family, different models are conceivable. Two important representatives are generalized estimating equations (GEE, Liang and Zeger, 1986) within the marginal model family and the generalized linear mixed-effects model (GLMM, Molenberghs and Verbeke, 2005) within the random-effects model family. Whereas

the latter is likelihood-based, the former is established upon frequentist statistics.

Data arising from longitudinal studies are often prone to incompleteness. This induces imbalance in the sense that not all planned observations are actually made. In the context of longitudinal studies, missingness predominantly occurs in the form of dropouts, in which subjects fail to complete the study for one reason or another. Since incompleteness usually occurs for reasons outside the control of the investigators and may be related to the outcome measurement of interest, it is generally necessary to address the process that governs incompleteness. Only in special but important cases it is possible to ignore the missingness process. Since one can never be certain about the precise form of the non-response process, certain assumptions have to be made.

In his 1976 paper, Rubin provided a formal framework for the field of incomplete data by introducing the important taxonomy of missing data mechanisms, consisting of missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). A non-response process is said to be MCAR if the missingness is independent of both unobserved and observed outcomes, but potentially depends on covariates. An MAR mechanism depends on the observed outcomes and perhaps also on the covariates, but not further on unobserved measurements. Finally, when an MNAR mechanism is operating, missingness does depend on unobserved measurements, maybe in addition to dependencies on covariates and/or on observed outcomes.

At the same time, the selection model, pattern-mixture model, and shared-parameter model frameworks have been established. In a selection model, the joint distribution of each subjects outcomes and the vector of missingness indicators is factored as the marginal outcome distribution and the conditional distribution of the missingness indicators given the outcomes. A pattern-mixture approach starts from the reverse factorization. In a shared-parameter model, a set of latent variables, latent classes, and/or random effects is assumed to drive both the measurement and non-response processes. An important version of such a model further asserts that, conditional on the latent variables, those two processes exhibit no further dependence. Rubin (1976) contributed the concept of ignorability, stating that under precise conditions, the missing data mechanism can be ignored when interest lies in inferences about the measurement process. Combined with regularity conditions, ignorability applies to MCAR and MAR combined, when likelihood or Bayesian inference routes are chosen, but the stricter MCAR condition is required for frequentist inferences to be generally valid.

First, the key examples that will be used throughout this thesis are introduced in **Chapter 2**. Next, in **Chapter 3** a detailed description of the main concepts regarding modeling incompleteness as well as longitudinal data is provided. Both the Gaussian and non-Gaussian case are considered.

Early work regarding missingness focused on the consequences of the induced lack of balance or deviations from the study design (Afifi and Elashoff, 1966; Hartley and Hocking, 1971). Later, algorithmic developments took place, such as the expectation-maximization algorithm (EM, Dempster, Laird and Rubin, 1977) and multiple imputation (Rubin, 1987).

Two simple approaches that are still commonly used are (1) a complete case analysis (CC), which restricts the analysis to those subjects for which all information has been measured according to the design of the study (2) simple imputation, such as last observation carried forward (LOCF), for which the last observed measurement is substituted for values at later points in time that are not observed. Claimed advantages include computational simplicity, no need for a full longitudinal model (for instance when the scientific question is in terms of the last planned measurement occasion only) and, for LOCF, compatibility with the intention-to-treat (ITT) principle. As explained in **Chapter 4**, it is unfortunate that so much emphasis has been given to these *ad hoc* methods. Besides the danger for bias and inefficiency, CC, LOCF and simple imputation methods require, at least, the missingness mechanism to be MCAR, a often too strong restriction. Further, we will argue that likelihood-based analyses, which are valid under the MAR missingness mechanism, not only enjoy much wider validity than the simple methods but moreover are simple to conduct, without additional data manipulation. Therefore, analysis of incomplete longitudinal data should shift away from the ad hoc methods and focus on likelihood-based ignorable primary analyses instead, that is, using the linear mixed model and the generalized linear mixed model for Gaussian and non-Gaussian data respectively.

As mentioned before, GEE is an attractive semi-parametric approach for non-Gaussian data within the marginal model family. However, it is based on frequentist methods and thus requires the missingness to be MCAR. Weighted GEE (WGEE) has been proposed by Robins, Rotnitzky and Zhao (1995) as a way to ensure validity under MAR. Alternatively, multiple imputation can be used to pre-process incomplete data, after which GEE is applied, resulting in so-called MI-GEE. In **Chapter 5**, both WGEE and MI-GEE are compared using asymptotic as well as small-sample simulations, in a variety of correctly and incorrectly specified models. In spite of the asymptotic unbiasedness of WGEE, results provide striking evidence that MI-GEE is both less biased and more accurate in the small to moderate sample sizes which

typically arise in real life settings.

So far, it is clear that not only it is advisable to avoid simple ad hoc methods, such as CC and LOCF, but there exist more appropriate flexible methods, which are valid under the weaker MAR assumption and easy to implement in statistical software, such as direct-likelihood, multiple imputation, WGEE and MI-GEE. However, one should consider possible departures from MAR and the consequences this might have on the inference and conclusions. In general, as mentioned before, reasons for non-response or dropout in particular are varied and therefore it is usually impossible to fully justify on a priori ground the assumption of MAR. At first sight, this suggests a need for MNAR models. However, some careful considerations have to be made, the most important one of which is that no modelling approach, whether either MAR or MNAR, can recover the lack of information that occurs due to incompleteness of the data. In the first part of **Chapter 6**, an overview is given of full selection models, such as the models proposed by Diggle and Kenward (1994) for continuous outcomes and by Baker, Rosenberger and DerSimonian (1992) for binary outcomes, as well as pattern-mixture models. The second part is devoted to the proof that the empirical distinction between MAR and MNAR is not possible, in the sense that each MNAR model fit to a set of observed data can be reproduced exactly by an MAR counterpart, a so-called MAR bodyguard.

Together with the fact that an MNAR model is not verifiable from the observed data, since it relies on modeling assumptions about the unobserved data which in general will never be known, rather then forgetting or blindly shifting to the MNAR framework, the optimal place for MNAR modeling is within a sensitivity analysis context. In **Chapter 7** different tools to perform such sensitivity analyses are discussed and proposed, such as methods to assess the influence of subjects based on global and local influence, or for instance using the MAR bodyguard as discussed in the previous chapter.

A modeling framework combining features from selection, pattern-mixture and shared-parameter models is proposed in **Chapter 8**. A flexible model is developed based on a common latent structure governing both the response and missingness process. This latent mechanism subdivides the subjects into different latent groups, which allows for classification of subjects. The resulting model is called a latent-class mixture model. Besides the fact that it allows for flexible MNAR modeling, it is also a useful tool for sensitivity analysis.

To conclude this thesis, a case study is reported in **Chapter 9** using several methods to perform a thorough sensitivity analysis, and concluding remarks are recapitulated in **Chapter 10**. In the final chapter (**Chapter 11**) it is shown how several methods to analyse incomplete longitudinal data can be implemented using the SAS and GAUSS software.

2

Key Examples

In this chapter, four key examples are introduced. Except for the Slovenian public opinion survey (Section 2.3), all are clinical studies. The orthodontic growth data, introduced in Section 2.1, while conducted in human subjects, is of more an epidemiological nature, as opposed to the two depression trials (Section 2.2) and the age-related macular degeneration trial (Section 2.4), which are clinical studies. Whereas the orthodontic growth data, the two depression trials, and the Slovenian public opinion survey are considered for illustrative purposes throughout the various chapters, a detailed analysis is performed of the age-related macular degeneration trial in Chapter 9.

2.1 Orthodontic Growth Data

The orthodontic growth data are introduced by Potthoff and Roy (1964) and contain growth measurements for 11 girls and 16 boys. For each subject, the distance in millimeters from the center of the pituitary to the pterygomaxillary fissure was recorded at ages 8, 10, 12, and 14. The research question is to determine whether dental growth is related to gender. The data were used by Jennrich and Schluchter (1986) to illustrate estimation methods for unbalanced data, where unbalancedness is now to be interpreted in the sense of an unequal number of boys and girls. The data are presented in Table 2.1. Individual profiles and sex group by age means are plotted in Figure 2.1.

Table 2.1: *Orthodontic growth data. Data for 11 girls and 16 boys. Measurements marked with ∗ were deleted by Little and Rubin (1987).*

	Age (in years)					Age (in years)			
Girl	8	10	12	14	Boy	8	10	12	14
1	21.0	20.0	21.5	23.0	1	26.0	25.0	29.0	31.0
2	21.0	21.5	24.0	25.5	2	21.5	22.5∗	23.0	26.5
3	20.5	24.0∗	24.5	26.0	3	23.0	22.5	24.0	27.5
4	23.5	24.5	25.0	26.5	4	25.5	27.5	26.5	27.0
5	21.5	23.0	22.5	23.5	5	20.0	23.5∗	22.5	26.0
6	20.0	21.0∗	21.0	22.5	6	24.5	25.5	27.0	28.5
7	21.5	22.5	23.0	25.0	7	22.0	22.0	24.5	26.5
8	23.0	23.0	23.5	24.0	8	24.0	21.5	24.5	25.5
9	20.0	21.0∗	22.0	21.5	9	23.0	20.5	31.0	26.0
10	16.5	19.0∗	19.0	19.5	10	27.5	28.0	31.0	31.5
11	24.5	25.0	28.0	28.0	11	23.0	23.0	23.5	25.0
					12	21.5	23.5∗	24.0	28.0
					13	17.0	24.5∗	26.0	29.5
					14	22.5	25.5	25.5	26.0
					15	23.0	24.5	26.0	30.0
					16	22.0	21.5∗	23.5	25.0

Sources: Potthoff and Roy (1964) and Jennrich and Schluchter (1986).

Little and Rubin (1987) deleted 9 of the $[(11+16)\times 4] = 108$ measurements, thereby producing 9 incomplete subjects. Deletion is confined to the age 10 measurements. They describe the mechanism to be such that subjects with a low value at age 8 are more likely to have a missing value at age 10. In Table 2.1, the measurements that were deleted are marked with an asterisk. The advantage of this example is that we have the complete data set, that is, the original data, as well as the incomplete data available.

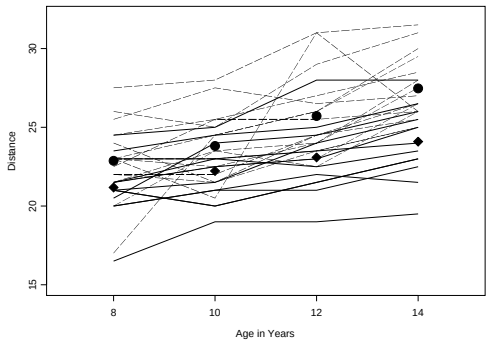
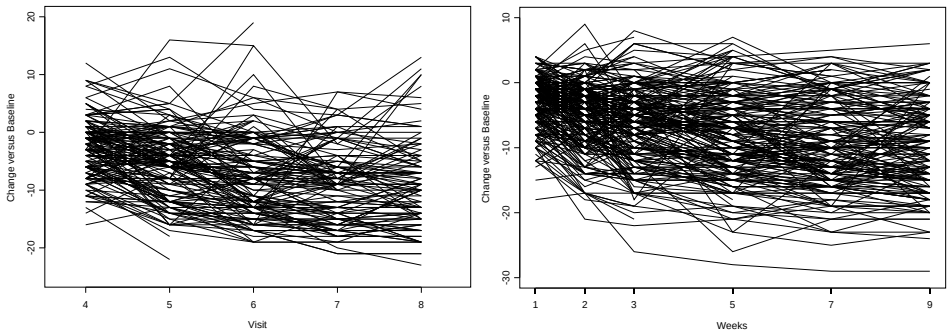


Figure 2.1: *Orthodontic growth data. Observed profiles and group by age means. Solid lines and diamonds are for girls, dashed lines and bullets are for boys.*

2.2 Two Depression Trials

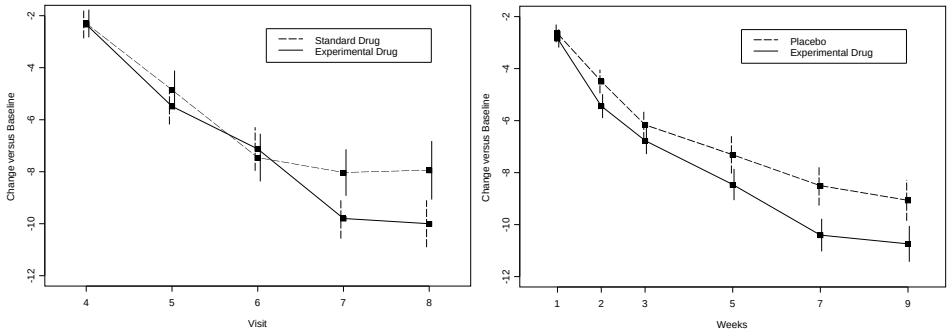
Two depression trials are considered, arising from two randomized, double-blind psychiatric clinical trials, conducted in the United States, and enrolling 342 and 357 patients, respectively. We refer to these clinical trials as respectively First and Second Depression Trial. Hamilton (1960) introduced the Hamilton Depression Rating Scale ($HAMD$), a 21-question multiple choice questionnaire which is used to measure the depression status of a patient. Presently, it is one of the most commonly used scales for rating depression in medical research. The questionnaire rates the severity of symptoms observed in depression such as low mood, insomnia, agitation, anxiety and weight-loss. The doctor must choose the possible responses to each question by interviewing the patient and observing their symptoms. Each question has between 3-5 possible responses which increase in severity. The first 17 questions contribute to the total score and questions 18-21 are recorded to give further information about the depression such as the presence of paranoid symptoms. Both depression trials consider this total $HAMD$ score, which is denoted by $HAMD_{17}$. Besides this continuous $HAMD_{17}$ score, we will also consider the dichotomized version, which distinguishes between patients diagnosed to be depressed ($HAMD_{17} > 7$), or not. For each patient, a baseline assessment is available.

Individual profiles are shown in Figure 2.2. Figure 2.3 pictures the mean profiles with standard errors for each treatment arm separately. Further, the dropout pattern is plotted in Figure 2.4, which shows the percentage of patients remaining on study at each time point. There are few dissimilarities between the First and Second Depression Trial.



(a) First depression trial (b) Second depression trial

Figure 2.2: *Two depression trials. Individual profiles.*

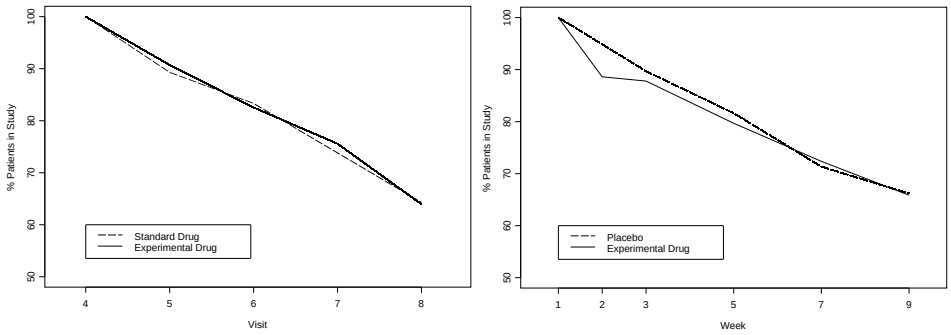


(a) First Depression Trial (b) Second Depression Trial

Figure 2.3: *Two depression trials. Mean profiles with standard errors by treatment arm.*

First Depression Trial In the first depression trial, patients have received either the primary dose of the experimental drug (treatment arm 1), or the secondary dose (treatment arm 2), or one of two non-experimental drugs (treatment arm 3 and 4). The primary objective of this study is the difference in treatment effect between treatment arm 1 and 4. Therefore, only observations corresponding to these treatment arms are included in the analyses, resulting in measurements of 170 patients. Further, the First Depression Trial contains measurements for 5 post-baseline visits going from visit 4 to 8. The exact time interval between visits is not recorded.

Figure 2.3(a) shows a similar mean profile for both treatment arms up to visit



(a) First Depression Trial (b) Second Depression Trial

Figure 2.4: *Two depression trials. Percentage of patients in the study at each time point by treatment arm.*

6, whereafter the standard drug mean profile flattens and the one of the experimental drug decreases, rendering an observed difference at the last visit. For both treatment arms, the dropout patterns are resembling (Figure 2.4(a)), resulting in a completion rate of about 64%.

Second Depression Trial The primary objective of the second depression trial is to compare the efficacy of an experimental anti-depressant with placebo to support a New Drug Application. Visits were scheduled once a week for the first 3 weeks after randomization and every 2 weeks thereafter, resulting in 6 post-baseline measurements taken at week 1, 2, 3, 5, 7, and 9.

The mean evolution over time appears to be quadratic (Figure 2.3(b)), and the difference between placebo and the new drug increases over time. Figure 2.4(b) shows a similar dropout pattern for both treatment arms with a completion rate around 66%. In contrast to the first depression trial, the second one has additional information about the reason of dropout. Adverse event and lack of efficacy are the main reasons for dropout in respectively the experimental and placebo treatment arm.

Results of this clinical trial are originally reported by Detke *et al.* (2002). The experimental drug was found to be significantly superior to placebo on the a priori declared primary efficacy analysis of mean change to endpoint on the $HAMD_{17}$ total score.

Table 2.2: *Slovenian public opinion survey. The* Don't Know *category is indicated by* $*$.

		Independence		
Secession	Attendance	Yes	No	$*$
Yes	Yes	1191	8	21
	No	8	0	4
	$*$	107	3	9
No	Yes	158	68	29
	No	7	14	3
	$*$	18	43	31
$*$	Yes	90	2	109
	No	1	2	25
	$*$	19	8	96

2.3 The Slovenian Public Opinion Survey

In 1991 Slovenians voted for independence from former Yugoslavia in a plebiscite. To prepare for this result, the Slovenian government collected data in the Slovenian public opinion survey, a month prior to the plebiscite. Rubin, Stern and Vehovar (1995) studied the three fundamental questions added to the survey and, in comparing it to the plebiscite's outcome, drew conclusions about the missing data process.

The three questions added were: (1) Are you in favour of Slovenian independence? (2) Are you in favour of Slovenia's secession from Yugoslavia? (3) Will you attend the plebiscite? In spite of their apparent equivalence, questions (1) and (2) are different since independence would have been possible in confederal form as well and therefore the secession question is added. Question (3) is highly relevant since the political decision was taken that not attending was treated as an effective NO to question (1). Thus, the primary estimand is the proportion θ of people that will be considered as voting YES, which is the fraction of people answering yes to both the attendance and independence question. The raw data are presented in Table 2.2.

Table 2.3: *Age-related macular degeneration trial. Mean (standard error) of visual acuity at baseline, at 4, 12, 24, and 52 weeks according to randomized treatment group (placebo* versus *interferon-α).*

Time point	Placebo	Interferon-α	Total
Baseline	55.3 (1.4)	54.6 (1.4)	55.0 (1.0)
4 weeks	54.0 (1.5)	50.9 (1.5)	52.5 (1.1)
12 weeks	52.9 (1.6)	48.7 (1.7)	50.8 (1.2)
24 weeks	49.3 (1.8)	45.5 (1.8)	47.5 (1.3)
1 year (52 weeks)	44.4 (1.8)	39.1 (1.9)	42.0 (1.3)

The data were introduced into the statistical literature by Rubin, Stern and Vehovar (1995) and used by Molenberghs, Kenward and Goetghebeur (2001a) to illustrate their sensitivity analysis tool, the interval of ignorance.

2.4 The Age Related Macular Degeneration Trial

These data arise from a randomized multi-centric clinical trial comparing an experimental treatment (interferon-α) with a corresponding placebo in the treatment of patients with age-related macular degeneration. In this thesis, we focus on the comparison between placebo and the highest dose (6 million units daily) of interferon-α, but the full results of this trial have been reported elsewhere (Pharmacological Therapy for Macular Degeneration Study Group, 1997). Patients with macular degeneration progressively lose vision. In the trial, the patients' visual acuity was assessed at different time points (4 weeks, 12 weeks, 24 weeks, and 52 weeks) through patients' ability to read lines of letters on standardized vision charts. These charts display lines of 5 letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). The patient's visual acuity is the total number of letters correctly read. In addition, one often refers to each line with at least four letters correctly read as a 'line of vision.' An endpoint of interest in this trial was the visual acuity at 1 year (treated as a continuous endpoint). Table 2.3 shows the visual acuity recorded as the numbers of letters read (mean and standard error) by treatment group at baseline, and at the four measurement occasions after baseline. An alternative way to measure visual acuity is by dichotomizing the continuous version. We will consider

Table 2.4: *Age-related macular degeneration trial. Overview of missingness patterns and the frequencies with which they occur. 'O' indicates observed and 'M' indicates missing.*

Measurement occasion					
4 wks	12 wks	24 wks	52 wks	Number	%
Completers					
O	O	O	O	188	78.33
Dropouts					
O	O	O	M	24	10.00
O	O	M	M	8	3.33
O	M	M	M	6	2.50
M	M	M	M	6	2.50
Non-monotone missingness					
O	O	M	O	4	1.67
O	M	M	O	1	0.42
M	O	O	O	2	0.83
M	O	M	M	1	0.42

the increase or decrease in numbers of letters read compared with baseline, however, another dichotomized version could be used as well, for instance, at least 3 lines of vision lost versus less than 3 lines of vision lost.

Regarding missingness in the ARMD data set, an overview of the different dropout patterns is given in Table 2.4. Clearly, both intermittent missingness as well as dropout occurs. It is observed that 188 of the 240 profiles are complete, which is a percentage of 78.33%, while 18.33% (44 subjects) exhibit monotone missingness. Out of the latter group, 2.50% or 6 subjects have no follow-up measurements. The remaining 3.34%, representing 8 subjects, have intermittent missing values. Although the group of dropouts is of considerable magnitude, the ones with intermittent missingness is much smaller. Nevertheless, it is cautious to include all into the analyses.

Both the original quasi-continuous outcome, visual acuity, as well as the binary indicator for increase or decrease in number of letters read compared to baseline, will be analysed in Chapter 9.

3

Fundamental Concepts of Incomplete Longitudinal Data

Longitudinal data are common in biomedical research and beyond. A typical longitudinal study would consist of observing a particular characteristic at several planned occasions, taken in relation to covariates of interest. Data arising from such investigations, however, are often prone to incompleteness, or missingness. In the context of longitudinal studies, missingness predominantly occurs in the form of dropout, in which subjects fail to complete the study for one reason or another. Since missingness usually occurs for reasons outside of the control of the investigators, and may be related to the outcome of interest, it is generally necessary to address the process that governs incompleteness. Only in special but important cases it is possible to ignore the missingness process.

In this chapter, we first introduce some general concepts of modelling incomplete data (Section 3.1). In Section 3.2 we discuss methods to model longitudinal data both in the Gaussian and non-Gaussian setting. For continuous repeated measurements, the linear mixed model is considered. Next, we focus on the situation of non-Gaussian outcomes, for which we distinguish between three model families: marginal, random-

effects, and conditional models. We highlight two important representatives, that is, generalized estimating equations (GEE, Liang and Zeger, 1986) within the marginal family, and the generalized linear mixed model (GLMM, Stiratelli, Laird and Ware, 1984; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) within the random-effects family. Further, we also display the weighted version of GEE, termed weighted generalized estimating equations (WGEE), introduced by Robins, Rotnitzky and Zhao (1995).

3.1 General Concepts of Modelling Incompleteness

The nature of the missingness mechanism can affect the analysis of incomplete data and its resulting statistical inference. Therefore we will introduce the terminology and notation necessary when modelling incomplete data, as well as the different missing data mechanisms. The important case where the missing data mechanism can be ignored, or excluded from the statistical analysis, will also be considered.

3.1.1 The Name of the Game

Let the random variable Y_{ij} denote the response of interest, for the ith study subject ($i = 1, \ldots, N$), designed to be to be measured at occasions t_{ij} ($j = 1, \ldots, n$). Independence across subjects is assumed. The outcomes are grouped into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})'$. In addition, for each occasion j, define R_{ij} as being equal to 1 if Y_{ij} is observed and 0 otherwise. The missing data indicators R_{ij} are grouped into a vector $\boldsymbol{R}_i$, which is of the same length as $\boldsymbol{Y}_i$. If the missingness is due to dropout, measurements for each subject are recorded up to a certain time point, after which all data are missing. In this case, a dropout indicator D_i for the occasion at which dropout occurs can be defined in terms of the missing data indicators, that is, $D_i = 1 + \sum_{j=1}^{n} R_{ij}$. We make the convention that $D_i = n+1$ for a complete sequence. Note that dropout is a particular case of monotone missingness. In order to have a monotone pattern of missingness, there has to exist a permutation of the measurement components such that a measurement earlier in the permuted sequence is observed for at least those subjects that are observed at later measurements. For this definition to be meaningful, we need to have a balanced design in the sense of a common set of measurement occasions for all subjects. Other patterns are called nonmonotone or intermittent missingness. When intermittent missingness occurs, one should use the vector of binary indicators $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in})'$ rather than the dropout indicator D_i.

In principle, n could vary by design across subjects, in which case it would be replaced by n_i. All methodology presented will be valid in such cases too, making the framework suitable for general longitudinal data and designed experiments with a fixed (and common) set of time points applying to all subjects. In many studies, including our examples, n will be constant and therefore our notation will be for this case.

It is often necessary to split the vector $\boldsymbol{Y}_i$ into observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components, respectively. The following terminology is adopted:

Complete data $\boldsymbol{Y}_i$: the scheduled measurements. This is the outcome vector that would have been recorded if there had been no missing data.

Full data $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$: the complete data, together with the missing data indicators. Note that one observes the measurements $\boldsymbol{Y}_i^o$ together with the missingness indicators $\boldsymbol{R}_i$.

Apart from the outcomes, additional information can be measured, which is collected before or during the study. This information is gathered in the covariate matrix $\boldsymbol{X}_i$ and is allowed to change for different measurement occasions. It can include both continuous and discrete outcomes. We assume the covariate vector $\boldsymbol{X}_i$ is fully observed for all subjects. Methods for the case of missing covariates have been explored by several authors (Little, 1992; Robins, Rotnitzky and Zhao, 1994; Zhao, Lipsitz and Lew, 1996).

3.1.2 Missing Data Mechanisms

In principle, one would like to consider the density of the full data $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ describe the measurement and missingness processes, respectively. Covariates are assumed to be measured but, for notational simplicity, suppressed from notation. This full density function can be factorized in different ways, each leading to a different framework. The *selection model* framework (SeM) is based on the following factorization (Rubin, 1976; Little and Rubin, 1987):

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}). \tag{3.1}$$

The first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. The second factor corresponds to the (self-)selection of individuals into "observed" and "'missing" groups. Alternatively, one can consider so-called *pattern-mixture models* (Little, 1993,

1994a, PMM), using the reversed factorization:

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{r}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{\psi}). \tag{3.2}$$

This density can be seen as a mixture of different populations, each of which characterized by the observed pattern of missingness.

Instead of using the selection or pattern-mixture model frameworks, the measurement and the dropout process can be jointly modelled using a *shared-parameter model* (Wu and Carroll, 1988; Wu and Bailey, 1988, 1989; TenHave *et al.*, 1998; Follmann and Wu, 1995; Little, 1995, SPM). In such a model the measurement and dropout process are assumed to be independent, conditional upon a certain set of shared parameters. This shared-parameter model is formulated by way of the following factorization:

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{b}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{b}_i, \boldsymbol{\psi}). \tag{3.3}$$

Here, $\boldsymbol{b}_i$ are shared parameters, often considered to be random effects and following a specific parametric distribution.

Within the selection model framework, Rubin (1976) developed a missing data taxonomy distinguishing between three missingness assumptions, which can be formulated using the second factor on the right hand side of selection model factorization (3.1), that is,

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi}). \tag{3.4}$$

The missingness process is said to be *missing completely at random* (MCAR) if the data are missing for reasons unrelated to the response or to characteristics of individuals. In this case the measurement and missingness process are independent, perhaps conditional on covariates, yielding $f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{\psi})$.

Data are *missing at random* (MAR) if the cause of missingness is allowed to depend on the subject's observed data, but not on their unobserved responses, resulting in $f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi})$.

If the cause of missing data is neither MCAR nor MAR, the data is *missing not at random* (MNAR). In the most general setting, the cause of a subject's missingness depends on their unobserved responses, even after allowing for the information of the observed data. In this case, (3.4) depends on the missing observations, implying the reason for dropout should be modelled simultaneously with the response.

Note that MCAR is equally trivial in the pattern-mixture model framework, where $\boldsymbol{r}_i$ does not influence the mixture components, and in the shared-parameter model framework, where no random-effects are shared between the two factors in (3.3).

Most strategies used to analyze such data are, implicitly or explicitly, based on two choices.

Model for Measurements. A choice has to be made regarding the modeling approach to the measurement sequence. Several views are possible.

View 1. One can choose to analyze the entire longitudinal profile, irrespective of whether interest focuses on the entire profile (e.g., difference in slope between groups) or on a specific time point (e.g., the last planned occasion). In the latter case, the motivation to model the entire profile is because, for example, earlier responses do provide statistical information on later ones. This is especially true when dropout is present. One would then make inferences about such an occasion within the posited full longitudinal model.

View 2. One states the scientific question in terms of the outcome at a well-defined point in time and restricts the corresponding analysis to this particular occasion. Several choices are possible:

View 2a. The scientific question is defined in terms of the *last planned occasion*. Of course, as soon as dropout occurs, such a measurement may not be available. In this case, one can either accept the dropout as it is or use one or other strategy (e.g., imputation, direct likelihood) to incorporate the missing outcomes.

View 2b. One can choose to define the question and the corresponding analysis in terms of the *last observed measurement*.

While Views 1 and 2a necessitate reflection on the missing data mechanism, View 2b avoids the missing data problem because the question is couched completely in terms of observed measurements. While View 2b is sometimes used as an alternative motivation for so-called *last observation carried forward* (LOCF) analysis (Siddiqui and Ali, 1998; Mallinckrodt *et al.*, 2003a,b), a common criticism is that the last observed measurement merges measurements at real stopping times (for dropouts) and at a purely design-based time (for completers). Thus, under View 2b, an LOCF analysis might be acceptable, provided it matched the scientific goals, but is then better described as a Last Observation analysis because nothing is carried forward. Such an analysis should properly be combined with an analysis of time to dropout, perhaps in a survival analysis framework. Of course, an investigator should reflect very carefully on whether View 2b represents a relevant and meaningful scientific question (Shih and Quan, 1997).

Method for Handling Missingness. A choice has to be made regarding the modeling approach for the missingness process. Luckily, under certain assumptions this process can be ignored (e.g., a likelihood-based ignorable analysis, for which MAR is a sufficient condition). Some simple methods, such as a complete case analysis and LOCF, do not explicitly address the missingness process either, but are nevertheless not ignorable. We will return to this issue in Chapter 4.

Let us now describe the measurement and missingness models in turn. The measurement model will depend on whether or not a full longitudinal analysis is done. In case View 2 is adopted, that is, when the focus is on the last observed measurement or on the last measurement occasion only, one typically opts for classical two- or multi-group comparisons (t test, Wilcoxon, etc.). When a longitudinal analysis is deemed necessary, the choice depends on the nature of the outcome. A variety of methods both for Gaussian and non-Gaussian longitudinal data will be discussed in Section 3.2.

Assume that incompleteness is due to dropout only, and that the first measurement Y_{i1} is obtained for everyone. Under the selection model framework, a possible model for the dropout process is a logistic regression for the probability of dropout at occasion j, given that the subject is still in the study. We denote this probability by $g(\boldsymbol{h}_{ij}, y_{ij})$ in which $\boldsymbol{h}_{ij}$ is a vector containing all responses observed up to but not including occasion j, as well as relevant covariates. We then assume that $g(\boldsymbol{h}_{ij}, y_{ij})$ satisfies

$$\text{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] = \text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \boldsymbol{y}_i)\right] = \boldsymbol{h}_{ij}\boldsymbol{\psi} + \omega y_{ij}, \qquad i = 1, \ldots, N, \tag{3.5}$$

(Diggle and Kenward, 1994). When ω equals zero, the dropout model is MAR, and all parameters can be estimated using standard software since the measurement model and the dropout model can then be fitted separately, as will be shown in the next section. If $\omega \neq 0$, the posited dropout process is MNAR. Model (3.5) provides the building blocks for the dropout process $f(d_i|\boldsymbol{y}_i, \boldsymbol{\psi})$. While it has been used, in particular, by Diggle and Kenward (1994), it is, at this stage, quite general and allows for a wide variety of modeling approaches. A review of the Diggle-Kenward model is provided in Section 6.1.1.

3.1.3 Ignorability

Rubin (1976) has shown that, under MAR and when the condition holds that parameters defining the measurement and dropout process, denoted by $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively,

are functionally independent, likelihood-based inference remains valid when the missing data mechanism is ignored. Practically speaking, the likelihood of interest is then based upon the factor $f(\boldsymbol{y}_i^o|\boldsymbol{\theta})$. This is called *ignorability*. Indeed, let us assume the statistical analysis and corresponding inference are likelihood-based. The contribution to the likelihood of a particular subject i, based on (3.1) is of the form

$$L_i^*(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}_i, \boldsymbol{r}_i) \propto f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}).$$

Since inference has to be based on what is observed, the full data likelihood L_i^* has to be replaced by the observed data likelihood L_i:

$$L_i(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}_i^o, \boldsymbol{r}_i) \propto f(\boldsymbol{y}_i^o, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) \tag{3.6}$$

with

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \\ &= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi}) d\boldsymbol{y}_i^m. \end{aligned}$$

Under an MAR process, we obtain

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \\ &= f(\boldsymbol{y}_i^o|\boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi}), \end{aligned} \tag{3.7}$$

that is, the likelihood factorizes into two components of the same functional form as the general factorization (3.1) of the complete data. If further $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct in the sense that the parameter space of the full vector $(\boldsymbol{\theta}', \boldsymbol{\psi}')'$ is the Cartesian product of the two component parameter spaces (separability condition), then inference of the measurement model parameters $\boldsymbol{\theta}$ can be made without explicitly formulating the missing data mechanism, that is, only based on the marginal observed data density $f(\boldsymbol{y}_i^o|\boldsymbol{\theta})$. For Bayesian inferences, the same holds if besides the separability condition, the priors are independent (Little and Rubin, 1987).

In conclusion, when the separability condition is satisfied, *within the likelihood framework*, ignorability is equivalent to the union of MAR and MCAR. Hence, non-ignorability and MNAR are synonyms in this context. A formal derivation is given in Rubin (1976) and Little and Rubin (1987), where it is also shown that the same requirements hold for Bayesian inference, but that frequentist inference is ignorable only under MCAR.

The practical implication of ignorability is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects,

manipulates the correct likelihood and thus provides valid parameter estimates, standard errors if based on the observed information matrix, and likelihood ratio values (Kenward and Molenberghs, 1998). This result makes so-called direct-likelihood analyses, valid under MAR, viable candidates for the status of primary analysis in clinical trials and a variety of other setting (Molenberghs *et al.*, 2004). This will be further discussed in Chapter 4.

A few cautionary remarks are warranted. First, when at least part of the scientific interest is directed towards the missingness process, for instance when one is interested in studying the reason for missingness, obviously both processes need to be considered. Under MAR, both processes can be modeled and parameters estimated separately. Second, likelihood inference is often surrounded with references to the sampling distribution (e.g., to construct measures of precision for estimators and for statistical hypothesis tests; Kenward and Molenberghs (1998)). However, the practical implication is that standard errors and associated tests are valid, when based on the observed rather than the expected information matrix and given that the parametric assumptions are correct. Third, it may be hard to rule out the operation of an MNAR mechanism. The reasons for missingness are varied and it is therefore difficult to fully justify on *a priori* grounds the assumption of MAR. Further, since it is not possible to test for MNAR against MAR (Jansen *et al.*, 2006b), one should always be open to the possibility that the data are MNAR. To explore the impact of deviations from the MAR assumption on the conclusions, one should ideally conduct a sensitivity analysis, within which models for the MNAR process can play a major role (Verbeke and Molenberghs, 2000). This point will be discussed further in Chapters 6 to 8. Fourth, such an analysis can proceed only under View 1, that is, a full longitudinal analysis is necessary, even when interest lies, for example, in a comparison between the two treatment groups at the last occasion. In the latter case, the fitted model can be used as the basis for inference at the last occasion. A common criticism is that a model needs to be considered, with the risk of model misspecification. However, it should be noted that in many clinical trial settings the repeated measures are balanced in the sense that a common (and often limited) set of measurement times is considered for all subjects, allowing the a priori specification of a saturated model (e.g., full group by time interaction model for the fixed effects and unstructured variance-covariance matrix).

3.2 Methodology for Longitudinal Data

Let us now turn attention to standard model frameworks for longitudinal data. First, the continuous case will be treated where the linear mixed model undoubtedly occupies the most prominent role. Then, we switch to the discrete setting, where important distinctions exist between three model families: the marginal, random-effects, and conditional model family. The mixed model parameters, both in the continuous and discrete cases, are usually estimated using maximum likelihood based methods which implies that the results are valid under MAR. A commonly encountered marginal approach to non-Gaussian data is generalized estimating equations (GEE, Liang and Zeger, 1986) which has a frequentist foundation. It is valid only under MCAR (Liang and Zeger, 1986), necessitating the need for extensions, such as weighted GEE (Robins, Rotnitzky and Zhao, 1995), and multiple-imputation based GEE (Schafer, 2003), which will be discussed as well.

3.2.1 Longitudinal Data

Laird and Ware (1982) proposed, for continuous outcomes, likelihood-based mixed-effects models. A broad discussion of such models is provided in Verbeke and Molenberghs (2000). The general linear mixed-effects model is the following:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{3.8}$$

where $\boldsymbol{Y}_i$ is the n-dimensional response vector for subject i, containing the outcomes at n various measurement occasions, $1 \leq i \leq N$, N is the number of subjects, $\boldsymbol{X}_i$ and Z_i are $(n \times p)$ and $(n \times q)$ known design matrices, $\boldsymbol{\beta}$ is the p-dimensional vector containing the fixed effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$ is the q-dimensional vector containing the random effects, and $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \Sigma)$ is a N-dimensional vector of residual components, combining measurement error and serial correlation. Further, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ are assumed to be independent. Finally, D and Σ are general covariance matrices of size $(q \times q)$ and $(n \times n)$ respectively. In case of no serial correlation, Σ reduces to $\sigma^2\boldsymbol{I}_n$. Inference is based on the marginal distribution of the response $\boldsymbol{Y}_i$ which, after integrating over random effects, can be expressed as

$$\boldsymbol{Y}_i \ \sim \ N(\boldsymbol{X}_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma). \tag{3.9}$$

Whereas the linear mixed model is seen as a unifying parametric framework for Gaussian repeated measures (Verbeke and Molenberghs, 2000), there are a variety of methods in common use in the non-Gaussian setting. In line with Fahrmeir and

Tutz (2001), Diggle, Heagerty, Liang and Zeger (2002) and Molenberghs and Verbeke (2005), we distinguish between three model families. In a *population-averaged* or *marginal model*, marginal distributions are used to describe the outcome vector, given a set of predictor variables. The correlation among the components of the outcome vector can then be captured either by adopting a fully parametric approach or by means of working assumptions, such as in GEE (Liang and Zeger, 1986).

Alternatively, in a *subject-specific* or *random-effects model*, the responses are assumed to be independent, given a collection of subject-specific parameters.

Finally, a *conditional model* describes the distribution of the components of the outcome vector, conditional on the predictor variables but also conditional on (a subset of) the other components of the response vector. Well-known members of this class of models are log-linear models (Agresti, 2002). Let us give an example of each for the case of Gaussian outcomes, or more generally for models with a linear mean structure.

A marginal model is characterized by a marginal mean function of the form

$$E(Y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}, \tag{3.10}$$

where $\boldsymbol{x}_{ij}$ is a vector of covariates for subject i at occasion j and $\boldsymbol{\beta}$ is a vector of regression parameters. In a random-effects model we focus on the expectation, additionally conditioning upon a random-effects vector $\boldsymbol{b}_i$:

$$E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i. \tag{3.11}$$

Finally, a simple first-order stationary transition model, which is a particular case of a conditional model, focuses on expectations of the form

$$E(Y_{ij}|Y_{i,j-1}, \ldots, Y_{i1}, \boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \alpha Y_{i,j-1}. \tag{3.12}$$

Alternatively, one might condition upon all outcomes except the one being modeled.

As shown by Verbeke and Molenberghs (2000) random-effects models imply a simple marginal model in the linear mixed model case. This is due to the elegant properties of the multivariate normal distribution. In particular, expectation (3.10) follows from (3.11) by either (a) marginalizing over the random effects or by (b) by conditioning upon the random-effects vector $\boldsymbol{b}_i = \boldsymbol{0}$. Hence, the fixed-effects parameters $\boldsymbol{\beta}$ have a marginal and a hierarchical model interpretation at the same time. Finally, certain auto-regressive models, in which later-time residuals are expressed in terms of earlier ones, lead to particular instances of the general linear mixed effects model as well, and hence have a marginal function of the form (3.10).

Since the linear mixed model has marginal, hierarchical, and conditional aspects, it is clear why it provides a unified framework in the Gaussian setting. However, there does not exist such a close connection when outcomes are of a non-Gaussian type, such as binary, categorical, or discrete.

We will consider the marginal and random-effects model families in turn and then point to some particular issues arising within them or when comparisons are made between them. Further, transition models, a particular type of conditional models, are useful within the longitudinal setting, and will therefore be discussed.

3.2.2 Marginal Models

Thorough discussions on marginal modeling can be found in Diggle, Heagerty, Liang and Zeger (2002) and in Molenberghs and Verbeke (2005). We introduce the marginal models, which will be considered in the subsequent chapters.

The Bahadur Model

Bahadur (1961) proposed a marginal model for binary outcomes, accounting for the association via marginal correlations. Define the marginal probability $\pi_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$, and define standardized deviations

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}, \tag{3.13}$$

where y_{ij} is an actual value of the binary response variable Y_{ij}. Further, let $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2})$, $\rho_{ij_1j_2j_3} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3})$,..., and $\rho_{i12\ldots n} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in})$. Then, the general Bahadur model can be represented by the expression

$$f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)\, c(\boldsymbol{y}_i), \tag{3.14}$$

where

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{1-y_{ij}},$$

and

$$c(\boldsymbol{y}_i) = 1 + \sum_{j_1<j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} + \ldots + \rho_{i12\ldots J} e_{i1} e_{i2} \ldots e_{iJ}.$$

Thus, the probability mass function is the product of the independence model $f_1(\boldsymbol{y}_i)$ and the correction factor $c(\boldsymbol{y}_i)$. One view-point is to consider the factor $c(\boldsymbol{y}_i)$ as a model for overdispersion.

Besides the Bahadur model, a broad set of marginal models have been proposed by Dale (1986), Plackett (1965), Lang and Agresti (1994), and Molenberghs and Lesaffre (1994, 1999). Even though a variety of flexible full-likelihood models exist, maximum likelihood can be unattractive due to excessive computational requirements, especially when high-dimensional vectors of correlated data arise, as alluded to in the context of the Bahadur model. As a consequence, alternative methods have been in demand.

Generalized Estimating Equations

Liang and Zeger (1986) proposed so-called *generalized estimating equations* (GEE), useful to circumvent the computational complexity of full likelihood, and which can be considered whenever interest is restricted to the mean parameters. This approach requires only the correct specification of the univariate marginal distributions, provided one is willing to adopt so-called *working assumptions* about the association structure of the vector of repeated measurements.

Let us introduce more formally the classical form of GEE (Liang and Zeger, 1986; Molenberghs and Verbeke, 2005). The score equations for a non-Gaussian outcome are

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \tag{3.15}$$

where $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i)$ and V_i is the so-called *working covariance matrix*, that is, V_i approximates $\mathrm{Var}(\boldsymbol{Y}_i)$, the true underlying covariance matrix for $\boldsymbol{Y}_i$. This working covariance matrix can be decomposed as $V_i = A_i^{1/2} C_i A_i^{1/2}$, in which $A_i^{1/2}$ is a diagonal matrix with standard deviations of $\boldsymbol{Y}_i$ along the diagonal, and $C_i = \mathrm{Corr}(\boldsymbol{Y}_i)$ is the correlation matrix. The variance of each Y_{ij} is $\mathrm{Var}(Y_{ij}) = \phi\, v(\mu_{ij})$, where $v(\mu_{ij})$ is a known variance function, that is, a known function of μ_{ij}, and ϕ is a scale parameter that may be known or should be estimated. Consequently, $A_i = A_i(\boldsymbol{\beta})$ depends upon the means, hence upon $\boldsymbol{\beta}$ through this variance function $v(\mu_{ij})$, and follows therefore directly from the marginal mean model. On the other hand, $\boldsymbol{\beta}$ commonly contains no information about C_i. Therefore, the correlation matrix C_i typically is written in terms of a vector $\boldsymbol{\alpha}$ of unknown parameters, $C_i = C_i(\boldsymbol{\alpha})$, and will need to be estimated. Liang and Zeger (1986) dealt with this set of nuisance parameters $\boldsymbol{\alpha}$ by allowing for specification of an incorrect structure for C_i or so-called *working correlation matrix*.

Assuming that the marginal mean $\boldsymbol{\mu}_i$ has been correctly specified as $h(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\beta$, they showed that, under mild regularity conditions, the estimator $\widehat{\beta}$ obtained from solving (3.15) is asymptotically normally distributed with mean β and with covariance

matrix

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}, \tag{3.16}$$

where

$$I_0 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \beta} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \beta'} \right) \tag{3.17}$$

$$I_1 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \beta} V_i^{-1} \text{Var}(\boldsymbol{Y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \beta'} \right). \tag{3.18}$$

Consistent estimates can be obtained by replacing all unknown quantities in (3.16) by consistent estimates. Observe that, when C_i is correctly specified, $\text{Var}(\boldsymbol{Y}_i) = V_i$ in (3.18), and thus $I_1 = I_0$. As a result, the expression for the covariance matrix (3.16) reduces to I_0^{-1}, corresponding to full likelihood, that is, when the first and second moment assumptions are correct. Thus, when the working correlation structure is correctly specified, it reduces to full likelihood, although generally it differs from it. On the other hand, when the working correlation structure differs strongly from the true underlying structure, there is no price to pay in terms of the consistency of the asymptotic normality of $\widehat{\boldsymbol{\beta}}$, but such a poor choice may result in loss of efficiency. With incomplete data that are MAR or MNAR, an erroneously specified working correlation matrix may additionally lead to bias (Molenberghs and Kenward, 2007).

Two further specifications are necessary before GEE is operational: $\text{Var}(\boldsymbol{Y}_i)$ on the one hand and $C_i(\boldsymbol{\alpha})$, with in particular estimation of α, on the other hand. Full modeling will not be an option, since we would then be forced to do what we want to avoid. In practice, $\text{Var}(\boldsymbol{Y}_i)$ in (3.18) is replaced by $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'$, which is unbiased on the sole condition of correct mean specification. Secondly, one also needs estimates of the nuisance parameters $\boldsymbol{\alpha}$. Liang and Zeger (1986) proposed moment-based estimates for the working correlation. To this end, deviations of the form

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1 - \pi_{ij})}},$$

are used. Note that $e_{ij} = e_{ij}(\boldsymbol{\beta})$ through $\mu_{ij} = \mu_{ij}(\boldsymbol{\beta})$ and therefore also through $v(\mu_{ij})$, the variance at time j, and hence the jth diagonal element of A_i.

Some of the more popular choices for the working correlations are independence ($\text{Corr}(Y_{ij}, Y_{ik}) = 0,\ j \neq k$), exchangeability ($\text{Corr}(Y_{ij}, Y_{ik}) = \alpha,\ j \neq k$), AR(1) ($\text{Corr}(Y_{ij}, Y_{i,j+t}) = \alpha^t,\ t = 0, 1, \ldots, n_i - j$), and unstructured ($\text{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk},\ j \neq k$).

An overdispersion parameter could be included as well, but we have suppressed it for ease of exposition. The standard iterative procedure to fit GEE, based on Liang and Zeger (1986), is then as follows: (1) compute initial estimates for $\boldsymbol{\beta}$, using a univariate GLM, that is, assuming independence; (2) compute Pearson residuals e_{ij}; (3) compute estimates for $\boldsymbol{\alpha}$; (4) compute $C_i(\boldsymbol{\alpha})$; (5) compute $V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = A_i^{1/2}(\boldsymbol{\beta})\, C_i(\boldsymbol{\alpha})\, A_i^{1/2}(\boldsymbol{\beta})$; (6) update the estimate for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}\, V_i^{-1}\, \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right]^{-1} \left[\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}\, V_i^{-1}\, (\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right].$$

Steps (2)–(6) are iterated until convergence. To illustrate step (3), consider compound symmetry, in which case the correlation is estimated by

$$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n(n-1)} \sum_{j \neq k} e_{ij} e_{ik}.$$

Note that, the GEE moments, that are specified, coincide with those of the Bahadur model, so that the former can be seen as a non-likelihood alternative of the latter, since no distributional assumptions regarding the full joint multivariate distribution of $\boldsymbol{Y}_i$. In summary, GEE for binary data can be seen as a moment-based version of the Bahadur model. Alternatively, it may be helpful to view it as a "correlation-corrected version of logistic regression."

Weighted Generalized Estimating Equations

As Liang and Zeger (1986) pointed out, GEE-based inferences are valid only under MCAR, due to the fact that they are based on frequentist considerations. An important exception, mentioned by these authors, is the situation where the working correlation structure happens to be correct, since then the estimates and model-based standard errors are valid under the weaker MAR assumption. This is because then, the estimating equations can be interpreted as likelihood equations. In general, the working correlation structure will not be correctly specified and hence Robins, Rotnitzky and Zhao (1995) proposed a class of weighted estimating equations to allow for MAR in case missingness is due to dropout.

The idea of weighted generalized estimating equations (WGEE) is to weigh each subject's contribution in the GEEs by the inverse probability that a subject drops out at the time he or she dropped out. Thus, anyone staying in the study is considered representative of himself as well as of a number of similar subjects that did drop out from the study. The incorporation of these weights, reduces possible bias in the

regression parameter estimates, $\hat{\beta}$. Such a weight can be expressed as

$$\begin{aligned}\nu_{ij} \equiv P[D_i = j] &= \prod_{k=2}^{j-1}(1 - P[R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1]) \times \\ &\quad P[R_{ij} = 0 | R_{i2} = \ldots = R_{i,j-1} = 1]^{I\{j \leq n\}},\end{aligned}$$

where $j = 2, 3, \ldots, n+1$.

Recall that we partitioned $\boldsymbol{Y}_i$ into the unobserved components $\boldsymbol{Y}_i^m$ and the observed components $\boldsymbol{Y}_i^o$. Similarly, we can make the exact same partition of $\boldsymbol{\mu}_i$ into $\boldsymbol{\mu}_i^m$ and $\boldsymbol{\mu}_i^o$. In the weighted GEE approach, the score equations to be solved are:

$$S(\beta) = \sum_{i=1}^{N}\sum_{d=2}^{n+1} \frac{I(D_i = d)}{\nu_{id}} \frac{\partial \boldsymbol{\mu}_i}{\partial \beta'}(d)(A_i^{1/2} R_i A_i^{1/2})^{-1}(d)(\boldsymbol{y}(d) - \boldsymbol{\mu}_i(d)) = \mathbf{0},$$

where $\boldsymbol{y}_i(d)$ and $\boldsymbol{\mu}_i(d)$ are the first $d-1$ elements of $\boldsymbol{y}_i$ and $\boldsymbol{\mu}_i$ respectively. We define $\frac{\partial \boldsymbol{\mu}_i}{\partial \beta'}(d)$ and $(A_i^{1/2} R_i A_i^{1/2})^{-1}(d)$ analogously, in line with the definitions of Robins, Rotnitzky and Zhao (1995).

3.2.3 Random-effects Models

While several nonequivalent random-effects models exist, one of the most popular ones is the *generalized linear mixed model* (GLMM, Breslow and Clayton, 1993). The focus will be on this one.

First, a general formulation of mixed-effects models is as follows. Assume that $\boldsymbol{Y}_i$ (possibly appropriately transformed) satisfies

$$\boldsymbol{Y}_i | \boldsymbol{b}_i \sim F_i(\boldsymbol{\theta}, \boldsymbol{b}_i), \tag{3.19}$$

that is, conditional on $\boldsymbol{b}_i$, $\boldsymbol{Y}_i$ follows a pre-specified distribution F_i, possibly depending on covariates, and parameterized through a vector $\boldsymbol{\theta}$ of unknown parameters, common to all subjects. Further, $\boldsymbol{b}_i$ is a q-dimensional vector of subject-specific parameters, called random effects, assumed to follow a so-called mixing distribution G which may depend on a vector $\boldsymbol{\xi}$ of unknown parameters, that is, $\boldsymbol{b}_i \sim G(\boldsymbol{\xi})$. The $\boldsymbol{b}_i$ reflect the between-unit heterogeneity in the population with respect to the distribution of $\boldsymbol{Y}_i$. In the presence of random effects, conditional independence is often assumed, under which the components Y_{ij} in $\boldsymbol{Y}_i$ are independent, conditional on $\boldsymbol{b}_i$. The distribution function F_i in (3.19) then becomes a product over the n independent elements in $\boldsymbol{Y}_i$.

In general, unless a fully Bayesian approach is followed, inference is based on the marginal model for $\boldsymbol{Y}_i$ which is obtained from integrating out the random effects,

over their distribution $G(\boldsymbol{\xi})$. If $f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)$ and $g(\boldsymbol{b}_i)$ denote the density functions corresponding to the distributions F_i and G, respectively, we have that the marginal density function of $\boldsymbol{Y}_i$ equals

$$f_i(\boldsymbol{y}_i) = \int f_i(\boldsymbol{y}_i|\boldsymbol{b}_i) g(\boldsymbol{b}_i) d\boldsymbol{b}_i, \tag{3.20}$$

which depends on the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. Assuming independence of the units, estimates of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\xi}}$ can be obtained from maximizing the likelihood function built from (3.20), and inferences immediately follow from classical maximum likelihood theory.

It is important to realize that the random-effects distribution G is crucial in the calculation of the marginal model (3.20). One often assumes G to be of a specific parametric form, such as a (multivariate) normal. Depending on F_i and G, the integration in (3.20) may or may not be possible analytically. Proposed solutions are based on Taylor series expansions of $f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)$, or on numerical approximations of the integral, such as (adaptive) Gaussian quadrature.

A general formulation of GLMM is as follows. Conditionally on random effects $\boldsymbol{b}_i$, it assumes that the elements Y_{ij} of $\boldsymbol{Y}_i$ are independent, with density function usually based on a classical exponential family formulation, that is, with mean $E(Y_{ij}|\boldsymbol{b}_i) = a'(\eta_{ij}) = \mu_{ij}(\boldsymbol{b}_i)$ and variance $\text{Var}(Y_{ij}|\boldsymbol{b}_i) = \phi\, a''(\eta_{ij})$, and where, apart from a link function h (e.g., the logit link for binary data or the log link for counts), a linear regression model with parameters $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ is used for the mean, that is, $h(\boldsymbol{\mu}_i(\boldsymbol{b}_i)) = \boldsymbol{X}_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i$. Note that the linear mixed model is a special case, with identity link function. The random effects $\boldsymbol{b}_i$ are again assumed to be sampled from a (multivariate) normal distribution with mean $\boldsymbol{0}$ and covariance matrix D. Usually, the canonical link function is used, i.e., $h = {a'}^{-1}$, such that $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i$. When the link function is chosen to be of the logit form and the random effects are assumed to be normally distributed, the familiar logistic-linear GLMM follows.

3.2.4 Marginal *versus* Random-Effects Models

Unlike for correlated Gaussian outcomes, the parameters of the random-effects and marginal models for correlated binary data describe different types of effects of the covariates on the response probabilities (Neuhaus, 1992). Therefore, the choice between population-averaged and subject-specific strategies should heavily depend on the scientific goals. Population-averaged or marginal models evaluate the success probability as a function of covariates only. With a random-effects or subject-specific approach, the response is modeled as a function of covariates and parameters, specific to the

subject. In such models, interpretation of fixed-effects parameters is conditional on a constant level of the random-effects parameter. Population-averaged comparisons, on the other hand, make no use of within cluster comparisons for cluster varying covariates and are therefore not useful to assess within-subject effects (Neuhaus, Kalbfleisch and Hauck, 1991).

It is useful to underscore the difference between the marginal and the random-effects model family, as well as the nature of this difference. To see the nature of the difference, consider a binary outcome variable and assume a random-intercept logistic model with linear predictor $\text{logit}[P(Y_{ij} = 1|t_{ij}, b_i)] = \beta_0 + b_i + \beta_1 t_{ij}$, where t_{ij} is the time covariate. The conditional means $E(Y_{ij}|b_i)$, as functions of t_{ij}, are given by

$$E(Y_{ij}|b_i) = \frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}, \tag{3.21}$$

whereas the marginal average evolution is obtained from averaging over the random effects:

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E\left[\frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}\right] \neq \frac{\exp(\beta_0 + \beta_1 t_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij})}. \tag{3.22}$$

This implies that the interpretation of the parameters in both types of model is completely different. Moreover, under the classical linear mixed model (Verbeke and Molenberghs, 2000), we have that $E(\boldsymbol{Y}_i)$ equals $\boldsymbol{X}_i\boldsymbol{\beta}$, such that the fixed effects have a subject-specific as well as a population-averaged interpretation, whereas under non-linear mixed models this does no longer hold in general. The fixed effects now only reflect the conditional effect of covariates, and the marginal effect is not easily obtained anymore as $E(\boldsymbol{Y}_i)$ is given by

$$E(\boldsymbol{Y}_i) \quad = \quad \int \boldsymbol{y}_i \left\{ \int f_i(\boldsymbol{y}_i|\boldsymbol{b}_i) g(\boldsymbol{b}_i) d\boldsymbol{b}_i \right\} d\boldsymbol{y}_i.$$

In the non-linear case, the interpretation of the parameters in both types of model (marginal or random-effects) is completely different. Depending on the model family (marginal or random-effects), one is led to either marginal or hierarchical inference. It is important to realize that in the general case the parameter resulting from a marginal model and from a random-effects model, say β^M and β^{RE} respectively, are different, even when the latter one is estimated using marginal inference. Some of the confusion surrounding this issue may result from the equality of these parameters in the very special linear mixed model case. When a random-effects model is considered, the marginal mean profile can be derived, but it will generally not produce a simple parametric form.

As an important example, consider our GLMM with logit link function, and where the only random effects are intercepts b_i. It can then be shown that the marginal mean $\boldsymbol{\mu}_i = E(Y_{ij})$ satisfies $h(\boldsymbol{\mu}_i) \approx \boldsymbol{X}_i\beta^M$ with

$$\frac{\beta^{RE}}{\beta^M} = \sqrt{c^2\sigma^2 + 1} > 1, \tag{3.23}$$

in which c equals $16\sqrt{3}/15\pi$. Hence, although the parameters β^{RE} in the generalized linear mixed model have no marginal interpretation, they do show a strong relation to their marginal counterparts. Note that, as a consequence of this relation, larger covariate effects are obtained under the random-effects model in comparison to the marginal model.

3.2.5 Conditional Models

Section 3.2.1 introduced the concept of conditional models as one where outcomes are modeled, conditional upon the value of other outcomes on the same unit. These other outcomes could encompass the entire set of measurements, like in a classical log-linear model (Agresti 2002), or a subset. A very specific class of conditional models are so-called transition models. In a transition model, a measurement Y_{ij} in a longitudinal sequence is described as a function of previous outcomes, or history $\boldsymbol{h}_{ij} = (Y_{i1}, \ldots, Y_{i,j-1})$ (Diggle, Heagerty, Liang and Zeger, 2002, p. 190). One can write a regression model for the outcome Y_{ij} in terms of $\boldsymbol{h}_{ij}$, or alternatively, the error term ε_{ij} can be written in terms of previous error terms. In the case of linear models for Gaussian outcomes, one formulation can be translated easily into another and specific choices give rise to well-known marginal covariance structures such as, for example, AR(1). The order of a transition model is the number of previous measurements that is still considered to influence the current one. A model is called stationary if the functional form of the dependence does not vary over time.

A particular version of a transition model is a stationary first-order autoregressive model for binary longitudinal outcomes, which follows a logistic-regression type model:

$$\text{logit}[P(Y_{ij} = 1|\boldsymbol{x}_{ij}, Y_{i,j-1} = y_{i,j-1}, \boldsymbol{\beta}, \alpha)] = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \alpha y_{i,j-1}. \tag{3.24}$$

Evaluating (3.24) to $y_{i,j-1} = 0$ and $y_{i,j-1} = 1$, respectively, produces the so-called transition probabilities between occasions $j-1$ and j. In this model, if there were no covariates, these would be constant across the population. When there are time-independent covariates only, the transition probabilities change in a relatively straightforward way with level of covariate. For example, a different transition structure may

apply to the standard and experimental arms in a two-armed clinical study. Extension to the second or higher orders is obvious.

4

Direct-Likelihood: Time to Leave Simplistic Methods Behind

In Chapter 3, the different missingness mechanisms have been discussed. An example of MCAR missingness is that a subject may move, their data may be lost due to an administrative mix-up, or they may simply tire of participating in the study. However, the reasons for missingness are not always easy to ascertain. For example, if a subject withdrew because they experienced a car accident, their outcome data might be considered MCAR, but perhaps should not be if the subject's treatment could have affected their ability to drive. As will be discussed in this chapter, methods like LOCF and CC are based on extremely strong assumptions about missingness and even the strong MCAR assumption does not suffice to guarantee that an LOCF analysis is valid.

Next, an example of MAR is a trial in which subjects are removed if their response has exceeded a pre-specified limit. Alternatively, subjects may quit the trial if they are either doing much better or significantly worse. Such scenarios are more common than MCAR. The MAR assumption implies that future behaviour for those who share the same past measurements and covariate values is on average identical whether or

not they drop out. This enables treatment effect to be estimated in longitudinal models without simultaneously modelling the cause of dropout.

As already pointed in Section 3.1.3, valid inference can be obtained under MAR through a likelihood-based analysis, without the need for modeling the dropout process. As a consequence, one can simply use, for example, linear or generalized linear mixed models as introduced in Section 3.2.1 (Verbeke and Molenberghs, 2000), without additional complication or effort. In this chapter, we will show that such an analysis not only enjoys much wider validity than the simplistic methods but in addition is easy to conduct, *without additional data manipulation* using such tools as the SAS procedures MIXED or NLMIXED, HLM4.0, the SPSS procedure MIXED, the SPlus functions lme and nlme and MLwiN, to name a few. Indeed, there is no reason to use *ad hoc* methods when direct-likelihood analyses can be implemented with standard software. A contribution of this chapter has also been published in Beunckens, Molenberghs and Kenward (2005).

The outline of this chapter is as follows. In Section 4.1, the commonly used simplistic methods are discussed, whereas Section 4.2 focusses on the use of direct-likelihood and its advantages compared to these *ad hoc* methods. The case of two measurements of a Gaussian outcome is considered in Section 4.3, in which the estimates of the mean at both time points are compared for both the simplistic methods and direct-likelihood approach. Further, we apply and compare these methods to both the orthodontic growth data and the first depression trial in Section 4.4.

4.1 Methods in Common Use

We will briefly review a few relatively simplistic methods, such as *complete case analysis* (CC), which restricts the analysis to those subjects for which all information has been measured according to protocol, and *last observation carried forward* (LOCF), for which the last observed measurement is substituted for values at later points in time that are not observed. So far, clinical trial practice has put a strong emphasis on such methods. Claimed advantages include computational simplicity, no need for a full longitudinal model (e.g., when the scientific question is in terms of the last observed measurement occasion only, that is, View 2b is adopted as introduced in Section 3.1.2) and, for LOCF, compatibility with the Intention-to-Treat (ITT) principle, (Schwartz and Lellouch, 1967; Pocock, 1983) since data on all patients randomized can be used.

4.1.1 Complete Case Analysis

A *complete case analysis* includes only those cases for analysis, for which all measurements - covariates and outcomes - were recorded (Verbeke and Molenberghs, 2000; Little and Rubin, 2002; Molenberghs and Verbeke, 2005). This method has obvious advantages. It is very simple to describe and since the data structure looks like if it resulted from a complete experiment, standard statistical software can be used without additional work. Further, since the entire estimation is done on the same subset of completers, there is a common basis for inference. Unfortunately, the method suffers from severe drawbacks. First, there can be a substantial loss of information, with adverse effects on precision and power, even if the frequency of missing data for single variables low. Further, such an analysis will only be representative for patients who remain on study and have complete data. A complete case analysis can have a role as an auxiliary analysis, especially if it relates to a scientific question. A final important issue about a complete case analysis is that it is only valid when the missingness mechanism is MCAR. Severe bias can result when the missingness mechanism is MAR but not MCAR. This bias can be positive or negative, as illustrated by Molenberghs *et al.* (2004).

4.1.2 Last Observation Carried Forward

A method that has received a lot of attention (Siddiqui and Ali, 1998; Verbeke and Molenberghs, 2000; Little and Rubin, 2002; Mallinckrodt *et al.*, 2003a,b; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007) is *last observation carried forward*. In the LOCF method, whenever a value is missing, the last observed value is substituted. For the LOCF approach, the MCAR assumption is necessary but not sufficient. This approach further assumes that subjects' responses would have been unchanged from the last observed value to the endpoint of the trial. This constant profile assumption seldom holds (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007). In a clinical trial setting, one might believe that the response profile *changes* as soon as a patient goes off treatment and plateaus thereafter. Therefore, carrying observations forward may bias estimates of treatment effects in either direction and will underestimate the associated standard errors (Heyting *et al.*, 1992; Gibbons *et al.*, 1993; Lavori *et al.*, 1995; Siddiqui and Ali, 1998; Verbeke and Molenberghs, 2000; Mallinckrodt *et al.*, 2001a,b; Molenberghs *et al.*, 2004; Beunckens *et al.*, 2005; Jansen *et al.*, 2006a). This method artificially increases the amount of information in the data by treating imputed and actually observed values on equal footing.

Despite its shortcomings, LOCF has been the longstanding method of choice for the primary analysis in clinical trials because of its simplicity, ease of implementation with standard software, and the belief that the potential bias from carrying observations forward leads to a "conservative" analysis. An analysis is deemed conservative when the treatment effect estimated is smaller in absolute value than the true one. However, examples of anti-conservative effect of LOCF are common (Little and Yau, 1996; Liu and Gould, 2002; Mallinckrodt *et al.*, 2004; Molenberghs *et al.*, 2004; Jansen *et al.*, 2006a), meaning an LOCF analysis can create the appearance of a treatment effect when none exists.

4.1.3 Available Case Analysis

In a traditional *available case analysis* (AC) (Little and Rubin, 2002), estimators are based on the subjects who have complete information available for a specific analysis, a subset that can change when different covariates or time points are considered. For example, a collection of such analyses could be the treatment-specific means at a series of designated measurement times. With increasing dropout over time, means later in the study would be calculated using fewer subjects than earlier means. If dropout is not MCAR, means at later measurement times would become increasingly biased.

4.2 Direct-likelihood Approaches When Data Are Incomplete

An alternative approach for handling missing data in a clinical trial setting is to use methods that are valid under the weaker MAR assumption (Verbeke and Molenberghs, 2000; Little and Rubin, 2002; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007), instead of the methods discussed in previous section for which the MCAR assumption, and more, is needed. Note that methods valid under MAR are also valid if data are MCAR, while the reverse does not hold. First, a comparison of MAR methods with the commonly used methods is provided using an artificial but insightful example. Next, these simplistic methods are contrasted with broadly valid and easy to implement direct-likelihood methods. We will also comment on alternatives such as multiple imputation and the expectation-maximization algorithm.

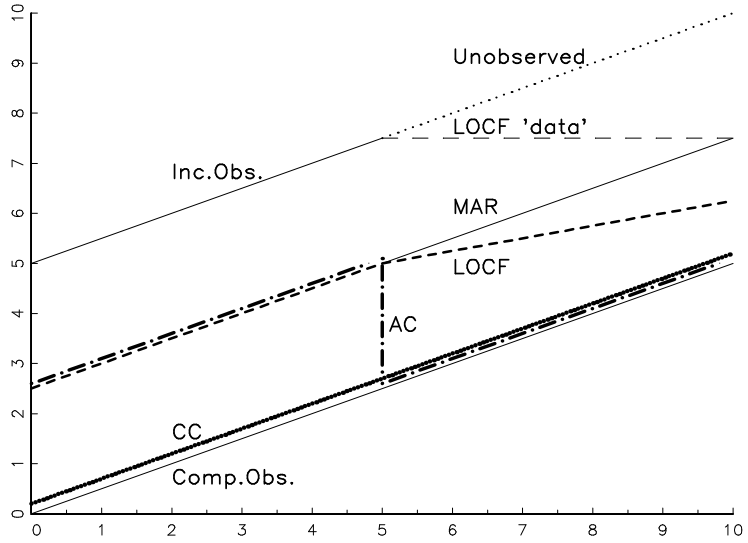


Figure 4.1: *Artificial example of a study where subjects drop out at time point 5 (x axis) after reaching a certain level of response (y axis). The dataset is composed of those with complete observations (bottom thin line) and those who have incomplete observations (top thin line). The estimated trajectories of this cohort using different analytic approaches to handle the incomplete data are shown. (MAR - middle bold line, LOCF - middle bold dashed line, AC - Bold dash-dot line, CC bottom bold line). The MAR line represents the correct result. The upper two lines show what would have happened to the subjects who dropped out; their actual unobserved average (dotted line) and the average assumed by the LOCF approach (long dash line).*

4.2.1 MAR Versus Commonly Used Methods

Let us take a look at an artificial but revealing example contrasting an MAR approach with the ones discussed earlier. Figure 4.1, displays the results of the traditional MCAR methods – complete case, available case and LOCF – with the result of an MAR method. In this example, the mean response is supposed to be a linear function of the variable on the abscissa. The slope is the same for patients with incomplete data and those with complete observations, but intercepts differ. We assume that patients with incomplete observations drop out half way through the study (time point 5) upon reaching a certain level of the response - an MAR missingness mechanism. Using a method valid under the MAR assumption, the analysis would yield the correct mean profile, that is, a straight line centered between the mean profiles of the completers and noncompleters. If one performed a complete case analysis, the fitted profile will

coincide with the mean profile of the complete cases (bold line). In Figure 4.1, the data that arise after carrying the last observed measurement forward is represented by a dashed line and clearly, under LOCF, a progressively increasing underestimate of the true mean is observed (bold dashed line). Finally, this figure shows how the AC approach (bold dash-dot line) can produce anomalous results in this situation; the trajectory becomes discontinuous at time point 5, with a mean identical to those who continue beyond that point. All of the MCAR methods produce incorrect results under this simple but plausible scenario.

4.2.2 Direct-likelihood Versus Commonly Used Methods

For longitudinal studies with missing data, a likelihood-based mixed-effects model only requires that missing data are MAR. These mixed-effects models permit the inclusion of subjects with missing values at some time points (both dropout and intermittent missingness). For the continuous-outcome setting, this amounts to the general linear mixed model, introduced in Section 3.2.1 (see also Verbeke and Molenberghs, 2000), which can be viewed both in the marginal and hierarchical framework. However, for the non-Gaussian case, such mixed-effect models are restricted to the random-effects model family. Here, we focus on the generalized linear mixed model as discussed in Section 3.2.3. On the other hand, as shown in Section 3.2.2, the weighted version of the frequentist GEE approach, WGEE, is a viable alternative within the marginal model framework, which is valid under MAR but does require to model the dropout process.

Such a likelihood-based MAR analysis is also termed likelihood-based ignorable analysis, or a direct-likelihood analysis. In the literature, names for these methods vary, and include hierarchical models, random-effects models, and random-coefficient models. In a direct-likelihood analysis, the observed data are used without deletion nor imputation. In doing so, appropriate adjustments valid under MAR are made to parameters at times when data are incomplete, due to the within-patient correlation. Even when interest lies in a comparison between the two treatment groups at the last measurement occasion, such a full longitudinal analysis is a good approach, since the fitted model can be used as the basis for inference.

In many clinical settings the repeated measures are balanced, in the sense that a common (and often limited) set of measurement times is considered for all subjects, which allows the a priori specification of a "saturated" model, such as, for example, a full group-by-time interaction for the fixed effects combined with an unstructured covariance matrix. For continuous outcomes, such a model specification is termed

Mixed-effects Model Repeated-Measures analysis (MMRM) by Mallinckrodt, Clark and Stacy (2001a,b). MMRM is a particular form of a linear mixed model, relevant for confirmatory clinical trials, fitting within the direct-likelihood paradigm. It has to be noted that this approach, for the special case where no dropout occurs, is fully equivalent to a one-way multivariate analysis of variance (MANOVA) analysis for repeated outcomes, with a class variable treatment effect. This observation provides a strong basis for such an approach, which is a very promising alternative for the simplistic ad hoc methods such as CC analysis or LOCF, described in Section 4.1.

These arguments, supplemented with the availability of software tools within which such multivariate models can be fitted to incomplete data, cast doubts regarding the usefulness of such simplistic methods as CC and LOCF. This issue has been discussed in detail, in the context of Gaussian outcomes on the one hand, by Molenberghs *et al.* (2004), and in the context of non-Gaussian outcomes on the other hand, by Jansen *et al.* (2006a). Apart from biases, as soon as the missing data mechanism is not MCAR, CC can suffer from severe efficiency losses. Especially since tools have become available to include incomplete sequences along with complete ones into the analysis, one should do everything possible to avoid wasting patient data. Next, LOCF, as other imputation strategies (Dempster and Rubin, 1983; Little and Rubin, 2002) can lead to artificially inflated precision. Further, as Molenberghs *et al.* (2004) have shown, the method can produce severely biased treatment comparisons and, perhaps contrary to some common belief, such biases can be conservative but also liberal. The method rests on the strong assumption that a patient's outcome profile remains flat, at the level of the last observed measurement, throughout the remainder of follow up. As a justification for the use of LOCF, proponents sometimes stated that it is a preferred approach when the ITT-principle is adhered to since data on all patients randomized can be used. However, direct-likelihood methods also use information on all subjects, including information from early dropouts, while avoiding the much stronger assumptions required to make LOCF valid. Thus, the direct-likelihood method is a sensible approach under ITT.

4.2.3 Alternatives to Direct-likelihood

There are a number of alternatives to direct-likelihood. One of these is multiple imputation (MI) (Rubin, 1987; Schafer, 1999; Little and Rubin, 2002). The MI method involves constructing a number of complete data sets from an incomplete one by drawing from the conditional distribution of the unobserved outcomes, given the observed ones. These data sets are then analysed and the results combined to produce

inferences. Verbeke and Molenberghs (2000) discuss the method in the context of continuous longitudinal data. Molenberghs and Verbeke (2005) illustrate how the SAS procedures MI and MIANALYZE can be used in this context. Multiple imputation is valid under the same conditions as direct-likelihood, and therefore does not suffer from the problems encountered in most single imputation methods. However, there are a number of situations where multiple imputation is particularly useful. For example, when outcomes as well as covariates are missing then multiple imputation is a sensible route. The method is also useful when several analyses, perhaps conducted by different analysts, have to be done on the same set of incomplete data. In such a case, all analyses could start from the same set of multiply-imputed sets of data and enhance comparability. Multiple imputation will be discussed in more detail in Chapter 5.

Another method that has seen a number of applications is the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977; Little and Rubin, 2002). Broadly speaking, the algorithm is a general method to fit a likelihood to incomplete data. When used as an alternative to the direct-likelihood method, it should give the exact same estimates, but computations are more difficult. Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005) show how the method can be used in SAS. For most standard longitudinal clinical trial settings, we recommend that direct-likelihood be the first choice.

4.3 Estimates in Case of Two Measurements

Using the simple setting of two repeated follow-up measures, the first of which is always observed while the second can be missing, we establish some properties of the LOCF and CC estimation procedures, both assuming MCAR, as well as the estimation procedure when the missingness mechanism is assumed to be MAR.

Let us assume each subject $i = 1, \ldots, N$ in the study is to be measured on two occasions. The responses are grouped in a 2-component vector (Y_{i1}, Y_{i2}). Assume a linear mixed model, with constant mean for both time points, and an unstructured variance-covariance matrix:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ & \sigma_2^2 \end{pmatrix}\right).$$

Further, assume the first r of these subjects complete the study, while for the remaining ones only the first measurement is observed. To obtain an explicit expression for the likelihood-based estimators, along the lines of Little and Rubin (2002), observe

Table 4.1: *Estimates of $\widehat{\mu}_1$ and $\widehat{\mu}_2$: MAR (all incomplete data under MAR), CC (complete cases only), versus LOCF (LOCF imputed data). Note that $\overline{y}_1^{(N)}$ represents $\frac{1}{N}\sum_{i=1}^{N} y_{i1}$, and analogous for $\overline{y}_1^{(r)}$, $\overline{y}_1^{(N-r)}$, $\overline{y}_2^{(N)}$, and $\overline{y}_2^{(r)}$.*

Estimate	MAR	CC	LOCF
$\widehat{\mu}_1$	$\overline{y}_1^{(N)}$	$\overline{y}_1^{(r)}$	$\overline{y}_1^{(N)}$
$\widehat{\mu}_2$	$\overline{y}_2^{(r)} + \frac{N-r}{N}\,\widehat{\beta}_1\,\left(\overline{y}_1^{(N-r)} - \overline{y}_1^{(r)}\right)$	$\overline{y}_2^{(r)}$	$\overline{y}_1^{(N)} + \frac{r}{N}\left(\overline{y}_2^{(r)} - \overline{y}_1^{(r)}\right)$

that conditional on the first observation, the second measurement will also be normally distributed with mean linearly related to the value of the first observation y_{i1} and with variance $\sigma^2_{2|1}$:

$$Y_{i2}|Y_{i1} = y_{i1} \sim N(\beta_0 + \beta_1 y_{i1}, \sigma^2_{2|1}),$$

where

$$\begin{cases} \beta_1 &= \rho\frac{\sigma_2}{\sigma_1}, \\ \beta_0 &= \mu_2 - \beta_1\mu_1 = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1, \\ \sigma^2_{2|1} &= \sigma_2^2(1-\rho^2). \end{cases}$$

The estimates of the mean parameters, $\widehat{\mu}_1$ and $\widehat{\mu}_2$, using either the CC or LOCF method, or a method valid under the MAR assumption, are listed in Table 4.1.

Note that, under LOCF, a correction is taking place without any adjustment for the correlation between the two measurements, whereas it is only correct in the unlikely case of correlation exactly equal to one and means constant across measurement occasions. Thus, LOCF would be inappropriate, and dramatically so in the zero correlation situation. Indeed, when there is no correlation between the first and second measurements, the regression coefficient $\beta_1 = 0$, and hence there is no correction under MAR. Unless the missingness mechanism is assumed to be MCAR, the means obtained under CC at both measurement times are incorrect, even though there is no need for a correction at the first one. The direct-likelihood method uses the difference between the means for complete and incomplete observations at time one, modified by the correlation between the two measurement occasions, to correct the mean at the second occasion. Note that, in practice, the estimators do not have to

be derived explicitly. While they are insightful, standard likelihood-based analyses using standard software will automatically ensure corrections of this type are used. Thus, for example, it would be sufficient to estimate the parameters in using the SAS procedure MIXED, as long as complete and incomplete sequences are passed on to the procedure. Further, note that the coefficient β_1 depends on the variance components implying that a misspecified variance structure may lead to bias in $\hat{\mu}_{2,MAR}$. Thus, the well-known independence between the distributions of the estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in multivariate normal population holds, once again, only when the data are balanced and complete.

4.4 Examples

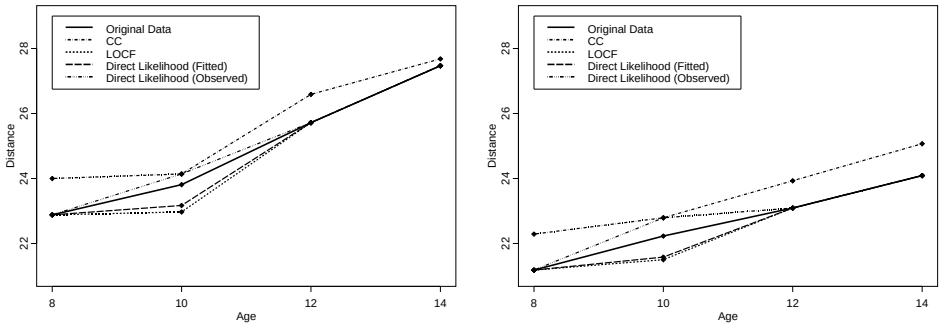
Let us apply the commonly used but simplistic methods, CC and LOCF, as well as MAR methods to both the orthodontic growth data and the first depression trial, introduced in Section 2.1 and 2.2 respectively. Since the orthodontic growth data is a study with longitudinal continuous measurements, we use the dichotomized outcome for the first depression trial. In this way, a comparison of MAR and simplistic methods is provided within the Gaussian and the non-Gaussian setting.

4.4.1 Orthodontic Growth Data

Let us first compare the simplistic methods and direct-likelihood method from Sections 4.1 and 4.2 using the orthodontic growth data. We analysed the original data, next to the CC data, the LOCF data, and the incomplete data as such. For this purpose, a linear mixed model is used, assuming unstructured mean, that is, assuming a separate mean for each of the eight age×sex combinations, together with an unstructured covariance structure, and using maximum likelihood (ML) as well as restricted maximum likelihood (REML). The mean profiles of the linear mixed model using maximum likelihood for all four data sets are given in Figure 4.2 for boys and girls separately. Next to this longitudinal approach, we will consider a MANOVA analysis and an ANOVA analysis per time point. For all these analyses, Table 4.2 shows the estimates and standard errors for boys at age 8 and 10, for the original data and all available incomplete data, as well as for the CC and the LOCF data.

First, we consider the group means for both sex groups for the original data set in Figure 4.2 (solid lines). Since we observe relatively straight lines both in left and right panel, there clearly seems to be a linear trend in both profiles.

In a complete case analysis of the growth data, the 9 subjects which lack one



(a) Boys (b) Girls

Figure 4.2: *Orthodontic growth data. Profiles for the Growth data using the original data, CC, LOCF and direct-likelihood method.*

measurement are deleted, resulting in a working data set with 18 subjects. This implies that 27 available measurements will not be used for analysis, a severe penalty on a relatively small data set. Observing the profiles for the CC data set in Figure 4.2, all group means increase relative to the original data set but mostly so at age 8. The net effect is that the profiles overestimate the average length. For the LOCF data set, the 9 subjects that lack a measurement at age 10 are completed by imputing the age 8 value. It is clear that this procedure will affect the linear but non-constant trend found for the original data set. Indeed, the imputation procedure forces the means at ages 8 and 10 to be very similar, thereby destroying the linear relationship. Hence, a simple, intuitively appealing interpretation of the trends is made impossible.

In case of direct-likelihood, we now see two profiles. One for the observed means (based on the available sample at each point in time) and one for the fitted means. These two coincide at all ages except age 10. At first sight, this is confusing because our model is a seemingly saturated one. However, the well-known fact that a saturated time-by-treatment group model reproduces the observed means is true only when the data are balanced, in the sense that all subjects have measurements at exactly the same times. Missingness disturbs this designed balance. This is a strength of the likelihood method, since it takes a correction into account, based on the observed data of a subject with incomplete data (see Table 4.1). As mentioned earlier, the complete observations at age 10 are those with a higher measurement at age 8. Due to the within-subject correlation, they are the ones with a higher measurement at age 10 as well, and therefore the fitted likelihood model corrects in the appropriate direction.

Table 4.2: *Orthodontic growth data. Estimates and standard errors for boys at age 8 and 10 for direct-likelihood analysis, MANOVA and ANOVA per time point, considering the original complete data, all available incomplete data, the complete cases only (Complete Case Analysis), and the imputed LOCF data (Last Observation Carried Forward Analysis).*

Method	Boys at Age 8		Boys at Age 10	
Original Data				
Direct likelihood, ML	22.88	(0.56)	23.81	(0.49)
Direct likelihood, REML	22.88	(0.58)	23.81	(0.51)
MANOVA	22.88	(0.58)	23.81	(0.51)
ANOVA per time point	22.88	(0.61)	23.81	(0.53)
All Available Incomplete Data				
Direct likelihood, ML	22.88	(0.56)	23.17	(0.68)
Direct likelihood, REML	22.88	(0.58)	23.17	(0.71)
MANOVA	24.00	(0.48)	24.14	(0.66)
ANOVA per time point	22.88	(0.61)	24.14	(0.74)
Complete Case Analysis				
Direct likelihood, ML	24.00	(0.45)	24.14	(0.62)
Direct likelihood, REML	24.00	(0.48)	24.14	(0.66)
MANOVA	24.00	(0.48)	24.14	(0.66)
ANOVA per time point	24.00	(0.51)	24.14	(0.74)
Last Observation Carried Forward Analysis				
Direct likelihood, ML	22.88	(0.56)	22.97	(0.65)
Direct likelihood, REML	22.88	(0.58)	22.97	(0.68)
MANOVA	22.88	(0.58)	22.97	(0.68)
ANOVA per time point	22.88	(0.61)	22.97	(0.72)

As an aside, note that in the case of direct-likelihood, the observed average at age 10 coincides with the CC average, while the fitted average does not coincide with anything else. Indeed, if the model specification is correct, then a direct-likelihood analysis produces a consistent estimator for the average profile, as if nobody had dropped out. This effect might be obscured in small data sets due to large variability. In spite of the small-sample behavior encountered here, the validity under MAR and the ease of implementation are good arguments that favor this direct-likelihood analysis over other techniques.

Let us now compare the different methods by looking at Table 4.2, which shows the estimates and standard errors for boys at age 8 and 10, for the original data and all available incomplete data, as well as for the CC data and the LOCF data.

Table 4.2 shows some interesting features. Whenever the data are balanced, the means are the same regardless of which estimation method is used. Standard errors are asymptotically the same and even in a small sample like the one considered here, differences are negligible. In all four cases, a CC analysis gives an upward biased estimate, for both age groups. This is obvious, since the complete observations at age 10 are those with a higher measurement at age 8, as we have seen before. The LOCF analysis gives a correct estimate for the average outcome for boys at age 8. This is not surprising since there were no missing observations at this age. As noted before, the estimate for boys of age 10 is biased downwards.

When the observed, incomplete data are analysed, we see from Table 4.2 that direct-likelihood, which is valid under MAR, produces good estimates, which diverge from the (M)ANOVA analyses, which are valid only under MCAR and give an overestimation of the average of age 10 in this case. MANOVA effectively reduces to a complete case analysis and therefore also yields an overestimation of the average at age 8. ANOVA produces a frequentist available case analysis, with correct inferences only at measurement occasions with complete data. Once again, we observe that direct-likelihood overcorrects, leading to mean estimates that are slightly too small. This is not due to bias, but rather to small-sample variability. It underscores the necessity to correctly specify the variance-covariance structure, and to ensure that its parameters are estimated sufficiently accurate (Molenberghs and Kenward, 2007).

In conclusion, it is clear that for balanced, complete data, the multivariate normal model has similar to identical behaviour to the frequentist (M)ANOVA analyses. However, when fitted to incomplete data, the likelihood-based methods are more broadly valid since they only require the missing data mechanism to be MAR, rather than MCAR. This provides a strong justification for the direct-likelihood method.

4.4.2 First Depression Trial

A second example will be situated in the non-Gaussian setting. To this end, we analyse the dichotomized response from the first depression trial.

The primary null hypothesis is zero difference between the experimental and the standard treatment in terms of proportion of the $HAMD_{17}$ total score above the level of 7, that is, in terms of depression status (yes/no). This will be tested using both marginal models (GEE and WGEE) and random-effects models (GLMM). According to the study protocol, the models include the fixed effects of treatment, visit, and treatment-by-visit interaction, all three considered as categorical covariates for which the standard drug and the last visit are considered as references, as well as the fixed effects of baseline score, a continuous covariate, and its interaction by visit. A random intercept is included when considering the random-effects models. Analyses are implemented using the SAS procedures GENMOD and NLMIXED.

Missing data are be handled in three different ways: (1) imputation using LOCF, (2) deletion of incomplete profiles, leading to a CC, and (3) analysing the data as they are, consistent with ignorability (for GLMM and WGEE). First we focus on a fully longitudinal approach (View 1), comparing the results regarding overall treatment effect obtained from the marginal and random-effects models, whereafter we also consider analyses of the treatment effect at the last visit, that is, the planned occasion (View 2a).

View 1: Longitudinal Analysis

Marginal Models. First, let us consider the GEE approach. Within the SAS procedure GENMOD the exchangeable working correlation matrix is used. In many cases, we observe the empirically-corrected standard errors to be larger than the model-based ones. This is because model-based standard errors are the ones that would be obtained if the estimating equations would be true likelihood equations, that is, when the working correlation structure is correct. In such cases, likelihood inference enjoys optimality. However, since the working correlation structure is allowed to be misspecified, model-based standard errors will be biased and it is advisable to base conclusions on empirically-corrected standard errors.

An inspection of parameter estimates and standard errors as shown in Table 4.3 reveals that the interaction between treatment and time is non-significant. At first sight, this suggests model simplification. However, there are a few reasons to prefer a different route.

Table 4.3: *First depression trial. GEE and WGEE : parameter estimates, standard errors and p-values.*

	CC (GEE)			**LOCF (GEE)**			**MAR (GEE)**			**MAR (WGEE)**		
Effect	Est.	(S.E.)	*p*-value	Est.	(S.E.)	*p*-value	Est.	(S.E.)	*p*-value	Est.	(S.E.)	*p*-value
Intercept	-2.00	(0.85)	0.019	-1.78	(0.63)	0.005	-1.92	(0.84)	0.021	-2.09	(0.89)	0.019
Treatment	0.69	(0.40)	0.088	0.64	(0.32)	0.043	0.71	(0.38)	0.063	0.69	(0.39)	0.079
Visit 4	1.51	(1.17)	0.197	0.94	(1.14)	0.406	0.89	(1.23)	0.467	-0.67	(2.00)	0.737
Visit 5	-0.09	(1.37)	0.946	0.15	(0.93)	0.873	0.14	(1.21)	0.909	0.62	(1.41)	0.663
Visit 6	0.79	(1.07)	0.462	0.82	(0.74)	0.266	0.93	(1.04)	0.368	1.13	(1.29)	0.382
Visit 7	0.28	(1.03)	0.785	0.17	(0.66)	0.791	0.03	(1.01)	0.975	-1.34	(1.42)	0.344
Treatment × Visit 4	-0.62	(0.65)	0.337	-0.45	(0.59)	0.453	-0.47	(0.64)	0.467	-0.34	(1.09)	0.775
Treatment × Visit 5	-0.65	(0.54)	0.227	-0.50	(0.40)	0.213	-0.62	(0.50)	0.214	-0.54	(0.63)	0.394
Treatment × Visit 6	-0.67	(0.41)	0.104	-0.71	(0.31)	0.023	-0.88	(0.41)	0.032	-1.55	(0.62)	0.012
Treatment × Visit 7	-0.45	(0.37)	0.219	-0.29	(0.24)	0.243	-0.28	(0.37)	0.453	0.49	(0.70)	0.483
Baseline	0.08	(0.04)	0.070	0.10	(0.03)	0.003	0.08	(0.04)	0.068	0.10	(0.05)	0.042
Baseline × Visit 4	0.07	(0.06)	0.267	0.09	(0.06)	0.164	0.12	(0.07)	0.087	0.24	(0.11)	0.032
Baseline × Visit 5	0.11	(0.08)	0.140	0.07	(0.05)	0.183	0.09	(0.07)	0.167	0.05	(0.08)	0.542
Baseline × Visit 6	0.01	(0.06)	0.899	-0.01	(0.04)	0.858	0.01	(0.05)	0.882	0.03	(0.07)	0.698
Baseline × Visit 7	0.01	(0.05)	0.845	0.01	(0.03)	0.845	0.02	(0.05)	0.643	0.10	(0.08)	0.172

First, as stated before, a longitudinal model used in a regulatory, controlled environment is ideally sufficiently generally specified to avoid driving conclusions through models that are too simple. Sticking to a single, pre-specified model also avoids dangers associated to model selection (e.g., inflated type I errors), recently reported in the literature (Hjort and Claeskens, 2003).

Second, a general model allows for, as a by-product, assessment of treatment effect at the last planned occasion. Third, one can still assess the important null hypothesis of (1) no average treatment effect, and (2) no treatment effect at any of the measurement occasions. These tests have been conducted and are reported in Table 4.4.

Unless one has strong believe that the MCAR assumption holds, it is careful to consider WGEE to perform an analysis that is correct under MAR. In terms of fitting the model to the data, it implies that weights have to be constructed, based on the probability to drop out at a given time, given the patient is still in the study, given his or her past measurements, and given covariates. We restrict attention the the previous outcome and treatment indicator. The code is exemplified in Section 11.3. The result of fitting this logistic regression did not reveal strong evidence for a dependence on the previous outcome (estimate 0.0974, s.e. 0.3513), nor on the treatment allocation (estimate -0.0652, s.e. 0.3137).

Let us now turn to the results. Apart from the effect of treatment, visit and treatment-by-visit interaction, results of the CC, LOCF, and standard GEE analyses are similar (Table 4.3). However, there are differences with the weighted GEE version, in parameter estimates and standard errors. The differences in standard errors, which are often larger under WGEE, are explained by the fact that additional sources of uncertainty, due to missingness, are taken into account. From Table 4.4, the marginal models reveal non-significant treatment effect in all cases, for either the hypothesis of no treatment effect or the hypotheses of no average effect. Corresponding to the one degree-of-freedom tests, parameter estimates and standard errors can be estimated as well for the mean treatment effect. For conciseness, only empirically corrected standard errors are shown. A strong difference is observed between the WGEE and other cases. Since this is the only one valid under MAR, it is clear that there are dangers associated to too simplistic methods. Furthermore, in the LOCF case, the p-value of the joint treatment effect is larger compared to WGEE, whereas the one for the mean treatment effect is smaller. This contradicts a common

Table 4.4: *First depression trial. View 1. GEE, WGEE and GLMM. Tests for (1) the joint null hypothesis of no treatment effect at any of the time points and (2) the hypothesis of no average treatment effect.*

Analysis	(1) Joint effect	(2) Mean effect		
	p-value	Est.	(S.E.)	p-value
CC (GEE)	0.6010	0.21	(0.32)	0.5216
LOCF (GEE)	0.2684	0.25	(0.28)	0.3570
MAR (GEE)	0.3047	0.26	(0.27)	0.3355
MAR (WGEE)	0.1694	0.30	(0.39)	0.4413
CC (GLMM)	0.6660	0.31	(0.61)	0.6125
LOCF (GLMM)	0.3111	0.56	(0.58)	0.3323
MAR (GLMM)	0.3564	0.44	(0.48)	0.3569

belief that LOCF is conservative. Molenberghs *et al.* (2004) and Jansen *et al.* (2006a) have shown that both conservative and liberal behavior is possible.

Random-effect Models. To fit generalized linear mixed models, we use the SAS procedure NLMIXED, which allows fitting a wide class of linear, generalized linear, and non-linear mixed models. It relies on numerical integration. Not only different integral approximations are available, the principal ones being (adaptive) Gaussian quadrature, but it also includes a number of optimization algorithms. The difference between non-adaptive and adaptive Gaussian quadrature is that for the first procedure the quadrature points are centered at zero for each of the random-effects and the current random-effects covariance matrix is used as the scale matrix, while for the latter the quadrature points will be appropriately centered and scaled, such that more quadrature points lie in the region of interest (Molenberghs and Verbeke, 2005). We will use both adaptive and non-adaptive quadrature, with several choices for the number of quadrature points, to check the stability of the results over a variety of choices for these numerical choices.

Precisely, we initiate the model fitting using non-adaptive Gaussian quadrature, together with the quasi-Newton optimization algorithm (step 1). The number of quadrature points is left to be determined by the procedure, and all starting

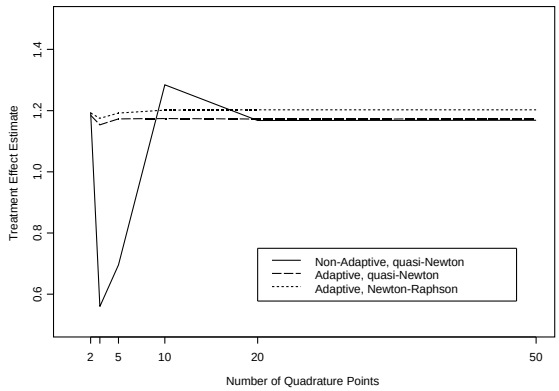


Figure 4.3: *First depression trial. The effect of adaptive versus non-adaptive quadrature, quasi-Newton versus Newton-Raphson, and the number of quadrature points on the treatment effect parameter.*

values are set equal to 0.5. Using the resulting parameter estimates, we keep these choices but hold the number of quadrature points fixed (2, 3, 5, 10, 20 and 50). Subsequently, we switch to adaptive Gaussian quadrature (step 2). Finally, the quasi-Newton optimization is replaced by the Newton-Raphson optimization (step 3). The effect of the method and the number of quadrature points is graphically represented in Figure 4.3 for a selected parameter (treatment effect). While the differences between these choices are purely numerical, we do notice differences between the results, illustrating that a numerical sensitivity analysis matters. The parameter estimates tend to stabilize with increasing number of quadrature points. However, non-adaptive Gaussian quadrature needs obviously more quadrature points than adaptive Gaussian quadrature.

Focusing on the results for 50 quadrature points, we have observed that the parameter estimates for step 1 and step 2 are only slightly different (order of 10^{-3}), whereas parameter estimates for step 3 differ in order of 10^{-1} compared to the previous steps (not shown). In spite of the differences in parameter estimates, is the noteworthy fact that the likelihood is the same in all steps, due to a flat likelihood.

This was confirmed by running all steps again, but now using the parameter estimates of step 3 as starting values, at which point the parameter estimates all coincide. Thus, it may happen that the optimization routine has only seemingly converged.

Table 4.5: *First depression trial. GLMM using adaptive Gaussian quadrature, Newton-Raphson optimization and 50 quadrature points : parameter estimates, standard errors and p-values.*

	CC			**LOCF**			**MAR**		
Effect	Est.	(S.E.)	p-value	Est.	(S.E.)	p-value	Est.	(S.E.)	p-value
Intercept	-3.78	(1.60)	0.020	-3.68	(1.43)	0.011	-3.51	(1.42)	0.014
Treatment	1.18	(0.80)	0.142	1.32	(0.72)	0.068	1.20	(0.72)	0.095
Visit 4	2.40	(2.21)	0.281	0.99	(1.82)	0.586	1.30	(1.88)	0.490
Visit 5	-0.37	(1.85)	0.842	-0.17	(1.53)	0.913	0.22	(1.61)	0.890
Visit 6	1.54	(1.61)	0.340	1.55	(1.36)	0.257	1.80	(1.49)	0.228
Visit 7	0.61	(1.57)	0.701	0.33	(1.36)	0.811	0.26	(1.51)	0.865
Treatment × Visit 4	-0.99	(1.08)	0.359	-0.64	(0.95)	0.504	-0.66	(0.95)	0.488
Treatment × Visit 5	-1.10	(0.92)	0.234	-0.94	(0.77)	0.229	-1.00	(0.81)	0.220
Treatment × Visit 6	-1.33	(0.82)	0.108	-1.56	(0.71)	0.029	-1.61	(0.77)	0.037
Treatment × Visit 7	-0.91	(0.81)	0.263	-0.64	(0.70)	0.357	-0.52	(0.76)	0.494
Baseline	0.15	(0.08)	0.072	0.21	(0.08)	0.006	0.15	(0.07)	0.051
Baseline × Visit 4	0.16	(0.13)	0.217	0.22	(0.11)	0.044	0.21	(0.11)	0.055
Baseline × Visit 5	0.23	(0.10)	0.032	0.18	(0.09)	0.045	0.17	(0.09)	0.061
Baseline × Visit 6	0.02	(0.08)	0.843	0.00	(0.07)	0.998	0.01	(0.08)	0.894
Baseline × Visit 7	0.02	(0.08)	0.824	0.02	(0.07)	0.792	0.03	(0.08)	0.664
σ	2.70	(0.39)	$< .0001$	3.11	(0.38)	$< .0001$	2.39	(0.32)	$< .0001$
-2ℓ		504.3			706.4			629.4	

Table 4.5 displays the results for the CC, LOCF and MAR analyses based on the GLMM using adaptive Gaussian quadrature with 50 quadrature points and Newton-Raphson optimization. Again similar results among CC, LOCF and the MAR analysis is observed, except for treatment, visit and the treatment-by-visit interaction.

Further, exactly as in the marginal model case, we assessed average treatment effect as well as treatment effect at any of the times. The results are reported in Table 4.4 as well. Again, the parameter p-values are different across methods as in the marginal model case, but all showing no significance.

Marginal *versus* Random-effects Models. In all cases, the variability of the random effect (standard deviation parameter σ) is highly significant. This implies that the GEE parameters and the random-effects parameters cannot be compared directly. If the conversion factor (3.23) is computed, then one roughly finds a factor of about 1.7 under MAR, 1.9 and 2.1 for the CC and LOCF analysis respectively. We note that this factor is not reproduced when directly comparing the two sets of estimates (Table 4.3 and 4.5). This is due to the fact that (3.23) operates at the true population parameter level, while we only have parameter estimates at our disposition. Since many of the estimates are not or only marginally significant, it is not unexpected to observe deviations from this relationship, even though the general tendency is preserved in most cases.

View 2: Single Time Point Analysis

When emphasis is on the last measurement occasion, LOCF and CC are straightforward to use. When the last observed measurement is of interest, while a different scientific question, the analysis is not different from the one obtained under LOCF but, of course, in this case CC is not an option.

Since the outcome is a dichotomous response, the data can be summarized in a $2\times k$ table, where k represents the number of treatments. The analysis essentially consists of comparing the proportions of success or failure in all groups. For this purpose, both Pearson's chi-squared test (Agresti, 1990) and Fisher's Exact test (Freeman and Halton, 1951) will be used. Nevertheless, it is still possible to obtain inferences from a full longitudinal model in this context. When an ignorable analysis is considered, one has to explicitly consider all incomplete profiles, in order to correctly incorporate all information available. Thus, one has to consider a longitudinal model. Both the marginal standard or weighted GEE analysis and the random-effects GLMM approach are considered. Table 4.6 shows a summary of the results in terms of p-values.

Table 4.6: *First depression trial. Views 2a : p-values for treatment effect at the last visit based on two longitudinal analyses (GLMM, (W)GEE), and two single time point analyses (Pearson's Chi-squared test, Fisher's Exact test).*

Method	Model	p-value
CC	GEE	0.0876
	GLMM	0.1424
	Pearson's Chi-squared Test	0.1506
	Fisher's Exact Test	0.1781
LOCF	GEE	0.0428
	GLMM	0.0676
	Pearson's Chi-squared Test	0.0851
	Fisher's Exact Test	0.0914
MAR	GEE	0.0633
	WGEE	0.0785
	GLMM	0.0949

Both endpoint analyses (that is, using the last available measurement) show insignificant treatment effect under the CC and LOCF method. The same holds for the GLMMs considering the three methods. However, notice a smaller p-value for LOCF compared with MAR on the one hand, and a larger p-value for CC compared with MAR on the other hand. The results from GEE under the CC analysis and from the WGEE analysis yield again a non-significant treatment effect, whereas now GEE under the LOCF analysis does show a significant borderline effect.

4.5 Conclusion

In this chapter, we compared the simplistic methods commonly used (CC, LOCF, AC) to analyse incomplete longitudinal data, against a direct-likelihood analysis. Such a direct-likelihood analysis uses all available information, without the need either to delete nor to impute measurements or entire subjects. It is theoretically justified

whenever the missing data mechanism is MAR, a less restrictive and more realistic assumption than MCAR, which is necessary (but not always sufficient) for simplistic analyses (AC, CC, LOCF). There is no distortion in the statistical information, since observations are neither removed (such as in CC analysis) nor added (such as in LOCF analysis). Indeed, an ignorable direct-likelihood analysis takes all information into account, not only from complete observations, but also from incomplete ones, through the conditional expectation of the missing measurements given the observed ones. In this chapter, this has been documented in the case of two measurements of a Gaussian outcome, by considering the estimates of the mean at both time points.

Further, we have exemplified the *ad hoc* methods and the direct-likelihood approach in both the Gaussian and the non-Gaussian setting. For continuous longitudinal data, the main mode of analysis is the likelihood-based linear mixed model (LMM). In the non-Gaussian case, the choice has to be made between marginal models and random-effects models. Generalized linear mixed models (GLMM) are a well-known set of random-effects model, which are likelihood-based. On the other hand, generalized estimating equations (GEE) is a non-likelihood marginal model, for which the stronger MCAR assumption is required. However, GEE can be extended to weighted GEE, making it also valid under MAR. To perform these direct-likelihood analyses (LMM or GLMM), standard software can be applied, and no additional programming is involved. For WGEE, a small amount of programming is necessary, which is easily done in standard software. These arguments justify a shift from the simplistic methods towards a direct-likelihood paradigm as primary analysis when analysing incomplete data from longitudinal clinical trials. In Sections 11.1–11.3 we illustrate how to perform simple analyses, direct-likelihood and WGEE using the SAS software.

5

Multiple Imputation and Weighting

In Chapter 3, three model families to analyse incomplete longitudinal data in a non-Gaussian setting were introduced. The random-effects family is represented by the commonly used generalized linear mixed effects model (GLMM), for which estimation is performed through maximum likelihood, implying that ignorability under MAR can be invoked. However, this is not the case for non-likelihood marginal models, such as the semi-parametric method of generalized estimating equations (GEE), which is a second prevalent modelling approach besides GLMMs. As pointed in previous chapters, such models are only valid under the restrictive assumption of MCAR. To ensure validity under MAR, Robins, Rotnitzky and Zhao (1995) proposed weighted generalized estimating equations (WGEE) as discussed in Section 3.2.2.

An alternative approach to handle MAR missingness when using GEE, as suggested by Schafer (2003), would be based on multiple imputation, a technique developed by (Rubin, 1987) and introduced in Section 4.2.3 as an alternative to direct-likelihood. This approach consists of multiply imputing the missing outcomes using a parametric model, followed by analysing the resulting complete datasets using a standard method. In case GEE is considered as the standard method, we refer to this combination of MI and GEE as 'MI-GEE'. Afterwards, the obtained inferences

are combined into a single one. Regarding the missingness process, standard multiple imputation requires MAR to hold, even though extensions exist.

In this chapter, the focus will be on the comparison between both GEE versions for incomplete data: WGEE and MI-GEE. Comparisons will be made by means of a simulation study, including both small-sample simulations, as well as so-called asymptotic simulations (Rotnitzky and Wypij, 1994). The behaviour of both methods in terms of mean squared error (MSE), variance and bias of the estimators will be studied, under correctly specified and misspecified models. In this way, robustness of both methods under misspecification of either the dropout model, the imputation model, or the measurement model, can be explored.

The outline of this chapter is as follows. In Section 5.1, we discuss methods for analysing incomplete longitudinal non-Gaussian data, which are valid under the MAR assumption, with main attention on WGEE and multiple imputation together with GEE as analysis method. A description of the asymptotic and small-sample simulation design, as well as the results of the simulation study, is provided in Section 5.2. Finally, we apply both approaches to the first depression trial data in Section 5.3. The contribution of this chapter is joint work with Cristina Sotto and has been published in Beunckens, Sotto and Molenberghs (2007b).

5.1 Non-Gaussian Incomplete Longitudinal Data and MAR

While full likelihood methods are appealing because of their flexible ignorability properties, their use for non-Gaussian outcomes can be problematic due to prohibitive computational requirements. Therefore, GEE is an attractive alternative within the marginal model family. Since GEE is based on frequentist considerations, the missing data mechanism needs to be MCAR for it to be ignorable. This motivates the use of *weighted generalized estimating equations* (Robins, Rotnitzky and Zhao, 1995), which is valid under the weaker MAR missingness mechanism. An alternative mode of analysis, proposed by Schafer (2003), consists of multiply imputing the missing outcomes using a full-parametric model, e.g., of a random-effects or conditional type, followed by analysis of the resulting completed data sets using a conventional marginal (e.g., GEE) or conditional model (e.g., a transition model), and finally performing multiple-imputation inference. This results in so-called *multiple imputation based generalized estimating equations* (MI-GEE) or *multiple imputation based transition model* (MI-

Transition), when respectively GEE or a transition model is used for the analysis of the completed data sets.

Since the marginal model GEE and its extension to WGEE have been outlined in Section 3.2.2, the focus in this section lies on the clarification of MI-GEE and MI-Transition. The main idea of multiple imputation has been described in Section 4.2.3, and a more detailed discussion is given below.

Multiple imputation (MI) was formally introduced by Rubin (1978). The key idea of the procedure is to first replace each missing value with a set of M plausible values drawn from the conditional distribution of the unobserved values, given the observed ones. This conditional distribution represents the uncertainty about the right value to impute. In this way, M imputed data sets are generated (imputation stage), which are then analysed using standard complete data methods (analysis stage). Finally, the results from the M analyses have to be combined into a single inference (pooling stage) by means of the method laid out in Rubin (1978). In its basic form, multiple imputation requires the missingness mechanism to be MAR, even though versions under MNAR have been proposed (Rubin, 1987; Molenberghs, Kenward and Lesaffre, 1997).

In line with the notation in Section 3.1, suppose the parameter vector of the distribution of the response $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$ is denoted by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})'$, in which $\boldsymbol{\beta}$ denotes the vector of fixed-effects parameters and $\boldsymbol{\alpha}$ the vector of covariance parameters. Multiple imputation uses the observed data $\boldsymbol{Y}^o$ to estimate the conditional distribution of $\boldsymbol{Y}^m$ given $\boldsymbol{Y}^o$. The missing data are sampled several times from this conditional distribution and augmented to the observed data. The resulting completed data are then used to estimate $\boldsymbol{\theta}$. If the distribution of $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$ were known, with parameter vector $\boldsymbol{\theta}$, then $\boldsymbol{Y}_i^m$ could be imputed by drawing a value of $\boldsymbol{Y}_i^m$ from the conditional distribution $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{\theta})$. The objective of the imputation phase is to sample from this true predictive distribution. However, $\boldsymbol{\theta}$ in the imputation model is unknown, and therefore needs to be estimated from the data first, say $\widehat{\boldsymbol{\theta}}$, after which $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \widehat{\boldsymbol{\theta}})$ is used to impute the missing data. Precisely, this implies one first generates draws from the distribution of $\widehat{\boldsymbol{\theta}}$, thereby taking sampling uncertainty into account. Generally, the parameter vector in the imputation model differs from the parameter vector that governs the analysis model. Alternatively, a Bayesian approach, in which uncertainty about $\boldsymbol{\theta}$ is incorporated by means of some prior distribution for $\boldsymbol{\theta}$, can also be taken.

In the last phase of multiple imputation, the results of the analyses for the M imputed data sets are pooled into a single inference. The combined point estimate for the parameter of interest from the multiple imputation is simply the average of

the M complete-data point estimates (Schafer, 1999). Let $\boldsymbol{\theta}$ denote the parameter of interest, then the estimate and its estimated variance are given by:

$$\widehat{\boldsymbol{\theta}} \equiv \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\theta}}^{\,m} \quad \text{and} \quad \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) \equiv \boldsymbol{V} = \boldsymbol{W} + \left(\frac{M+1}{M}\right)\boldsymbol{B},$$

where

$$\boldsymbol{W} = \frac{\sum_{m=1}^{M}\boldsymbol{U}^{\,m}}{M} \quad \text{and} \quad \boldsymbol{B} = \frac{\sum_{m=1}^{M}(\widehat{\boldsymbol{\theta}}^{\,m} - \widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^{\,m} - \widehat{\boldsymbol{\theta}})'}{M-1},$$

with $\boldsymbol{W}$ denoting the average *within* imputation variance and $\boldsymbol{B}$ the *between* imputation variance (Rubin, 1987).

Since in WGEE all subjects are given weights, calculated using the hypothesized dropout model, any misspecification of this dropout model will affect all subjects, and thus the results. On the other hand, one can consider MI together with GEE or with a transition model, resulting in MI-GEE and MI-Transition, respectively. In essence, this method comes down to first using the predictive distribution of the unobserved outcomes given the observed ones and perhaps covariates. After this step, the missingness mechanism can be further ignored, provided it is MAR. In these MI cases, a misspecification made in the imputation step will only effect the unobserved (that is, imputed) but not the observed part of the data. Meng's (1994) results show that, as long as the imputation model is not grossly misspecified, this approach will perform well. Considering all this, one might be inclined to expect the MI-GEE or MI-Transition to be more robust against model misspecification than WGEE. In the next section, we will use a simulation study to investigate this idea.

5.2 A Simulation Study

In the previous section, we pointed to various methods to overcome the bias occurring in GEE under MAR. WGEE is unbiased for a correctly specified dropout and mean structure of the measurement model. MI-GEE requires compatibility between the imputation and estimation model to be correctly specified. Therefore, it is of interest to quantify the bias and precision under various types of misspecification. To this end, an asymptotic simulation study, as well as small-sample simulations, are conducted on various underlying data-generating models. Whereas asymptotic simulations give a nice paradigm to explore the situation of "large" samples, small-sample simulations give insight into the behaviour of the methods in real-life settings.

In the simulation study, we distinguish between two stages: (1) the data-generating stage and (2) the analysis stage. In the first stage, a data-generating model is defined.

Under the selection model framework, this generating model consists of a measurement model on the one hand, and a dropout model, given the measurement model on the other. In the analysis stage, a distinction should be made among three types of models: a measurement model, a dropout model and an imputation model. For the WGEE approach, only a marginal measurement model and a dropout model need to be specified. In contrast, the analysis stage for MI-GEE would entail the specification of an imputation model, rather than a dropout model, as well as a marginal measurement model. Finally, for MI-Transition, a conditional rather than marginal measurement model is needed, as well as an imputation model.

To assess the distinctive and relative merits of the methods of interest, we consider their performance, first in the case without any misspecification, then under various misspecifications. Since interest lies in comparing WGEE and MI-GEE as methods for dealing with missing data in a binary longitudinal setting, the misspecification can be made either in the dropout model, in the imputation model, or in the measurement model. Misspecification in the missingness mechanism, however, e.g., using MCAR for an underlying MAR mechanism, is not further explored, as this is not the main focus here and has already been investigated extensively (Jansen *et al.*, 2006a).

First, the various generating models employed for the simulations are defined in Section 5.2.1. Section 5.2.2 is devoted to the description of the design of the simulation study, after which the results of the simulation, under each of the various scenarios, are presented in Section 5.2.3.

5.2.1 Data-generating Models

For the simulation study, we generated an outcome at 3 time points using three different measurement models: first, three-dimensional binary outcomes were generated from a Bahadur model, second from a second-order autoregressive, AR(2), transition model, and finally, a three-dimensional continuous outcome (that was later dichotomized) was generated from a trivariate Gaussian distribution. Whereas the choice of the first two is obvious, since our focus lies on binary repeated measures, the third case depicts real-life settings for which a continuous outcome is available, but the scientific question is based on a dichotomized version of it. For all three cases, the measurement model incorporated a binary treatment indicator, such as a treatment *versus* placebo classification. In addition, for the dropout model, an MAR mechanism was considered. Assuming that dropout can occur only after the first time point, there are three possible dropout patterns: (1) dropout at the second time point, (2) dropout at the third time point, or (3) no dropout. The combination of the various

measurement models and the dropout model gives rise to three data-generating models, which will hereinafter be denoted as GM I (Bahadur measurement model and MAR dropout model), GM II (AR(2) measurement model and MAR dropout model) and GM III (Gaussian measurement model and MAR dropout model).

Note that we restrict the simulation setting to short sequences, since the higher-order Bahadur models would become prohibitive to generate from. Nevertheless, both the WGEE, as well as the MI-GEE methods, and then especially also the MI-Transition models, can be used, and in fact are very appealing, for longer sequences of repeated measures. When sequences become very long, the transition model is preferable owing to its computational convenience.

Let us now define these three data-generating mechanisms in turn. Denote by t_j the time point at which measurement j is taken and by x_i the treatment indicator. GM I is based on a Bahadur model, which follows general formulation (3.14), with

$$\text{logit}(\pi_{ij}) = \text{logit}[P(Y_{ij} = 1|x_i, t_j)] = \beta_0 + \beta_x\, x_i + \beta_t\, t_j + \beta_{xt}\, x_i\, t_j, \tag{5.1}$$

where we choose $\beta_0 = -0.25$, $\beta_x = 0.5$, $\beta_t = 0.2$ and $\beta_{xt} = -0.8$, with two- and three-way correlation coefficients equal to $\rho_{ij_1j_2} = 0.2$ and $\rho_{ij_1j_2j_3} = 0$, respectively. The latter define an exchangeable correlation structure. The missingness process for GM I is assumed to be MAR, and the probability of dropout at time point j given x_i and the measurement at the previous time point, is modelled by a logistic regression of the form

$$\text{logit}[P(D_i = j|x_i, y_{i,j-1}, D_i \geq j)] = \psi_0 + \psi_x\, x_i + \psi_{prev}\, y_{i,j-1},$$

where $j = 2, 3$, $\psi_0 = -0.5$, $\psi_x = -0.6$ and $\psi_{prev} = -3.5$. Combining this dropout model with the measurement model yields, for GM I, 68% completers, 15% with the last outcome missing (7% for $x = 0$ and 8% for $x = 1$), and 18% with only the first outcome observed (10% for $x = 0$ and 8% for $x = 1$).

The same dropout model is used to generate the missingness for GM II, but now combined with the AR(2) transition model. Such a model can be described as follows:

$$\begin{aligned}
P(x_i) &= \mu_x, \\
\text{logit}[P(Y_{i1} = 1|x_i)] &= \alpha_0 + \alpha_x\, x_i, \\
\text{logit}[P(Y_{i2} = 1|x_i, y_{i1})] &= \phi_0 + \phi_x\, x_i + \phi_1 y_{i1}, \\
\text{logit}[P(Y_{i3} = 1|x_i, y_{i1}, y_{i2})] &= \gamma_0 + \gamma_x\, x_i + \gamma_1\, y_{i1} + \gamma_2\, y_{i2},
\end{aligned}$$

where $\mu_x = 0.5$, $\alpha_0 = -0.2$, $\alpha_x = 0.3$, $\phi_0 = -0.1$, $\phi_x = 0.5$, $\phi_1 = 0.7$, $\gamma_0 = -0.25$, $\gamma_x = 0.35$, $\gamma_1 = 0.4$ and $\gamma_2 = 0.6$. For this generation model, the missingness

proportions are 73% for completers, 11% with the last outcome missing (7% for $x = 0$ and 4% for $x = 1$), and 17% with only the first outcome observed (11% for $x = 0$ and 6% for $x = 1$).

Since the methods of interest, WGEE and MI-GEE, involve marginal models, so as to allow comparison, the above conditional model needs to be further marginalized to obtain so-called marginalized "true" parameters, which then approximately describe a marginal logistic function. This marginalization assumes that the corresponding underlying marginal model is of the form given in (5.1). Inasmuch as the underlying measurement model is in fact conditional, rather than marginal, there is no way to verify whether this assumed underlying marginal model is "true". This marginalization was done by computing the marginal probabilities from the underlying conditional AR(2) transition model probabilities, that is, for a given outcome vector and treatment level, $(y_{i1}, y_{i2}, y_{i3}, x_i)$,

$$P(y_{i1}, y_{i2}, y_{i3}, x_i) = P(y_{i3}|x_i, y_{i1}, y_{i2})P(y_{i2}|x_i, y_{i1})P(y_{i1}|x_i)P(x_i). \tag{5.2}$$

On a hypothetical dataset consisting of all 16 possible combinations of the form $(y_{i1}, y_{i2}, y_{i3}, x_i)$, with corresponding probability weights $P(y_{i1}, y_{i2}, y_{i3}, x_i)$, we fitted a GEE model of the form (5.1). The resulting marginalized "true" parameters of GM II are $\beta_0 = -0.3658$, $\beta_x = 0.2673$, $\beta_t = 0.2265$ and $\beta_{xt} = 0.0790$.

Finally, for GM III, we assume a Gaussian outcome, W_{ij}, at three time points, where:

$$\mu_{ij} = E(W_{ij}|x_i, t_j) = \eta_0 + \eta_x\, x_i + \eta_t\, t_j + \eta_{xt}\, x_i\, t_j,$$

for $i = 0, 1$ and $j = 1, 2, 3$, with $\eta_0 = 3.5$, $\eta_x = 0$, $\eta_t = 1.75$ and $\eta_{xt} = 0.5$, yielding

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_0 \\ \boldsymbol{\mu}_1 \end{pmatrix} = \begin{pmatrix} (\mu_{01}, \mu_{02}, \mu_{03})' \\ (\mu_{11}, \mu_{12}, \mu_{13})' \end{pmatrix} = \begin{pmatrix} (5.75, 8.00, 10.25)' \\ (5.25, 7.00, 8.75)' \end{pmatrix}.$$

Moreover, we assume the following unstructured covariance structure:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.80 & 0.35 \\ 0.80 & 1 & 0.50 \\ 0.35 & 0.50 & 1 \end{pmatrix}.$$

The missingness process for this GM is given by:

$$\text{logit}[P(D_i = j|x_i, w_{i,j-1}, D_i \geq j)] = \delta_0 + \delta_x\, x_i + \delta_{prev}\, w_{i,j-1},$$

where $j = 2, 3, 4$, $\delta_0 = -0.15$, $\delta_x = 0.8$ and $\delta_{prev} = -0.35$. Combining this dropout model with the measurement model yields, on average, over all the 500 samples generated from GM III, 76% completers, 7% with the last outcome missing (3% for $x = 0$

and 4% for $x = 1$), and 17% with only the first outcome observed (7% for $x = 0$ and 10% for $x = 1$).

The binary outcome Y_{ij} was then obtained from the continuous outcome W_{ij} by defining a cut-off value of 6.5, that is, $Y_{ij} = 1$, if $W_{ij} \geq 6.5$, and 0, otherwise. Although the generated outcomes are continuous in nature, the focus here is on the analysis of the binary version Y_{ij}. For this reason, we need to obtain "true" parameters corresponding to this dichotomized response by fitting a GEE model of the form (5.1) to the 500 complete samples. Note however that this model is not necessarily the unknown underlying marginal model for the binary outcomes. The resulting parameters are $\beta_0 = -3.0373$, $\beta_x = 0.0095$, $\beta_t = 1.7812$ and $\beta_{xt} = 0.4828$.

The choice for linear time evolutions, at the scale of the linear predictor and within each of the treatment arms, allows us to distinguish between misspecification effects on cross-sectional parameters (β_0 and β_x), longitudinal parameters (β_t), and parameters combining aspects of both (β_{xt}). In practice, for example in a clinical trial, it might be advisable to allow for an unstructured, saturated treatment-by-time model, reducing the risk of model misspecification and in line with recommendations made by Molenberghs *et al.* (2004) and several references listed therein.

5.2.2 Design of the Simulation Study

We now proceed to describe the details of the simulation study. Given that the sequence of outcomes and the missing data process for GM I and GM II are discrete, quantification of this bias under specific assumptions about the nonresponse process can be done via an algorithm first proposed by Rotnitzky and Wypij (1994). This so-called asymptotic simulation method entails first creating a hypothetical data set consisting of all possible outcome sequences for each level of the covariate(s). In addition, for each of these, there are J possible missingness patterns. The probability mass with which each of these outcome sequences occurs can be computed based on the assumed data-generating model (measurement and dropout models).

For our case, we consider a binary outcome at 3 time points, denoted by $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})'$, and a binary treatment indicator, x_i, that is, a single covariate with 2 levels. This gives rise to $2^3 = 8$ possible sequences at each level of the covariate, yielding a total of 16 possibilities. From the assumed measurement model, the probability masses for each of these 16 sequences can be computed, $P(\boldsymbol{y}_i, x_i)$ say. Now, for each such case, there are 3 possible dropout patterns – dropout at second time point, dropout at the third time point, and no dropout – yielding a total of 48 possibilities. The probabilities $P(\boldsymbol{y}_i, x_i)$ are thus further split among the 3

missingness patterns according to the dropout probabilities. Specifically, denoting by $P(D_i = j|D_i \geq j, \boldsymbol{y}_i, x_i)$, $j = 2, 3, 4$, the probability of dropout at time point j, given the subject is still in the study, we obtain:

$$\begin{aligned}
P(\boldsymbol{y}_i, x_i, D_i = 4|D_i \geq 4) &= P(\boldsymbol{y}_i, x_i) \textstyle\prod_{j=2}^{4}[1 - P(D_i = j|D_i \geq j, \boldsymbol{y}_i, x_i)], \\
P(\boldsymbol{y}_i, x_i, D_i = 3|D_i \geq 3) &= P(\boldsymbol{y}_i, x_i) \textstyle\prod_{j=2}^{3}[1 - P(D_i = j|D_i \geq j, \boldsymbol{y}_i, x_i)] \\
&\quad \times P(D_i = 4|D_i \geq 4, \boldsymbol{y}_i, x_i), \\
P(\boldsymbol{y}_i, x_i, D_i = 2|D_i \geq 2) &= P(\boldsymbol{y}_i, x_i) \textstyle\prod_{j=3}^{4} P(D_i = j|D_i \geq j, \boldsymbol{y}_i, x_i) \\
&\quad \times [1 - P(D_i = 2|D_i \geq 2, \boldsymbol{y}_i, x_i)].
\end{aligned}$$

The estimating equations are then applied to this hypothetical data set with the application of the resulting probability weighting. The solutions obtained are the limiting (that is, asymptotic) solutions, which can then be compared with the known parameters of the simulation model, so as to conveniently derive the asymptotic bias of the estimators.

For the small-sample simulations, we assume a sample of size $N = 100$ subjects, equally divided between the two treatment groups. Based on the underlying probabilities from GM I or GM II, 50 observations were generated randomly for each treatment group. $S = 500$ such samples were then generated. Similarly, for GM III, we generated $S = 500$ samples with $n_0 = 50$ observations from $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $n_1 = 50$ observations from $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. While asymptotic simulations were conducted only for GM I and GM II, small-sample simulations were done for all three generation models. When using a GEE approach for analysis, the same working correlation structure as assumed during data generation is employed.

5.2.3 Results of the Simulation Study

For the ensuing discussion, in assessing and comparing WGEE and imputation-based GEE, various properties are quantified. First, we define bias as the difference between the estimate and the true value of the parameter, that is, $\mathrm{Bias}(\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. For the asymptotic version, probability weights, computed from the underlying GM, are applied in solving the estimating equations (Rotnitzky and Wypij, 1994). The resulting estimates are the limiting solutions, which can then be used to compute the asymptotic bias (Bias_∞), while the resulting variances are the asymptotic variances (Var_∞) of the parameter estimators.

For the small-sample simulations, the average ($\overline{\text{Est}}$) of the estimators over all $S = 500$ samples, its true variance for a sample of size N (Var_N), its estimated variance for a sample of size N ($\widehat{\text{Var}}_N$) and MSE are computed as:

$$\begin{aligned}
\overline{\text{Est}} &\equiv \sum_{i=1}^{S} \frac{\widehat{\boldsymbol{\beta}_i}}{S} \\
\text{Var}_N &\equiv \text{Var}_N(\overline{\text{Est}}) = \frac{\text{Var}_\infty}{N} \\
\widehat{\text{Var}}_N &\equiv \widehat{\text{Var}}_N(\overline{\text{Est}}) = \sum_{i=1}^{S} \frac{(\widehat{\boldsymbol{\beta}_i} - \overline{\text{Est}})^2}{S-1} \\
\text{MSE} &\equiv \text{MSE}(\overline{\text{Est}}) = \text{Bias}_N^2(\overline{\text{Est}}) + \widehat{\text{Var}}_N(\overline{\text{Est}})
\end{aligned}$$

Everything Correctly Specified

We first investigate the individual merits of each method when every one of its aspects is correctly specified. Recall that GM I is based on a Bahadur measurement model and a logistic model for dropout that is reflective of an MAR mechanism, that is, depending on the previous measurement as well as the treatment indicator. An appropriate analysis model would consist of a measurement model and a dropout model that match those of this GM. Since GEE methods are moment-based versions of the Bahadur model (Section 3.2.2), a GEE-based version, with the same structure as that of the underlying measurement model would be suitable. To address the MAR nature of the missingness, the GEE-based approach is supplemented with a weighting scheme, obtained from a model of the same form as that of the underlying dropout model, resulting now in WGEE. Thus, WGEE was fitted for GM I, using weights taken from fitting a logistic dropout model with the treatment indicator and the previous measurement as predictors. It should be noted that under WGEE the imputation model is not relevant since the missingness is addressed, not by imputation, but rather, by means of the dropout model. The results for both the asymptotic and small-sample simulations are shown in Table 5.1.

Clearly, the asymptotic unbiasedness of the WGEE estimators under a correctly specified mean structure is demonstrated by our asymptotic simulation. The same cannot be said, however, for the small-sample simulation, under which a substantial amount of bias is observed. Moreover, the estimated variances of the parameter estimators are considerably larger than the true variances, demonstrating the inefficiency of WGEE for small samples. These observations indicate that, for a sample of size $N = 100$, the consistency of the WGEE estimators does not seem to be achieved, at

Table 5.1: *Asymptotic and small-sample simulation results for WGEE, with everything correctly specified, under GM I. Asymptotic results include asymptotic bias ($Bias_\infty$) and asymptotic variance (Var_∞), while small-sample simulation results (for 500 simulations) include the average ($\overline{Est}$), bias ($Bias_N$), estimated variance ($\widehat{Var}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators, for $N = 100$.*

	Asymptotic		Small-Sample				
Parameter	Bias_∞	Var_∞	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
$\beta_0 = -0.25$	-1.87E-06	0.44095	-0.6457	-0.3956	1.0779	0.0044	1.2345
$\beta_x = 0.5$	1.99E-07	1.10959	0.6225	0.1225	2.1108	0.0111	2.1258
$\beta_t = 0.2$	2.02E-07	0.11942	0.3018	0.1018	0.2388	0.0012	0.2491
$\beta_{xt} = 0.8$	-1.66E-07	0.27815	-0.9355	-0.1356	0.4441	0.0028	0.4625

least not for this particular generating model.

For GM II, which uses an AR(2) transition model for the mean structure and a conditional logistic model for dropout, we considered fitting an AR(2) transition model, which is consistent with the underlying measurement model, after multiple imputation (MI-Transition). The multiple imputations are carried out with the SAS procedure MI, which employs a conditional logistic imputation model for binary outcomes, a model in line with the underlying measurement model of GM II and fully parametric, admitting valid inferences under MAR (Schafer, 2003). Thus, the analysis model, the imputation model, and the measurement model, are correctly specified. Note also that a dropout model need not to be defined for this mode of analysis, since imputations, rather than dropout weights, are used to cope with the missingness. For the asymptotic simulation, $M = 500$ datasets were imputed, while for the small-sample simulations, since efficient results can be obtained even under a small number of imputations (Rubin, 1987), we chose a more practically relevant value of $M = 5$. Table 5.2 gives the results for both types of simulations.

The first panel shows asymptotically unbiased parameter estimates, since, for this outcome, data for all subjects are assumed available and are thus not imputed. The estimates of the small-sample simulations for this outcome, on the other hand, show some bias as can be expected whenever taking finite samples. For the second and third panels, some bias is observed, asymptotically and for small samples, but the

Table 5.2: *Asymptotic and small-sample simulation results for MI-Transition, with everything correctly specified, under GM II. Asymptotic results include asymptotic bias ($Bias_\infty$) and asymptotic variance (Var_∞), while small-sample simulation results (for 500 simulations) include the average ($\overline{Est}$), bias ($Bias_N$), estimated variance ($\widehat{Var}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators, for $N = 100$.*

	Asymptotic		Small-Sample				
Parameter	Bias_∞	Var_∞	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
$\alpha_0 = $ -0.2	-0.0000	8.0803	-0.2313	-0.0313	0.0925	0.0808	0.0935
$\alpha_x = 0.3$	-0.0000	16.1003	0.3369	0.0369	0.1791	0.1610	0.1805
$\phi_0 = $ -0.1	-0.0096	12.0926	-0.0683	0.0317	0.2046	0.1209	0.2056
$\phi_x = 0.5$	-0.0666	18.0194	0.5041	0.0041	0.2635	0.1802	0.2635
$\phi_1 = 0.7$	0.0343	18.1493	0.7241	0.0241	0.2692	0.1815	0.2698
$\gamma_0 = $ -0.25	0.0236	17.4438	-0.1702	0.0798	0.3472	0.1744	0.3535
$\gamma_x = 0.35$	-0.0568	18.5632	0.3590	0.0090	0.3023	0.1856	0.3024
$\gamma_1 = 0.4$	-0.0594	19.7766	0.5029	-0.0971	0.2354	0.1978	0.2448
$\gamma_2 = 0.6$	0.0072	18.9333	0.4382	0.0382	0.3249	0.1893	0.3264

amounts are generally of small magnitudes. Some degree of difference can also be observed between the estimated and true variances, pointing to certain inefficiency of the estimators. This might be attributed to the fact that, when applying multiple imputation, small-sample behaviour stems from both the actual sample size, N, as well as from the number of imputations, M. Thus, in cases where the former is large while the latter is relatively small, it should not come as a surprise that the estimated variance is relatively large.

Finally, we consider GM III, which is based on a Gaussian measurement model and a logistic dropout model. The analysis model used for this GM was MI-GEE, which requires an imputation model and a measurement model, but not a dropout model. Multiple imputations of the missing Gaussian outcomes were first obtained using a

Table 5.3: *Small-sample simulation results for MI-GEE, with everything correctly specified, under GM III. Results include the average ($\overline{Est}$), bias ($Bias_N$), estimated variance ($\widehat{Var}_N$) and mean squared error (MSE), of the parameter estimators, for $N = 100$.*

	Small-Sample			
Parameter	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	MSE
β_0 = -3.0373	-3.0358	0.0015	0.1978	0.1978
β_x = 0.0095	0.0151	0.0056	0.3968	0.3968
β_t = 1.7812	1.7808	-0.0004	0.0601	0.0601
β_{xt} = 0.4828	0.4767	-0.0061	0.1480	0.1481

Gaussian imputation model, thereby ensuring a correctly specified imputation model, that is, one that uses the underlying measurement process to generate the imputations for the missing observations. The Gaussian outcome was then dichotomized based on the previously defined cutoff value, after which standard GEE, using a probit link, was applied to the dichotomized outcome of the imputed datasets. Since the underlying distribution for the outcomes is not discrete, only small-sample simulations are possible. Although initially $S = 500$ samples were generated, after dichotomization of the Gaussian outcome, there were 51 samples for which convergence was not attained. Inspection of these samples showed that the treatment-by-time interaction could not be estimated because, at one time point, all dichotomized outcomes belonged to only one treatment group.

Table 5.3 gives the results of the simulation only for the $S' = 449$ convergent samples. The "true" parameter values used to compute the bias were obtained by fitting the same measurement model using the complete (binary) data from the $S' = 449$ samples. Consistent with the theory on MI, we obtained only very small bias for the estimates, which might be expected to decrease even further under larger samples.

Dropout and Measurement Models Correct, Imputation Model Incorrect

We now consider a comparison between WGEE and MI-GEE, both having a correctly specified measurement model, but the latter using an incorrectly specified imputation model and the former specifying the dropout model correctly. For the two methods,

Table 5.4: *Small-sample simulation results for WGEE, with correctly specified dropout, and MI-GEE, with incorrectly specified imputation model, under GM I. Results include the bias ($Bias_N$), estimated variance ($\widehat{Var}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators (Parm), for $N = 100$.*

	WGEE				MI-GEE			
Parm	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
$\beta_0 = $ -0.25	-0.3956	1.0779	0.0044	1.2345	-0.0169	0.2332	0.1896	0.2335
$\beta_x = 0.5$	0.1225	2.1108	0.0111	2.1258	0.0195	0.4835	0.3938	0.4839
$\beta_t = 0.2$	0.1018	0.2388	0.0012	0.2491	0.0088	0.0548	0.0414	0.0548
$\beta_{xt} = $ -0.8	-0.1356	0.4441	0.0028	0.4625	-0.0058	0.1172	0.0885	0.1172

the measurement model used is consistent with the underlying Bahadur measurement model of GM I. Fitting WGEE for GM I, using the same mean structure as that of the underlying measurement model and with weights obtained from a logistic dropout model with the treatment indicator and the previous measurement as predictors, ensures every aspect is correctly specified. For MI-GEE, imputations are done using a conditional logistic imputation model for binary outcomes – a model that is *not* consistent with the marginal nature of the underlying Bahadur measurement model and is, therefore, incorrectly specified. Thus, the said comparison, of WGEE with correctly specified dropout and measurement models against MI-GEE with correctly specified measurement model but incorrectly specified imputation model, is possible under GM I. The results are given in Table 5.4.

As was already noted above, WGEE does not yield unbiased and consistent estimators for the particular sample size used, whereas the bias is considerably smaller for MI-GEE. The latter also leads to more precise estimators than those obtained for WGEE, as evidenced by smaller differences between the estimated and true variances for MI-GEE, despite the fact that the WGEE analysis model used was entirely correctly specified. Moreover, comparison of the MSEs indicate more efficient estimators for MI-GEE. All of these observations suggest a certain amount of robustness of MI-GEE when misspecifying the imputation model.

Table 5.5: *Small-sample simulation results for WGEE, with incorrectly specified dropout, and MI-GEE, with correctly specified imputation model, under GM III. Results include the bias ($Bias_N$), estimated variance ($\widehat{Var}_N$) and mean squared error (MSE), of the parameter estimators (Parm), for $N = 100$.*

	WGEE			MI-GEE		
Parm	Bias_N	$\widehat{\text{Var}}_N$	MSE	Bias_N	$\widehat{\text{Var}}_N$	MSE
$\beta_0 = -3.0373$	-0.1855	0.3113	0.3457	0.0015	0.1978	0.1978
$\beta_x = 0.0095$	-0.1380	0.5644	0.5834	0.0056	0.3968	0.3968
$\beta_t = 1.7812$	0.3100	0.1376	0.2336	-0.0004	0.0601	0.0601
$\beta_{xt} = 0.4828$	0.0367	0.2312	0.2325	-0.0061	0.1480	0.1481

Imputation and Measurement Models Correct, Dropout Model Incorrect

Whereas above the relative performances of WGEE with correctly specified dropout and MI-GEE with incorrectly specified imputation model were compared, we now proceed to look at the reverse. That is, we consider a comparison of WGEE with incorrectly specified dropout model against MI-GEE with correctly specified imputation model. In both cases, the measurement model corresponds to the assumed underlying measurement model for the dichotomized version of the continuous response. For this assessment, we apply the methods under GM III. For GM III, imputing the missing observations using a Gaussian imputation model and subsequently fitting standard GEE to dichotomized outcomes of the completed sets of data, results in MI-GEE with everything correctly specified. To enable comparison with WGEE using an incorrectly specified dropout model, we obtain weights from a logistic dropout model with the treatment indicator and the binary version of the previous measurement as predictors. The latter is a clear misspecification in the dropout model, since the underlying dropout model uses the continuous form of the previous measurement as predictor. The results of this comparison are given in Table 5.5. Only small-sample simulations are possible since the underlying GM does not consist of a discrete set of outcomes.

Bias is much smaller for MI-GEE, which can be expected as this is a correctly specified analysis model. With respect to the estimated precision ($\widehat{\text{Var}}_N$) for a sample of size $N = 100$, the estimates obtained from MI-GEE are superior to those from WGEE. Overall, the MI-GEE estimates are more efficient, with MSEs for the WGEE

Table 5.6: *Asymptotic and small-sample simulation results for marginalized MI-Transition, with everything correctly specified, under marginalized GM II. Asymptotic results include asymptotic bias ($Bias_\infty$) and asymptotic variance (Var_∞), while small-sample simulation results (for 500 simulations) include the average ($\overline{Est}$), bias ($Bias_N$), estimated variance ($\widehat{Var}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators, for $N = 100$.*

	Asymptotic		Small-Sample				
Parameter	Bias_∞	Var_∞	$\overline{\text{Est}}$	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
$\beta_0 = -0.3658$	-0.0045	1.11659	-0.4253	-0.0595	1.1230	0.0112	1.1265
$\beta_x = 0.2673$	0.0285	2.39565	0.3134	0.0461	2.4702	0.0240	2.4723
$\beta_t = 0.2265$	-0.0022	0.20405	0.2644	0.0379	0.2060	0.0020	0.2074
$\beta_{xt} = 0.0790$	-0.0363	0.43727	0.0648	-0.0142	0.4524	0.0044	0.4526

estimates about 1.5 times those of MI-GEE. These results seem to highlight the sensitivity of WGEE to misspecifications in the dropout model, in contrast to MI-GEE, which was noted to be somewhat robust to misspecifications in the imputation model.

Imputation and Dropout Models Correct, Measurement Model Incorrect

We finally proceed to looking at a comparison between WGEE and MI-GEE when the measurement model is specified incorrectly. For this setting, we consider GM II. We first present the results of the asymptotic and small-sample simulations for the marginalized version of MI-Transition, with which WGEE and MI-GEE are subsequently compared. Recall that the resulting parameter estimates, from the correctly specified MI-Transition model (Table 5.2), define three sets of conditional probabilities, from which marginal probabilities can be derived as in (5.2). These estimated probabilities were then used as weights in fitting a GEE model of the form (5.1) on a data set consisting of all possible combinations of outcome sequences and treatment level, yielding the marginalized version of MI-Transition. The resulting parameter estimates were subsequently compared with these "marginal" parameters; the results are shown in Table 5.6. Asymptotic bias for the parameter estimates is generally small, while its small-sample counterpart is larger. Estimated and true variances for a sample of size

Table 5.7: *Small-sample simulation results for WGEE, with correctly specified dropout, and MI-GEE, with correctly specified imputation model, under marginalized GM II. Results include the bias ($Bias_N$), estimated variance ($\widehat{Var}_N$), true variance (Var_N) and mean squared error (MSE), of the parameter estimators (Parm), for $N = 100$.*

	WGEE				MI-GEE			
Parm	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE	Bias_N	$\widehat{\text{Var}}_N$	Var_N	MSE
$\beta_0 = $ -0.3658	-0.4223	1.1310	0.0047	1.3098	-0.0562	0.2508	0.1901	0.2539
$\beta_x = 0.2673$	-0.1451	2.9804	0.0104	3.0014	0.0530	0.4927	0.3841	0.4955
$\beta_t = 0.2265$	0.1241	0.2149	0.0014	0.2303	0.0343	0.0608	0.0414	0.0620
$\beta_{xt} = 0.0790$	0.0792	0.5877	0.0030	0.5940	-0.0233	0.1184	0.0847	0.1190

$N = 100$ differ substantially, indicating some degree of inefficiency under this sample size.

Assuming now that these "marginal" parameters define some underlying marginal model for GM II, we fit both WGEE and MI-GEE, with a correctly specified dropout model and a correctly specified imputation model, respectively. For WGEE, weights are obtained from a dropout model consistent with the underlying dropout model of GM II, while for MI-GEE, imputations are generated from a conditional AR(2) transition model, which is in line with the underlying measurement model of GM II. In this way, both the dropout and imputation models are correctly specified. However, the fitted measurement models for both WGEE and MI-GEE are clearly misspecified, in the sense that the outcomes are modelled marginally (that is, GEE), rather than conditionally (that is, AR(2)). Let us compare the results of both of these misspecified models (Table 5.7) as well as the correctly specified marginalized MI-Transition (Table 5.6). Clearly, MI-GEE produces less biased estimates compared to WGEE and even to marginalized MI-Transition. In addition, MI-GEE outperforms both WGEE and marginalized MI-Transition in terms of precision and efficiency.

Table 5.8: *First depression trial. GEE, WGEE and MI-GEE. View 1: tests for (1) the joint null hypothesis of no treatment effect at none of the time points and (2) the hypothesis of no average treatment effect, and View 2a: test for (3) treatment effect at the last visit.*

	View 1				View 2a		
Analysis	Joint effect	Mean effect			Effect at last visit		
	p-value	Est.	(S.E.)	p-value	Est.	(S.E.)	p-value
GEE	0.3047	0.26	(0.27)	0.3355	0.71	(0.38)	0.0633
WGEE	0.1694	0.30	(0.39)	0.4413	0.69	(0.39)	0.0785
MI-GEE (binary)	0.3661	0.24	(0.19)	0.2134	0.74	(0.44)	0.1109
MI-GEE (continuous)	0.5142	0.17	(0.18)	0.3288	0.53	(0.33)	0.1048

5.3 First Depression Trial

In Section 4.4.2, we already applied weighted GEE to analyse the dichotomized response from the first depression trial data. Additionally, in this section, we consider multiple imputation based GEE, in which the imputation is performed both on the continuous $HAMD_{17}$ score itself and on its dichotomized version. The SAS code used for this purpose is illustrated in Section 11.4. First, when imputing the missing observation of the continuous $HAMD_{17}$ response variable, a multivariate Gaussian imputation model is considered. Afterwards, the obtained Gaussian outcome is dichotomized according to depression status, that is, the patient is diagnosed as being depressed in case his/her score is larger than 7. In the second case, in which multiple imputation is implemented on the dichotomized version of the $HAMD_{17}$ score, imputations are obtained through a transition model thereby adopting all observations of previous visits.

In line with Sections 3.1.2 and 4.4.2, two routes will be taken regarding the choice of the measurement model. First, we follow a fully longitudinal approach (View 1), focussing on the hypothesis of (1) no average treatment effect as well as (2) no treatment effect at any of the measurement occasions. Further, we also consider the analysis of the treatment effect at the last visit, that is, the planned occasion (View 2a).

Results are shown in Table 5.8. We observe a lower estimate and standard error for the mean treatment effect for both MI-GEE approaches compared to WGEE, yielding a lower p-value. On the other hand, for the joint treatment effect MI-GEE provides higher p-values. When considering the treatment effect at the last visit, the estimate of the binary MI-GEE method is similar to the one base on WGEE, whereas the estimate of MI-GEE after imputing the continuous $HAMD_{17}$ score is slightly different. However, the corresponding p-values of both MI-GEE approaches are comparable, and result in a less significant effect compared to both GEE and WGEE.

5.4 Concluding Remarks

When the analysis of incomplete binary longitudinal data is envisaged, several routes are available. Apart from likelihood-based methods, such as the generalized linear mixed-effects model, which were discussed in Chapters 3 and 4, non-likelihood methods are attractive, especially when a so-called marginal model is of interest. Because standard generalized estimating equations (Liang and Zeger, 1986) are unbiased only under MCAR, a variety of modifications and alternatives to GEE are available. Undoubtedly the most popular route is through weighted estimating equations, as proposed by Robins, Rotnitzky and Zhao (1995). A combination of GEE and multiple imputation methods methods (MI-GEE) provides an alternative route (Schafer, 2003). Once multiple imputation is considered an option, it has the merit of allowing for a variety of imputation techniques, whereafter several analysis methods can be considered. Two such routes considered in this chapter are MI-GEE and MI-Transition.

In this chapter we have provided quantitative evidence, based on asymptotic, as well as small-sample, simulations, that can be usefully applied in the decision making process. We have considered WGEE, MI-GEE, and MI-Transition under a variety of scenarios. While simulations are necessarily limited, we believe both methods have been put to the test in a fair fashion. Although asymptotically WGEE exhibits the desirable properties that it theoretically is known to possess, these are barely reproduced for small samples, even when every aspect of the analysis is correctly specified. Moreover, the observed sensitivity of WGEE to misspecification in either the dropout or measurement model renders these asymptotic properties meaningless. On the other hand, MI-GEE and MI-Transition demonstrate a certain degree of robustness to misspecification in either the imputation or measurement model, this, despite a further marginalization for the MI-Transition case. Furthermore, WGEE's applicability to

the case where also covariates are missing is less straightforward, while application of MI is relatively easy. Moreover, one can do MI under MAR with intermittent missingness. Although the results of this study provide insight about the methods under consideration, it is always wise to try a couple of different methods, by way of sensitivity analysis.

WGEE is merely the incorporation of inverse probability weighting (IPW) into the conventional GEE setup. In general, IPW is a method for correcting for missingness mechanisms that are not strictly MCAR by using information about the missingness probabilities. In its basic form the method has the advantage of robustness, since it does not depend on the knowledge of the distribution of the unobserved data. However, the price that is paid for this is inefficiency (Clayton *et al.*, 1998), which was also shown in this chapter. Specifically, Clayton *et al.* (1998) investigated the use of inverse probability weighting (IPW) and multiple imputation, among others, in the context of longitudinal binary data in a multi-phase sampling setting. They found that, while IPW was inefficient for such a 2×2-phase design, MI showed remarkable efficiency. Moreover, this, along with possible extension to data arising from other designs, indicates the substantial strengths of MI. To overcome this problem of inefficiency, Carpenter, Kenward and Vansteelandt (2006) developed so-called doubly robust IPW, a modified version of IPW, introduced in the discussion rejoinder in Scharfstein, Rotnizky and Robins (1999). Carpenter, Kenward and Vansteelandt (2006) used simulations to study these doubly robust IPW estimators in comparison with standard IPW, maximum likelihood, and MI. IPW estimators were again found to be inefficient and sensitive to the choice of the weight model, but the doubly robust version proved to be as efficient as MI and robust to misspecification. However, its applicability to the case where also covariates are missing is less straightforward, for which application of MI is still possible. Although applied to continuous Gaussian data, Carpenter, Kenward and Vansteelandt (2006) expect the results to generalize to the discrete case. Whereas Clayton *et al.* (1998) used actual data and Carpenter, Kenward and Vansteelandt (2006) used simulations of a small-sample nature, we complement a small-sample simulation study with asymptotic simulations. Through our simulations, we reinforce the strength of MI over IPW, specifically in application to GEE. Indeed, WGEE can be viewed as a type of IPW scheme that uses as weights the inverse of the probability of dropout (taken from some dropout model), while MI-GEE uses imputations for the missing data. WGEE was found to be inefficient for small-samples, in line with the findings of these two papers regarding the inefficiency of such IPW schemes. However, this (lack of) efficiency might well be addressed by adopting the doubly robust IPW version in obtaining the WGEE solutions.

Misspecifications are common in practice and it is seldom the case that one would have an entirely correctly specified analysis model. This, along with the fact that the nice properties of WGEE are not attained for modest sample sizes, which is common in typical clinical trials, discourages its recommendation. On the other hand, although theoretically MI-GEE does not provide consistent results when there is a misspecification, overall, it still yields more precise estimates than WGEE.

Thus, we provided evidence for the important fact that MI-GEE is less biased and more precise in small and moderate samples, in spite of the asymptotic unbiasedness of WGEE. As a consequence, in practice, MI-GEE would be the preferred method for analysis over WGEE. Moreover, although the focus of this thesis is on missingness in the response, in real-life settings, missingness in covariates is often encountered. In such cases, the choice for MI-GEE is even more convincing, since the use of WGEE would be ruled out. Finally, with MI, the imputation model is not restricted to the use of covariates that are available, without necessarily being of interest in the measurement model, can be incorporated in the imputation model, thereby yielding presumably better imputations as well as wider applicability.

As a final remark, recall that asymptotic simulations were done to obtain the asymptotic bias and asymptotic variance. These have theoretical use only, and may provide guidance as to what happens in large to very large samples. Supplementing them with small-sample simulations is therefore an attractive route. Needless to say the method is of no use with conventional data analysis.

6

MNAR and Its Relation with MAR

In Chapters 4 and 5 it has been shown that, if the MAR assumption is guaranteed to hold, a standard analysis will follow. This is certainly true for likelihood methods (Chapter 4), while for the marginal GEE methods it needs to be adjusted to the MAR case (Chapter 5).

However, in realistic settings, the reasons for missingness are varied and it is therefore hard to fully justify on *priori* grounds the assumption of MAR. Moreover, since it is not possible to test for MNAR against MAR (Jansen *et al.*, 2006b), one can never exclude the possibility that MNAR models may be operating. Nevertheless, ignorable analyses may provide reasonably stable results, even when the assumption of MAR is violated, in the sense that such analyses constrain the behavior of the unseen data to be similar to that of the observed data (Mallinckrodt, Clark and Stacy, 2001a,b). Further, such an MAR analysis can be specified beforehand without additional work relative to a situation with complete data.

In the most general MNAR setting, the cause of a subject's missingness depends on their unobserved responses, even after allowing for the information of the observed data. In this case, the missingness process should be modelled simultaneously with the response. An example of MNAR data would be a subject who had been doing

well until midway in a trial but relapsed after the last observed visit and was lost to follow-up. While MNAR models are more general and explicitly incorporate the dropout mechanism, the inferences they produce are typically highly dependent on untestable and often implicit assumptions regarding the distribution of the unobserved measurements given the observed ones. The quality of the fit to the observed data does not reflect at all the appropriateness of the implied structure governing the unobserved data. This point is irrespective of the MNAR route taken implying a definitive MNAR analysis does not exist. To explore the impact of deviations from the MAR assumption on the conclusions, one should ideally conduct a sensitivity analysis, within which models for the MNAR process can play a major role (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007). Such analyses will be discussed in Chapters 8 and 7.

In the first section of this chapter, we describe the full selection model for continuous outcomes in the MNAR situation proposed by Diggle and Kenward (1994) and give a concise review of selection models for the non-Gaussian setting, thereby focussing on the model family proposed by Baker, Rosenberger and DerSimonian (1992). Further, a brief overview is given of pattern-mixture models in Section 6.2. However, since such MNAR models ar not fully verifiable from the data, the empirical distinction between MNAR and MAR is not possible unless one is prepared to accept the posited MNAR model in an unquestioning way. Regarding this issue, Section 6.3 is devoted to the proof that an empirical distinction between MNAR and MAR is not possible, in the sense that each MNAR model fit to a set of observed data can be reproduced exactly by an MAR counterpart. Of course, such a pair of models will produce different predictions of the unobserved outcomes given the observed ones. The latter can be found in Beunckens *et al.* (2007c) and is achieved in collaboration with Cristina Sotto.

6.1 Full Selection MNAR Modeling

For continuous outcomes, Diggle and Kenward (1994) proposed a full selection model which is valid under MNAR. In the discrete case, Molenberghs, Kenward and Lesaffre (1997) considered a global odds ratio (Dale) model. Within the selection model framework, models have been proposed for non-monotone missingness as well (Baker, Rosenberger and DerSimonian, 1992; Jansen and Molenberghs, 2006), and further a number of proposals have been made for non-Gaussian outcomes (Molenberghs and Verbeke, 2005).

In this section, we first describe the Diggle and Kenward model for continuous longitudinal data. Next, we provide a brief perspective on counterparts for discrete data and picture the family of MNAR selection models for non-monotone missingness proposed by Baker, Rosenberger and DerSimonian (1992).

6.1.1 Diggle and Kenward Model for Continuous Longitudinal Data

Diggle and Kenward (1994) proposed a model for longitudinal Gaussian data with non-random dropout, that is, the missingness mechanism was assumed to be MNAR, which combines the multivariate normal model for longitudinal Gaussian data with a logistic regression for dropout. To maximize the resulting likelihood, integration over the missing data is needed.

The likelihood contribution of the ith subject, based on the observed data $(\boldsymbol{y}_i^o, d_i)$, is proportional to the marginal density function

$$f(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}) \;=\; \int f(\boldsymbol{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\psi})\; d\boldsymbol{y}_i^m \quad = \quad \int f(\boldsymbol{y}_i|\boldsymbol{\theta}) f(d_i|\boldsymbol{y}_i, \boldsymbol{\psi})\; d\boldsymbol{y}_i^m, \quad (6.1)$$

in which a marginal model for $\boldsymbol{Y}_i$ is combined with a model for the dropout process, conditional on the response - since we are considering the selection model framework - and where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are vectors of unknown parameters in the measurement model and dropout model, respectively.

Let $\boldsymbol{h}_{ij} = (y_{i1}, \ldots, y_{i;j-1})$ denote the observed history of subject i up to time $t_{i,j-1}$. The Diggle-Kenward model for the dropout process allows the conditional probability for dropout at occasion j, given that the subject was still observed at the previous occasion, to depend on the history $\boldsymbol{h}_{ij}$ and the possibly unobserved current outcome y_{ij}, but not on future outcomes y_{ik}, $k > j$. These conditional probabilities $P(D_i = j|D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})$ can now be used to calculate the probability of dropout at each occasion:

$$f(d_i|\boldsymbol{y}_i, \boldsymbol{\psi}) \;=\; P(D_i = d_i|\boldsymbol{y}_i, \boldsymbol{\psi}) \;=\; P(D_i = d_i|\boldsymbol{h}_{id_i}, y_{id_i}, \boldsymbol{\psi}) \quad (6.2)$$

$$= \begin{cases} P(D_i = d_i|D_i \geq d_i, \boldsymbol{h}_{id_i}, y_{id_i}, \boldsymbol{\psi}), & d_i = 2, \\ P(D_i = d_i|D_i \geq d_i, \boldsymbol{h}_{id_i}, y_{id_i}, \boldsymbol{\psi}) & \\ \qquad \times \prod_{j=2}^{d_i-1} \left[1 - P(D_i = j|D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})\right], & d_i = 3, \ldots, n, \\ \prod_{j=2}^{n} \left[1 - P(D_i = j|D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})\right], & d_i = n+1. \end{cases}$$

Diggle and Kenward (1994) combine a multivariate normal model for the measurement process with a logistic regression model for the dropout process. More specifically, the measurement model assumes that the vector $\boldsymbol{Y}_i$ of repeated measurements for the ith subject satisfies the linear regression model

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, V_i), \qquad i = 1, \ldots, N, \tag{6.3}$$

in which $\boldsymbol{\beta}$ is a vector of population-averaged regression coefficients. The matrix V_i can be left unstructured or assumed to be of a specific form, such as resulting from a linear mixed model.

The logistic dropout model can, for example, take the form

$$\text{logit}\left[P(D_i = j \mid D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})\right] \quad = \quad \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}. \tag{6.4}$$

More general models can easily be constructed by including the complete history $\boldsymbol{h}_{ij} = (y_{i1}, \ldots, y_{i;j-1})$, as well as external covariates, in the above conditional dropout model. Note also that, strictly speaking, one could allow dropout at a specific occasion to be related to all future responses as well. However, this is rather counter-intuitive in many cases. Moreover, including future outcomes seriously complicates the calculations since computation of the likelihood (6.1) then requires evaluation of a possibly high-dimensional integral. Note also that special cases of model (6.4) are obtained from setting $\psi_2 = 0$ or $\psi_1 = \psi_2 = 0$, respectively. In the first case, dropout is no longer allowed to depend on the current measurement, implying MAR. In the second case, dropout is independent of the outcome, which corresponds to MCAR. In both cases, all parameters can be estimated using standard software since the multivariate normal measurement model and the dropout model can then be fitted separately.

Diggle and Kenward (1994) obtained parameter and precision estimates by maximum likelihood. The likelihood involves marginalization over the unobserved outcomes $\boldsymbol{Y}_i^m$, for which subject-by-subject integration is required. Practically, this involves relatively tedious and computationally demanding forms of numerical integration. Diggle and Kenward (1994) used the Nelder and Mead simplex algorithm (Nelder and Mead, 1965). We use the Newton-Raphson Ridge optimization method instead.

6.1.2 Analysis of the Second Depression Trial Data

As an illustration, we apply the Diggle-Kenward selection modelling approach to the second depression trial data, introduced in Section 2.2. The data are analysed under MCAR, MAR, and MNAR, respectively. The six post-baseline visits correspond to the measurements taken at weeks 1, 2, 3, 5, 7, and 9. In the measurement model, we include an intercept, and assume as fixed effects the following covariates: *treatment*, *time*, $time^2$, and the interactions of *treatment* with *time* and $time^2$. Random effects are modeled as part of the within-subject error correlations, with the covariance structure to be of the heterogeneous first-order autoregressive type. Further, dropout model (6.4) is considered. Apart from the explicit MCAR, MAR, and MNAR versions of this model, we will also conduct an ignorable analysis (that is, an analysis based on the measurement model only, ignoring the dropout model). The results for the measurement model parameters have to coincide, on theoretical grounds, with these of the MCAR and MAR analyses. Analyses are implemented using SAS IML, and will be exemplified in Section 11.5.

In Table 6.1, parameter estimates and standard errors are listed for the four analyses, as well as the estimate of the difference in treatment effect at the endpoint, that is, week 9, which was the primary objective of the study, together with its p-value. The coincidence of MCAR, MAR, and ignorable measurement parameter estimates is observed, except for very small numerical instability. The p-value of the difference at the endpoint does not change much, it being significant in all four cases.

Note that for the MNAR analysis, the estimates of the ψ_1 and ψ_2 parameter are approximately equally large, but with different sign. This is in line with the argument of Molenberghs *et al.* (2001b), saying that the dropout oftentimes depends on the increment $y_{ij} - y_{i,j-1}$. This is because two subsequent measurements are usually positively correlated. By rewriting the fitted dropout model in terms of the increment, we obtain

$$\text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \boldsymbol{y}_i)\right] = -2.46 + 0.03 y_{i,j-1} - 0.08(y_{ij} - y_{i,j-1}),$$

This suggests that the probability of dropout increases with larger negative increments; that is, those patients who showed or would have shown a greater decrease in $HAMD_{17}$ score from the previous visit are more likely to drop out, given the decrease from baseline at the previous visit is not large. In other words, patients with a large improvement compared with the previous visit, a sudden shift on profile, are more likely to drop out.

Table 6.1: *Second Depression Trial. Estimates and standard errors of model parameters and the difference of treatment effect at the last visit, assuming ignorability, as well as explicitly modeling the missing data mechanism under MCAR, MAR, and MNAR assumptions.*

	Ignorable		MCAR		MAR		MNAR	
Parameters	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Mean Parameters								
β_0 : *intercept*	6.93	(1.49)	6.93	(1.48)	6.93	(1.48)	6.99	(1.48)
β_1 : *baseline*	-0.37	(0.07)	-0.37	(0.07)	-0.37	(0.07)	-0.37	(0.07)
β_2 : *treatment*	-0.34	(0.66)	-0.34	(0.65)	-0.34	(0.65)	-0.35	(0.67)
β_3 : *time*	-2.40	(0.29)	-2.40	(0.29)	-2.40	(0.29)	-2.49	(0.31)
β_4 : $time^2$	0.14	(0.03)	0.14	(0.03)	0.14	(0.03)	0.15	(0.03)
β_5 : $time \times treatment$	0.59	(0.40)	0.59	(0.40)	0.59	(0.40)	0.60	(0.41)
β_6 : $time^2 \times treatment$	-0.03	(0.04)	-0.03	(0.04)	-0.03	(0.04)	-0.04	(0.04)
Variance Parameters								
σ_1 : *std at time* 1	4.05		4.02	(0.17)	4.02	(0.17)	4.01	(0.17)
σ_2 : *std at time* 2	5.29		5.27	(0.24)	5.27	(0.24)	5.25	(0.24)
σ_3 : *std at time* 3	5.96		5.94	(0.27)	5.94	(0.27)	5.92	(0.27)
σ_4 : *std at time* 4	6.52		6.49	(0.29)	6.49	(0.29)	6.55	(0.30)
σ_5 : *std at time* 5	6.24		6.21	(0.28)	6.21	(0.28)	6.18	(0.27)
σ_6 : *std at time* 6	6.33		6.29	(0.30)	6.29	(0.30)	6.26	(0.29)
ρ : *common correlation*	0.73		0.72	(0.02)	0.72	(0.02)	0.72	(0.02)
Missing Data Parameters								
ψ_0			-2.46	(0.11)	-2.21	(0.14)	-2.46	(0.27)
ψ_1					-0.05	(0.02)	0.11	(0.05)
ψ_2							-0.08	(0.06)
-2 log-likelihood			7949.4		7943.1		7941.6	
diff. at endpoint (p-val.)	2.20	(.0179)	2.19	(.0176)	2.19	(.0176)	2.18	(.0177)

6.1.3 Models for Discrete Longitudinal Data

First and foremost, let us observe that the generalized linear mixed model discussed in Section 3.2.3 and the weighted and multiple imputation based estimating equations of Sections 3.2.2 and 5.1 can be embedded in MNAR models. In addition, a thorough review of full likelihood-based methods for non-Gaussian data can be found in Molenberghs and Verbeke (2005), many of which can be used when the missing data mechanism is MNAR. Let us briefly discuss a few of these.

Molenberghs, Kenward and Lesaffre (1997) proposed a model for longitudinal ordinal data with non-random dropout, that is, the missingness mechanism was assumed to be MNAR, which combines the multivariate Dale model for longitudinal ordinal data with a logistic regression model for dropout. The resulting likelihood can be maximized relatively simply, using the fact that all stochastic outcomes are of a categorical type, using the EM algorithm. It means that the integration over the missing data, needed to maximize the likelihood of Diggle and Kenward (1994), is replaced by finite summation. This is certainly not the only model available. The work on incomplete categorical data is vast. Baker and Laird (1988) develop the original work of Fay (1986) and give a thorough account of the modeling of contingency tables in which there is one response dimension and an additional dimension indicating whether the response is absent. Baker and Laird use loglinear models and the EM algorithm for the analysis. They pay particular attention to the circumstances in which no solution exists for the non-random dropout models. Such non-estimability is also a feature of the models we use below, but the more complicated setting makes a systematic account more difficult. Stasny (1986) and Conaway (1992, 1993) consider non-random missingness models for categorical longitudinal data. Baker (1995) allows for intermittent missingness in repeated categorical outcomes. Baker, Rosenberger and DerSimonian (1992) present a method for incomplete bivariate binary outcomes with general patterns of missingness, and we will provide more details on these so-called BRD models in turn. The model was adapted for the use of covariates by Jansen *et al.* (2003).

6.1.4 BRD Selection Models

Baker, Rosenberger and DerSimonian (1992) proposed a log-linear based family of models for two binary outcomes, possibly subject to non-monotone missingness. They use a four-way classification of both outcomes, together with their respective missingness indicators.

Table 6.2: *Theoretical distribution of the probability mass over complete and observed cells, respectively, for a bivariate binary outcome with non-monotone missingness. Tables correspond to completely observed subjects and subjects with the second, the first and both measurements missing, respectively.*

(a) Complete cells

$\pi_{11,11}$	$\pi_{11,12}$
$\pi_{11,21}$	$\pi_{11,22}$

$\pi_{10,11}$	$\pi_{10,12}$
$\pi_{10,21}$	$\pi_{10,22}$

$\pi_{01,11}$	$\pi_{01,12}$
$\pi_{01,21}$	$\pi_{01,22}$

$\pi_{00,11}$	$\pi_{00,12}$
$\pi_{00,21}$	$\pi_{00,22}$

(b) Observed cells

$\pi_{11,11}$	$\pi_{11,12}$
$\pi_{11,21}$	$\pi_{11,22}$

$\pi_{10,1+}$
$\pi_{10,2+}$

$\pi_{01,+1}$	$\pi_{01,+2}$

$\pi_{00,++}$

The generic expressions for the counts and corresponding probabilities are $Z_{r_1r_2j_1j_2}$ and $\pi_{r_1r_2j_1j_2}$ respectively, where $r_\ell = 0(1)$ if the measurement at occasion ℓ is missing (observed) and $j_\ell = 1(2)$ if the value for the binary variable ℓ is 1 (2).

The complete data and observed data cell probabilities are presented in Table 6.2. The models can be written as:

$$\begin{aligned} E(Z_{11,j_1j_2}) &= \nu_{11,j_1j_2}, & E(Z_{10,j_1j_2}) &= \nu_{11,j_1j_2}\tilde{\beta}_{j_1j_2}, \\ E(Z_{01,j_1j_2}) &= \nu_{11,j_1j_2}\tilde{\alpha}_{j_1j_2}, & E(Z_{00,j_1j_2}) &= \nu_{11,j_1j_2}\tilde{\alpha}_{j_1j_2}\tilde{\beta}_{j_1j_2}\tilde{\gamma}, \end{aligned}$$

with $\nu_{11,j_1j_2} = Z_{++++}\pi_{11,j_1j_2} = N\pi_{11,j_1j_2}$ and

$$\tilde{\alpha}_{j_1j_2} = \frac{P(r_1=0,r_2=1|j_1j_2)}{P(r_1=1,r_2=1|j_1j_2)}, \qquad \tilde{\beta}_{j_1j_2} = \frac{P(r_1=1,r_2=0|j_1j_2)}{P(r_1=1,r_2=1|j_1j_2)},$$

$$\tilde{\gamma} = \frac{P(r_1=1,r_2=1|j_1j_2)\,P(r_1=0,r_2=0|j_1j_2)}{P(r_1=1,r_2=0|j_1j_2)\,P(r_1=0,r_2=1|j_1j_2)},$$

such that the $\tilde{\alpha}$ ($\tilde{\beta}$) parameters describe missingness in the first (second) variable, and $\tilde{\gamma}$ captures the interaction between both non-response indicators. The subscripts are missing from $\tilde{\gamma}$ since Baker, Rosenberger and DerSimonian (1992) have shown that this quantity is independent of j_1 and j_2 in every identifiable model. From the expressions for $\tilde{\alpha}_{j_1j_2}$, $\tilde{\beta}_{j_1j_2}$, and $\tilde{\gamma}$, we see that selection model quantities are employed. Baker, Rosenberger and DerSimonian (1992) considered nine models, based on setting $\tilde{\alpha}_{j_1j_2}$ and $\tilde{\beta}_{j_1j_2}$ constant in one or more indices, and enumerated using the 'BRD'

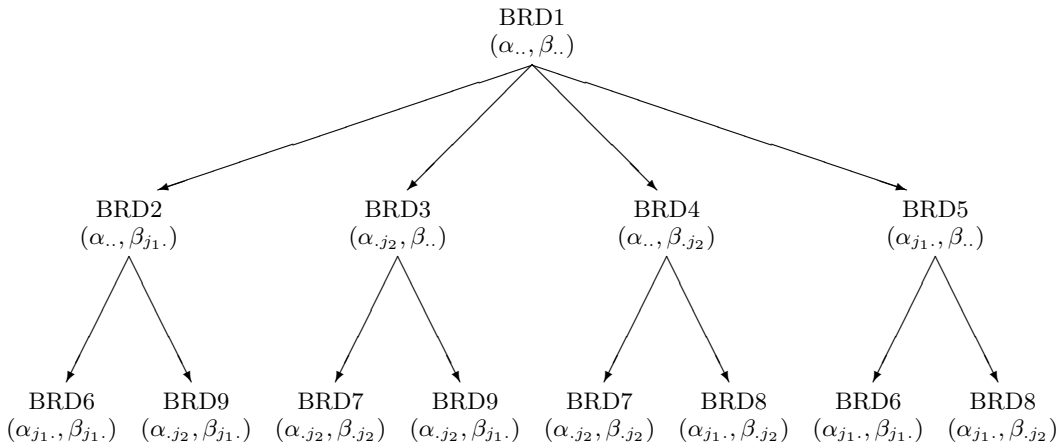


Figure 6.1: *Graphical representation of the BRD model nesting structure.*

abbreviation:

BRD1 :	$(\tilde{\alpha}_{..}, \tilde{\beta}_{..})$	BRD4 :	$(\tilde{\alpha}_{..}, \tilde{\beta}_{.j_2})$	BRD7 :	$(\tilde{\alpha}_{.j_2}, \tilde{\beta}_{.j_2})$
BRD2 :	$(\tilde{\alpha}_{..}, \tilde{\beta}_{j_1.})$	BRD5 :	$(\tilde{\alpha}_{j_1.}, \tilde{\beta}_{..})$	BRD8 :	$(\tilde{\alpha}_{j_1.}, \tilde{\beta}_{.j_2})$
BRD3 :	$(\tilde{\alpha}_{.j_2}, \tilde{\beta}_{..})$	BRD6 :	$(\tilde{\alpha}_{j_1.}, \tilde{\beta}_{j_1.})$	BRD9 :	$(\tilde{\alpha}_{.j_2}, \tilde{\beta}_{j_1.})$.

The nesting structure of these models is schematically represented in Figure 6.1. Interpretation is straightforward, for example, BRD1 is MCAR, and in BRD4 missingness in the first variable is constant, while missingness in the second variable depends on its – possibly unobserved – value. BRD6–BRD9 saturate the observed data degrees of freedom, while the lower numbered ones leave room for a non-trivial model fit to the observed data.

Jansen *et al.* (2003) extended the original BRD models to accommodate (possibly continuous) covariates. The index distinguishing between different covariate levels will be suppressed from notation. A selection model parameterization is used, differing from and extending the original one:

$$\pi_{r_1 r_2, j_1 j_2} = p_{jk} q_{r_1 r_2 | j_1 j_2}, \tag{6.5}$$

where $p_{j_1 j_2}$ parameterizes the measurement process and $q_{r_1 r_2 | j_1 j_2}$ describes the missingness mechanism, conditional on the measurements.

In particular, these authors assume

$$p_{j_1 j_2} = \frac{\exp(\boldsymbol{\eta}_{j_1 j_2})}{\sum_{j_1,j_2=1}^{2} \exp(\boldsymbol{\eta}_{j_1 j_2})}, \tag{6.6}$$

$$q_{r_1 r_2|j_1 j_2} = \frac{\exp[\alpha_{j_1 j_2}(1-r_1) + \beta_{j_1 j_2}(1-r_2) + \gamma(1-r_1)(1-r_2)]}{1+\exp(\alpha_{j_1 j_2}) + \exp(\beta_{j_1 j_2}) + \exp(\alpha_{j_1 j_2} + \beta_{j_1 j_2} + \gamma)}, \tag{6.7}$$

where $\alpha_{j_1 j_2}$, $\beta_{j_1 j_2}$ and γ have the same interpretation as $\tilde{\alpha}_{j_1 j_2}$, $\tilde{\beta}_{j_1 j_2}$ and $\tilde{\gamma}$.

No a priori ordering is imposed on the outcomes. The advantage is that genuine multivariate settings (e.g., several questions in a survey) can be handled as well. To identify the model, we set $\boldsymbol{\eta}_{22} = \mathbf{0}$ and further $\boldsymbol{\eta}_{j_1 j_2} = X_{j_1 j_2}\boldsymbol{\theta}$. This allows inclusion of covariate effects which, together with (6.6), is related to the multigroup logistic model (Albert and Lesaffre, 1986). In case no covariates are included in the model, the measurement model is modeled through $p_{j_1 j_2}$ as such, with $p_{22} = 1 - p_{11} - p_{12} - p_{21}$ to identify the model.

6.1.5 Analysis of the Slovenian Public Opinion Survey Data

In this section, we present an overview of the analyses of the Slovenian public opinion survey data, conducted by Rubin, Stern and Vehovar (1995) and Molenberghs, Kenward and Goetghebeur (2001a). Their main emphasis was on determining the proportion θ of the population that would attend the plebiscite and vote for independence. Therefore, we collapse Table 2.2 over the secession question, producing Table 6.3.

Let us first revise the estimates obtained by Rubin, Stern and Vehovar (1995), which are reproduced in Table 6.4. The complete case estimate for θ, $\widehat{\theta} = 0.928$, is based on the subjects answering all three questions and the available case estimate, $\widehat{\theta} = 0.929$, is based on the subjects answering the two questions of interest here. Apart from these simple models, Rubin, Stern and Vehovar (1995) considered two

Table 6.3: *Slovenian public opinion survey. Observed cells collapsed over the secession question. A simplified cell indexing system has been used.*

$m_1 : 1439$	$m_2 : 78$	$m_5 : 159$	$m_7 : 144$	$m_8 : 54$	$m_9 : 136$
$m_3 : 16$	$m_4 : 16$	$m_6 : 32$		

Table 6.4: *Slovenian public opinion survey. Some estimates of the proportion θ attending the plebiscite and voting for independence, as presented in Rubin, Stern and Vehovar (1995) and Molenberghs, Kenward and Goetghebeur (2001a).*

Estimation method	Voting in favour of independence: $\widehat{\theta}$
Non-parametric bounds	[0.694;0.905]
Complete cases	0.928
Available cases	0.929
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782
Plebiscite	0.885

MAR models, the first one solely based on the two questions of direct interest, the second one using the secession question as an auxiliary variable, producing $\widehat{\theta} = 0.883$ and $\widehat{\theta} = 0.782$, respectively. Finally, they considered a single MNAR model, based on the assumption that missingness on a question depends on the answer to that question but not on the other questions. Rubin, Stern and Vehovar (1995) concluded, owing to the proximity of the MAR analysis to the plebiscite value ($\theta_{\text{Pleb}} = 0.885$), that MAR in this and similar cases may be considered a plausible assumption.

Molenberghs, Kenward and Goetghebeur (2001a) supplemented these analysis with a so-called pessimistic-optimistic interval, also reported in Table 6.4. These pessimistic (optimistic) bounds, or non-parametric bounds, are obtained by setting all incomplete data that can be considered a yes (no), as yes (no). It is noteworthy that both estimates of the simple complete case and available case analyses are out of these bounds, underscoring the growing conviction that they should routinely be disregarded and a move towards, at least, MAR should be in place (Molenberghs and Kenward, 2007). Further, Molenberghs, Kenward and Goetghebeur (2001a) considered all nine BRD models, producing a range for θ from 0.741 to 0.892. Let us reflect on the results obtained from fitting each of these nine BRD models. Molenberghs, Kenward and Goetghebeur (2001a) presented a summary table but unfortunately

Table 6.5: *Slovenian public opinion survey. Summaries on each of the Models BRD1–BRD9 are presented, with obvious column labels.*

Model	Structure	d.f.	loglik	$\widehat{\theta}$	C.I.
BRD1	(α, β)	6	-2495.29	0.892	[0.878;0.906]
BRD2	(α, β_{j_1})	7	-2467.43	0.884	[0.869;0.900]
BRD3	(α_{j_2}, β)	7	-2463.10	0.881	[0.866;0.897]
BRD4	(α, β_{j_2})	7	-2467.43	0.765	[0.674;0.856]
BRD5	(α_{j_1}, β)	7	-2463.10	0.844	[0.806;0.882]
BRD6	$(\alpha_{j_1}, \beta_{j_1})$	8	-2431.06	0.819	[0.788;0.849]
BRD7	$(\alpha_{j_2}, \beta_{j_2})$	8	-2431.06	0.764	[0.697;0.832]
BRD8	$(\alpha_{j_1}, \beta_{j_2})$	8	-2431.06	0.741	[0.657;0.826]
BRD9	$(\alpha_{j_2}, \beta_{j_1})$	8	-2431.06	0.867	[0.851;0.884]

there was a small computational error that had to be corrected, for which reason the corrected results are reproduced in Table 6.5 (Molenberghs *et al.*, 2007).

A graphical representation of the original analyses and the BRD models combined is given in Figure 6.2. BRD1 produces $\widehat{\theta} = 0.892$, exactly the same as the first MAR estimate obtained by Rubin, Stern and Vehovar (1995). This does not come as a surprise, since both models assume MAR and use information from the two main questions. However, before continuing with the models' interpretation, it is necessary to assess their fit. Conducting likelihood ratio tests for BRD1 versus the ones with 7 parameters, that is, BRD2–BRD5, and then in turn for BRD2–BRD5 versus the saturated models BRD6–BRD9, suggests the lower numbered models do not fit well, leaving us with BRD6–BRD9. The impression might be generated that the poor model fit of BRD1 might be seen as evidence for discarding the MAR-based value 0.892. We will come back to this issue in Section 6.3.3.

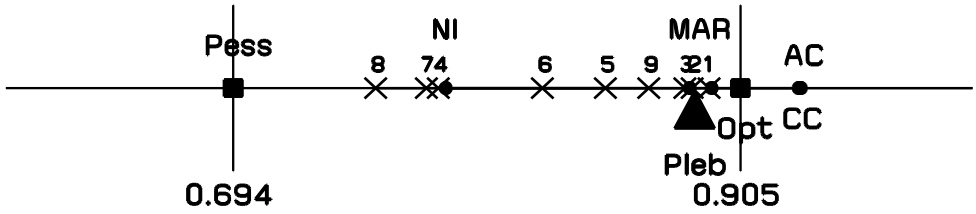


Figure 6.2: *Slovenian public opinion survey. Relative position for the estimates of "proportion of YES votes", based on the models considered in Rubin, Stern and Vehovar (1995) and on the BRD models. The vertical lines indicate the non-parametric pessimistic–optimistic bounds. (Pess: pessimistic boundary; Opt: optimistic boundary; MAR: Rubin* et al*'s MAR model; NI: Rubin* et al*'s MNAR model; AC: available cases; CC: complete cases; Pleb: plebiscite outcome. Numbers refer to the BRD models.)*

6.2 Pattern-Mixture Modeling

Pattern-mixture models were introduced in Section 3.1.2 as one of the three major frameworks within which missing data models can be developed. In this section we provide a brief overview of pattern-mixture models. More details can be found in Verbeke and Molenberghs (2000) and Molenberghs and Kenward (2007).

Early references include Rubin (1977), who mentioned the concept of a sensitivity analysis within a fully Bayesian framework, Glynn, Laird and Rubin (1986), and Little and Rubin (1987). Important early development was provided by Little (1993, 1994a, 1995).

Pattern-mixture models can be considered for their own sake to answer a particular scientific question. Further, several authors have contrasted selection models and pattern-mixture models. This is done either (1) to answer the same scientific question, such as marginal treatment effect or time evolution, based on these two rather different modeling strategies, or (2) to gain additional insight by supplementing the selection model results with those from a pattern-mixture approach.

Examples of pattern-mixture applications can be found in Verbeke, Lesaffre and Spiessens (2001a) or Michiels *et al.* (2002) for continuous outcomes, and Molenberghs, Michiels and Lipsitz (1999), or Michiels, Molenberghs and Lipsitz (1999) for categorical outcomes.

An important issue is that pattern-mixture models are by construction underidentified, that is, overspecified. Little (1993, 1994a) solves this problem through the use of identifying restrictions: inestimable parameters of the incomplete patterns

are set equal to (functions of) the parameters describing the distribution of the completers. Identifying restrictions are not the only way to overcome under-identification, and we will discuss alternative approaches. Although some authors perceive this under-identification as a drawback, we believe it is an asset because it forces one to reflect on the assumptions made. Pattern-mixture models can serve important roles in sensitivity analysis.

Fitting pattern-mixture models can be approached in several ways. It is important to decide whether pattern-mixture and selection models are to be contrasted with one another or rather the pattern-mixture modeling is the central focus. In the latter case, it is natural to conduct an analysis, and preferably a sensitivity analysis, *within* the pattern-mixture family. Basically we will consider three strategies to deal with under-identification.

- **Strategy 1.** As mentioned before, Little (1993, 1994a) advocated the use of identifying restrictions and presented a number of examples. A general framework for identifying restrictions is discussed in more detail in Thijs *et al.* (2002), with three special but important cases: *complete case missing values* (CCMV) (proposed by Little (1993)), *neighboring case missing values* (NCMV), and *available case missing values* (ACMV). Note that ACMV is the natural counterpart of MAR in the pattern-mixture model framework (Molenberghs *et al.*, 1998b). This provides a way to compare ignorable selection models with their counterpart in the pattern-mixture setting. Kenward, Molenberghs and Thijs (2003) focus on restrictions avoiding dependence of dropout on measurements made at future occasions.

  The procedure to apply identifying restrictions is discussed in full detail in Thijs *et al.* (2002). The key steps are as follows:

  1. Fit a model to the pattern-specific identifiable densities: $f_t(y_1, \ldots, y_t)$. This results in a set parameter estimates, $\boldsymbol{\beta}_p$ say, for each pattern p.
  2. Select an identification method of choice (ACMV, CCMV, NCMV).
  3. Using this identification method, determine the conditional distributions of the unobserved outcomes, given the observed ones: $f_t(y_{t+1}, \ldots, y_t | y_1, \ldots, y_t)$.
  4. Using standard multiple imputation methodology (Rubin, 1987; Schafer, 1997; Verbeke and Molenberghs, 2000; Molenberghs and Kenward, 2007), draw multiple imputations for the unobserved components, given the observed outcomes and the correct pattern-specific density.

5. Analyse the multiply-imputed data sets using the method of choice. This can be another pattern-mixture model, but also a selection model or any other desired model.
6. Inferences can be conducted in the standard multiple imputation way.

- **Strategy 2.** As opposed to identifying restrictions, model simplification can be done in order to identify the parameters. The advantage is that the number of parameters decreases, which is desirable since the length of the parameter vector is a general issue with pattern-mixture models. Indeed, Hogan and Laird (1997) noted that in order to estimate the large number of parameters in general pattern-mixture models, one has to make the awkward requirement that each dropout pattern occurs sufficiently often. Broadly, we distinguish between two types of simplifications.
  - ▷ **Strategy 2a.** Trends can be restricted to functional forms supported by the information available within a pattern. For example, a linear or quadratic time trend is easily extrapolated beyond the last obtained measurement. One only needs to provide an ad hoc solution for the first or the first few patterns. In order to fit such models, one simply has to carry out a model building exercise within each of the patterns separately.
  - ▷ **Strategy 2b.** Next, one can let the parameters vary across patterns in a controlled parametric way. Thus, rather than estimating a separate time trend within each pattern, one could for example assume that the time evolution within a pattern is unstructured, but parallel across patterns. This is effectuated by treating pattern as a covariate. The available data can be used to assess whether such simplifications are supported within the time ranges for which there is information.

Although the second strategy is computationally simple, it is important to note that there is a price to pay. Indeed, simplified models, qualified as *assumption rich* by Sheiner, Beal and Dunne (1997), also make untestable assumptions, just as in the selection model case. From a technical point of view, Strategy 2 only requires to either consider 'pattern' as an extra covariate in the model, or to conduct an analysis 'by pattern,' such that a separate analysis is obtained for each of the dropout patterns. In the identifying restrictions setting on the other hand (Strategy 1), the assumptions are clear from the start. Precisely for these reasons it is stated in Thijs *et al.* (2002) that the use of simplified models is not the best strategy and can be rather dangerous as well.

Pattern-mixture models do not always automatically provide estimates and standard errors of marginal quantities of interest, such as overall treatment effect or overall time trend. Hogan and Laird (1997) provided a way to derive selection model quantities from the pattern-mixture model. Several authors have followed this idea to formally compare the conclusions from a selection model with the selection model parameters in a pattern-mixture model (Michiels, Molenberghs and Lipsitz, 1999; Verbeke, Lesaffre and Spiessens, 2001a).

6.3 Every MNAR Model Has Got an MAR Bodyguard

As mentioned before, one can never exclude the possibility that MNAR models may be operating. Even though a variety of statistical models have been proposed for the MNAR situation, a few of which has been discussed in Sections 6.1 and 6.2, such models are prone to considerable sensitivity due to the unverifiable modeling assumptions.

In this section, we will show that, strictly speaking, the correctness of the alternative model can only be verified in as far as it fits the *observed* data. Thus, evidence for or against MNAR can only be provided within a particular, predefined parametric family, the plausibility of which cannot be verified in empirical terms alone. We show that the formal data-based distinction between MAR and MNAR is not possible, in the sense that each MNAR model fit to a set of observed data can be reproduced exactly by an MAR counterpart. Of course, such a pair of models will produce different predictions of the unobserved outcomes, given the observed ones. We show that, while this so-called MAR bodyguard generally does not belong to a conventional parametric family, its existence has important ramifications.

Such a position is in contrast to the view that one can test for an MNAR mechanism using the data under analysis. Such tests, comparing MAR and MNAR mechanisms, can of course be constructed using conventional statistical methodology as done, for example, by Diggle and Kenward (1994). It is very important to realize that such tests are conditional upon the alternative model holding, which can only be assessed as far as it fits the observed data, not the unobserved.

First, we show formally that every MNAR model can be doubled up with a uniquely defined MAR counterpart, producing exactly the same fit to the observed data as the original MNAR model, in the sense that it produces exactly the same pre-

dictions to the observed data (e.g., fitted counts in an incomplete contingency table) as the original MNAR model, and depending on exactly the same parameter vector. Next, the specific case of incomplete contingency tables is studied, after which we apply the ideas developed to data from the Slovenian public opinion survey.

6.3.1 General Result

In this section, we will show that for every MNAR model fitted to a set of data, there is an MAR counterpart providing exactly the same fit to the data. Here, the concept of model fit should be understood as measured using such conventional methods as deviance measures and, of course, in as far as the observed data are concerned. The following steps are involved: (1) fitting an MNAR model to the data; (2) reformulating the fitted model in PMM form; (3) replacing the density or distribution of the unobserved measurements given the observed ones and given a particular response pattern by its MAR counterpart; (4) establishing that such an MAR counterpart uniquely exists. Throughout this section, we will suppress covariates $\boldsymbol{x}_i$ from notation, but assume them to be present.

In the first step, we fit an MNAR model to the observed set of data. In line with the notation introduced in Section 3.1.2, the observed data likelihood is:

$$L = \prod_i \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{y}_i^m. \tag{6.8}$$

Upon denoting the obtained parameter estimates, e.g., obtained by likelihood-based or Bayesian methods, by $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\psi}}$ respectively, the fit to the hypothetical full data is

$$f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \widehat{\boldsymbol{\theta}}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \widehat{\boldsymbol{\psi}}). \tag{6.9}$$

To undertake the second step, full density (6.9) can be re-expressed in PMM form as:

$$f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f(\boldsymbol{y}_i^o | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}). \tag{6.10}$$

A similar reformulation can be considered for an SPM. In a PMM, the model will have been expressed in this form to begin with.

Note that, in line with PMM theory, the final term on the right hand side of (6.10), $f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, d_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$, is not identified from the observed data. In this case, it is determined solely from modeling assumptions. Within the PMM framework, identifying restrictions can be considered as mentioned in previous section.

The third step requires replacing this factor by the appropriate MAR counterpart. To this end, we need the following lemma, formulating MAR equivalently within the PMM framework.

Lemma 1 *In the PMM framework, the missing data mechanism is MAR if and only if*

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i,\boldsymbol{\theta}) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{\theta}).$$

This means that, in a given pattern, the conditional distribution of the unobserved components given the observed ones equals the corresponding distribution marginalized over the patterns.

The proof, which is rather straightforward and similar to what can be found in Molenberghs *et al.* (1998b), is shown below. Note that, owing to this result, MAR can be formulated in terms of $\boldsymbol{R}$ given $\boldsymbol{Y}$, but also in terms of $\boldsymbol{Y}$ given $\boldsymbol{R}$.

Proof of Lemma 1 Suppressing parameters and covariates from notation, the decomposition of the full data density, in both SeM and PMM fashion, whereby MAR is applied to the SeM version, produces:

$$f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m)f(\boldsymbol{r}_i|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m|\boldsymbol{r}_i)f(\boldsymbol{r}_i). \tag{6.11}$$

Further factoring the right hand side and moving the second factor on the left to the right as well gives:

$$\begin{aligned} f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i)\frac{f(\boldsymbol{y}_i^o|\boldsymbol{r}_i)f(\boldsymbol{r}_i)}{f(\boldsymbol{r}_i|\boldsymbol{y}_i^o)} \\ f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i)\frac{f(\boldsymbol{y}_i^o,\boldsymbol{r}_i)}{f(\boldsymbol{r}_i|\boldsymbol{y}_i^o)} \\ f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o)f(\boldsymbol{y}_i^o) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i)f(\boldsymbol{y}_i^o), \end{aligned}$$

and hence

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i).$$

⊠

Using Lemma 1, it is clear that $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i,\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}})$ needs to be replaced with

$$h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i) = h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}}), \tag{6.12}$$

where the $h(\cdot)$ notation is used for shorthand purposes. Note that the density in (6.12) follows from the SeM-type marginal density of the complete data vector. Sometimes, therefore, it may be more convenient to replace the notation $\boldsymbol{y}_i^o$ and $\boldsymbol{y}_i^m$ by one that explicitly indicates which components are observed and missing in pattern $\boldsymbol{r}_i$ under consideration:

$$h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i) = h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f[(y_{ij})_{r_j=0}|(y_{ij})_{r_j=1},\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}}]. \tag{6.13}$$

Thus, (6.13) provides a unique way of extending the model fit to the observed data, belonging to the MAR family. As stated before, the above construction does not lead to a member of a conventional parametric family. Also, it helps to understand that an overall, definitive conclusion about the nature of the missing data mechanism is not possible, even though one can make progress if attention is confined to a given parametric family, in which one puts sufficiently strong prior belief.

To show formally that the fit remains the same, we consider the observed-data likelihood based on (6.8) and (6.10):

$$\begin{aligned}
\widehat{L} &= \prod_i \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \widehat{\boldsymbol{\theta}}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \widehat{\boldsymbol{\psi}}) d\boldsymbol{y}_i^m \\
&= \prod_i \int f(\boldsymbol{y}_i^o | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) d\boldsymbol{y}_i^m \\
&= \prod_i f(\boldsymbol{y}_i^o | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) \\
&= \prod_i \int f(\boldsymbol{y}_i^o | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) h(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o) d\boldsymbol{y}_i^m.
\end{aligned}$$

The above results justify the following theorem:

Theorem 1 *Every fit to the observed data, obtained from fitting an MNAR model to a set of incomplete data, is exactly reproducible from an MAR decomposition.*

The key computational consequence is the need to compute $h(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o)$ in (6.12) or (6.13). This means, for each pattern, the conditional density of the unobserved measurements given the observed ones needs to be extracted from the marginal distribution of the complete set of measurements. Molenberghs *et al.* (1998b) have shown that, for the case of dropout, the so-called *available case missing value restrictions* (ACMV) provide a practical computational scheme. Precisely, ACMV states that

$$\forall t \geq 2, \forall s < t : f(y_{it} | y_{i1}, \cdots, y_{i,t-1}, d_i = s) = f(y_{it} | y_{i1}, \cdots, y_{i,t-1}, d_i \geq t). \quad (6.14)$$

In other words, the density of a missing measurement, conditional on the measurement history, is determined from the corresponding density over all patterns for which all of these measurements are observed. For example, the density of the third measurement in a sequence, given the first and second ones, in patterns with only 1 or 2 measurements taken, is determined from the corresponding density over all patterns with 3 or more measurements. Thijs *et al.* (2002) and Verbeke and Molenberghs

(2000, p. 347) derived a practical computational method for the factors in (6.14):

$$f(y_{it}|y_{i1},\cdots,y_{i,t-1},d_i=s) \tag{6.15}$$

$$= \frac{\sum_{d=s}^{n}\alpha_d f_d(y_{i1},\ldots,y_{is})}{\sum_{d=s}^{n}\alpha_d f_d(y_{i1},\ldots,y_{i,s-1})} \tag{6.16}$$

$$= \sum_{d=s}^{n}\left(\frac{\alpha_d f_d(y_{i1},\ldots,y_{i,s-1})}{\sum_{d=s}^{n_i}\alpha_d f_d(y_{i1},\ldots,y_{i,s-1})}\right) f_d(y_s|y_{i1},\ldots,y_{i,s-1}). \tag{6.17}$$

Here, α_d is the probability to belong to pattern d.

The above identifications for the monotone case are useful when an MNAR pattern-mixture model has been fitted to begin with, since then the identifications under MAR can be calculated from the pattern-specific marginal distributions. In case a selection model has been fitted in the initial step, $f(y_{i1},\ldots,y_{in_i}|\widehat{\boldsymbol{\theta}})$ has been estimated, from which all conditional distributions, needed in (6.13), can be derived. When the initial model is an MNAR PMM model and the missing data patterns are non-monotone, then it is necessary to first rewrite the PMM in SeM form, and derive the required conditional distributions from the so-obtained SeM measurement model. This essentially comes down to calculating a weighted average of the pattern-specific measurement models. In some cases, such as for contingency tables, this step can be done in an alternative way by fitting a saturated MAR selection model to the fit obtained from the PMM model.

We will illustrate and contrast the monotone and non-monotone cases using a bivariate and trivariate outcome with dropout on the one hand and a bivariate non-monotone outcome on the other hand. While the theorem applies to both the monotone and non-monotone settings, it is insightful to see that only for the former relatively simple and intuitively appealing expressions arise, while the latter setting involves the need for iterative computation. In the next section, the aforementioned general contingency table setting to which a PMM has been fitted, will be studied.

A Bivariate Outcome With Dropout

Here and in the following examples, we will present and equate the SeM and PMM decompositions, enabling us to derive expressions for the MAR bodyguards. It is interesting and straightforward to derive results for the MCAR case, and hence these will be presented, too.

Dropping covariates, parameters, and the subject index i from notation, the SeM-PMM equivalence for the case of two outcomes, the first of which is always observed

but the second one partially missing, is given by:

$$\begin{aligned} f(y_1, y_2)\widetilde{g}(d = 2|y_1, y_2) &= f_2(y_1, y_2)\widetilde{\alpha}(d = 2), \\ f(y_1, y_2)\widetilde{g}(d = 1|y_1, y_2) &= f_1(y_1, y_2)\widetilde{\alpha}(d = 1). \end{aligned}$$

Note that this is the setting considering by Glynn, Laird and Rubin (1986). Here, $\widetilde{g}(\cdot)$ is used for the SeM dropout model, with $\widetilde{\alpha}(\cdot)$ denoting the PMM probabilities to belong to one of the patterns. Since $\widetilde{\alpha}(d = 1) + \widetilde{\alpha}(d = 2) = 1$ and a similar result holds for the $\widetilde{g}(\cdot)$ functions, it is convenient to write:

$$f(y_1, y_2)g(y_1, y_2) = f_2(y_1, y_2)\alpha \tag{6.18}$$

$$f(y_1, y_2)[1 - g(y_1, y_2)] = f_1(y_1, y_2)[1 - \alpha]. \tag{6.19}$$

Assuming MCAR, it is clear that $\alpha = g(y_1, y_2)$, producing, without any difficulty:

$$f(y_1, y_2) = f_2(y_1, y_2) = f_1(y_1, y_2). \tag{6.20}$$

Under MAR, y_2 has to be removed from $g(\cdot)$ for incomplete observations, but since we assume a single parametric function for the missingness model, it follows that $g(y_1, y_2) = g(y_1)$ and hence (6.18) produces

$$f(y_1)f(y_2|y_1)g(y_1) = f_2(y_1)f_2(y_2|y_1)\alpha.$$

Upon reordering, we find:

$$\frac{f(y_1)g(y_1)}{f_2(y_1)\alpha} = \frac{f_2(y_2|y_1)}{f(y_2|y_1)}, \tag{6.21}$$

yielding $f(y_2|y_1) = f_2(y_2|y_1)$. The same arguments can be applied to (6.19), and combined with the previous finding we obtain:

$$f(y_2|y_1) = f_2(y_2|y_1) = f_1(y_2|y_1). \tag{6.22}$$

Note that (6.22) is strictly weaker than (6.20). The last term in (6.22) is not identified by itself, and hence, we see it needs to be set equal to its counterpart from the completers which, in turn, is equal to the marginal distribution. This is in agreement with (6.13) as well as with the specific identifications applicable in the monotone and hence ACMV setting.

A Trivariate Outcome With Dropout

Note that identification (6.22) does not involve mixtures. This changes as soon as there are three or more outcomes. The equations corresponding to (6.18)–(6.19),

specialized to the MAR case, are:

$$\begin{aligned} f(y_1, y_2, y_3)g_0 &= f_0(y_1, y_2, y_3)\alpha_0, && (6.23)\\ f(y_1, y_2, y_3)g_1(y_1) &= f_1(y_1, y_2, y_3)\alpha_1, && (6.24)\\ f(y_1, y_2, y_3)g_2(y_1, y_2) &= f_2(y_1, y_2, y_3)\alpha_2, && (6.25)\\ f(y_1, y_2, y_3)g_3(y_1, y_2) &= f_3(y_1, y_2, y_3)\alpha_3. && (6.26) \end{aligned}$$

We have chosen to include pattern 0, the one without follow-up measurements, as well, and will return to this one. We could write $g_3(\cdot)$ as a function of y_3 as well, but because the sum of the $g_d(\cdot)$ equals one, it is clear that $g_3(\cdot)$ ought to be independent of y_3. With arguments similar to the ones developed in the case of two measurements, we can rewrite (6.26) as:

$$\frac{f(y_1, y_2)}{f_3(y_1, y_2)} \cdot \frac{g_3(y_1, y_2)}{\alpha_3} = \frac{f_3(y_3|y_1, y_2)}{f(y_3|y_1, y_2)}.$$

Exactly the same consideration can be made based on (6.25), and hence

$$f_3(y_3|y_1, y_2) = f(y_3|y_1, y_2) = f_2(y_3|y_1, y_2). \quad (6.27)$$

The first factor identifies the second one, and hence also the third one. Starting from (6.24), we obtain:

$$f_1(y_2, y_3|y_1) = f(y_2, y_3|y_1),$$

which produces, in fact, two separate identities:

$$\begin{aligned} f_1(y_2|y_1) &= f(y_2|y_1), && (6.28)\\ f_1(y_3|y_1, y_2) &= f(y_3|y_1, y_2) = f_3(y_3|y_1, y_2) = f_2(y_3|y_1, y_2). && (6.29) \end{aligned}$$

For the latter one, identity (6.27) has been used as well. The density $f(y_2|y_1)$, needed in (6.28), is determined from the general ACMV result (6.17):

$$f(y_2|y_1) = \frac{\alpha_2 f_2(y_2|y_1) + \alpha_3 f_3(y_2|y_1)}{\alpha_2 + \alpha_3}.$$

Finally, turning attention to (6.23), it is clear that $g_0 = \alpha_0$ and hence also $f_0(y_1, y_2, y_3) = f(y_1, y_2, y_3)$. From the latter density, only $f(y_1)$ has not been determined yet, but this one follows again very easily from the general ACMV result:

$$f(y_1) = \frac{\alpha_1 f_1(y_1) + \alpha_2 f_2(y_1) + \alpha_3 f_3(y_1)}{\alpha_1 + \alpha_2 + \alpha_3}.$$

In summary, the necessary MAR identifications easily follow from both the PMM and the SeM formulations of the model.

A Bivariate Outcome With Non-Monotone Missingness

The counterparts to (6.18)–(6.19) and (6.23)–(6.26) for a bivariate outcome with non-monotone missingness are

$$\begin{aligned} f(y_1,y_2)g_{00}(y_1,y_2) &= f_{00}(y_1,y_2)\alpha_{00}, && (6.30)\\ f(y_1,y_2)g_{10}(y_1,y_2) &= f_{10}(y_1,y_2)\alpha_{10}, && (6.31)\\ f(y_1,y_2)g_{01}(y_1,y_2) &= f_{01}(y_1,y_2)\alpha_{01}, && (6.32)\\ f(y_1,y_2)g_{11}(y_1,y_2) &= f_{11}(y_1,y_2)\alpha_{11}. && (6.33) \end{aligned}$$

Clearly, under MCAR, the $g_{r_1r_2}(\cdot)$ functions do not depend on the outcomes and hence $f_{r_1r_2}(y_1,y_2) = f(y_1,y_2)$ for all four patterns. For the MAR case, (6.30)–(6.33) simplify to

$$\begin{aligned} f(y_1,y_2)g_{00} &= f_{00}(y_1,y_2)\alpha_{00}, && (6.34)\\ f(y_1,y_2)g_{10}(y_1) &= f_{10}(y_1,y_2)\alpha_{10}, && (6.35)\\ f(y_1,y_2)g_{01}(y_2) &= f_{01}(y_1,y_2)\alpha_{01}, && (6.36)\\ f(y_1,y_2)g_{11}(y_1,y_2) &= f_{11}(y_1,y_2)\alpha_{11}. && (6.37) \end{aligned}$$

Observe there are four identifications across the $g_{r_1r_2}(y_1,y_2)$ functions:

$$g_{00} + g_{10}(y_1) + g_{01}(y_2) + g_{11}(y_1,y_2) = 1,$$

for each (y_1,y_2). Also $\sum_{r_1,r_2}\alpha_{r_1,r_2} = 1$. Applying the usual algebra to (6.34)–(6.37), we obtain three identifications for the unobservable densities:

$$\begin{aligned} f_{00}(y_1,y_2) &= f(y_1,y_2), && (6.38)\\ f_{10}(y_1|y_2) &= f(y_1|y_2), && (6.39)\\ f_{01}(y_2|y_1) &= f(y_2|y_1). && (6.40) \end{aligned}$$

Using these in conjunction with the identifiable parts of the distributions yields the MAR bodyguard.

6.3.2 The General Case of Incomplete Contingency Tables

In Section 6.3.1, we have derived general identification schemes for an MAR extension of a fitted model to a binary or trivariate outcome with dropout, as well as to a bivariate outcome with non-monotone missingness. Whereas the monotone cases provide explicit expressions in terms of the pattern-specific densities, the three identifications (6.38)–(6.40) obtained in the case of non-monotone missingness provide an

identification only in terms of the marginal probability. This in itself is not a problem, since the marginal density is always available, either directly when a SeM is fitted, or through marginalization when a PMM or an SPM is fitted.

In the specific case of contingency tables, further progress can be made. Indeed, we can show a saturated MAR model is always available, for any incomplete contingency table setting. This implies one can start from the fit of an MNAR model to the observed data, and then extend it, using this result, towards MAR. We will present the general result and then discuss its precise implications for practice.

Assume we have a $\prod_{k=1}^{n} c_k$ contingency table with supplemental margins, where k indexes the n dimensions in the table and c_k is the number of alternatives the kth categorical variable can take. The table of completers is indexed by $\boldsymbol{r} = \mathbf{1} = (1, \ldots, 1)$. A particular incomplete table is indexed by a $\boldsymbol{r} \neq \mathbf{1}$. The full set of tables can but does not have to be present. The number of cells is:

$$\#\text{cells} = \sum_{\boldsymbol{r}} \prod_{k=1}^{n} c_k^{r_k}. \tag{6.41}$$

Denote the measurement model probabilities by $p_{\boldsymbol{j}} = p_{j_1 \ldots j_n}$ for $j_k = 1, \ldots c_k$ and $k = 1, \ldots, n$. Clearly, these probabilities sum to one. The missingness probabilities, assuming MAR, are:

$$p(\boldsymbol{r}|\boldsymbol{j}) = \begin{cases} p(\boldsymbol{r}|j_k \text{ with } r_k = 1) & \text{if } \boldsymbol{r} \neq \mathbf{1}, \\ 1 - \sum_{\boldsymbol{r} \neq \mathbf{1}} p(\boldsymbol{r}|\boldsymbol{j}) & \text{if } \boldsymbol{r} = \mathbf{1}. \end{cases} \tag{6.42}$$

Summing over $\boldsymbol{r}$ implies summing over those patterns for which actual observations are available. The number of parameters in the saturated model is

$$\#\text{parameters} = \left(\prod_{k=1}^{n} c_k - 1 \right) + \sum_{\boldsymbol{r} \neq \mathbf{1}} \prod_{k=1}^{n} c_k^{r_k}. \tag{6.43}$$

The first term in (6.43) is for the measurement model, the second one is for the missingness model. Clearly, the number of parameters equals one less than the number of cells, establishing the claim. The situation where covariates are present is covered automatically, merely by considering one extra dimension in the contigency table, $j = 0$ say, with c_0 referring to the total number of covariate levels in the set of data.

We will now study the implications for the simple but important setting of a bivariate contingency table with dropout, as well as a bivariate contingency table with non-monotone missingness.

A Bivariate Contingency Table With Dropout

For bivariate contingency tables with dropout, identifications can be derived by fitting the saturated MAR model, described in the previous section, to the fit obtained from the original MNAR model. Denote the counts obtained from the fit of the original model by $z_{2,jk}$ and $z_{1,j}$, for the completers and dropouts, respectively. Denote the measurement model probabilities by p_{jk} and the dropout probabilities by q_j. Then, due to ignorability, the likelihood factors into two components:

$$\ell_1 = \sum_{j,k} z_{2,jk} \ln p_{jk} + \sum_j z_{1,j} \ln p_{j+} - \lambda \left(\sum_{j,k} p_{jk} - 1 \right), \tag{6.44}$$

$$\ell_2 = \sum_{j,k} z_{2,jk} \ln q_j + \sum_j z_{1,j} \ln(1 - q_j). \tag{6.45}$$

We have used an undetermined Lagrange multiplier λ to incorporate the sum constraint on the marginal probabilities. Solving the score equations for (6.44) and (6.45) produces, with simple and well-known algebra:

$$\widehat{p_{jk}} = \frac{1}{n} z_{2,jk} \left(\frac{z_{2,j+} + z_{1,j}}{z_{2,j+}} \right), \tag{6.46}$$

$$\widehat{q_j} = \frac{z_{2,j+}}{z_{2,j+} + z_{1,j}}, \tag{6.47}$$

where n is the total sample size. Combining parameter estimates leads to the new, MAR-based, fitted counts:

$$\widehat{z_{2,jk}} = n\widehat{p_{jk}}\widehat{q_j} = z_{2,jk}, \tag{6.48}$$

$$\widehat{z_{1,jk}} = n\widehat{p_{jk}}(1 - \widehat{q_j}) = z_{1,j} \frac{z_{2,jk}}{z_{2,j+}}, \tag{6.49}$$

$$\widehat{z_{1,j+}} = z_{1,j+}. \tag{6.50}$$

From (6.48) and (6.50) it is clear that the fit in terms of the observed data has not changed. The expansion of the incomplete data into a complete one is described by (6.49). Equations (6.48) and (6.49) can be used to produce the MAR counterpart to the original model, without any additional calculations. This is not so simple for the non-monotone case, as we will show next.

A Bivariate Contingency Table With Non-Monotone Missingness

The counterparts to (6.44)–(6.45) for the case of a bivariate contingency table with non-monotone missingness are:

$$\begin{aligned}\ell_1 &= \sum_{j,k} z_{11,jk} \ln p_{jk} + \sum_j z_{10,j} \ln p_{j+} + \sum_k z_{01,k} \ln p_{+k} \\ &\quad + z_{00} \ln p_{++} - \lambda \left(\sum_{j,k} p_{jk} - 1 \right), \end{aligned} \tag{6.51}$$

$$\begin{aligned}\ell_2 &= \sum_{j,k} z_{11,jk} \ln(1 - q_{10,j} - q_{01,k} - q_{00}) + \sum_j z_{10,j} \ln q_{10,j} \\ &\quad + \sum_k z_{01,k} \ln q_{01,k} + z_{00} \ln g_{00}. \end{aligned} \tag{6.52}$$

Notation has been modified in accordance with the design. The q quantities correspond to the $g(\cdot)$ model in Section 6.3.1.

While $p_{++} = 1$ and hence z_{00} does not contribute information to the measurement probabilities, it does add to the estimation of the missingness model.

Deriving the score equations from (6.52) and (6.53) is straightforward but, unlike in the monotone case, no closed form exists. Chen and Fienberg (1974) derived an iterative scheme for the probabilities p_{jk}, based on setting the expected sufficient statistics equal to their *complete-data* counterparts:

$$np_{jk} = z_{11,jk} + z_{10,j} \frac{p_{jk}}{p_{j+}} + z_{01,k} \frac{p_{jk}}{p_{+k}} + z_{00} \frac{p_{jk}}{p_{++}},$$

(with $p_{++} = 1$) and hence

$$(n - z_{00}) p_{jk} = z_{11,jk} + z_{10,j} \frac{p_{jk}}{p_{j+}} + z_{01,k} \frac{p_{jk}}{p_{+k}}. \tag{6.53}$$

The same equation is obtained from the first derivative of (6.52). Chen and Fienberg's iterative scheme results from initiating the process with a set of starting values for the p_{jk}, e.g., from the completers, and then evaluating the right hand side of (6.53). Equating it to the left hand side provides an update for the parameters. The process is repeated until convergence.

While there are no closed-form counterparts to (6.46) and (6.47), the expressions equivalent to (6.48)–(6.50) are

$$\widehat{z_{11,jk}} = z_{11,jk}, \tag{6.54}$$
$$\widehat{z_{10,jk}} = z_{10,j}\frac{p_{jk}}{p_{j+}}, \tag{6.55}$$
$$\widehat{z_{01,jk}} = z_{01,k}\frac{p_{jk}}{p_{+k}}, \tag{6.56}$$
$$\widehat{z_{00,jk}} = z_{00}p_{jk}. \tag{6.57}$$

However, there is an important difference between (6.48)–(6.50) on the one hand and (6.54)–(6.57) on the other hand. In the monotone case, the expressions on the right hand side are in terms of the counts z only, whereas here the marginal probabilities p_{jk} intervene, which have to be determined from a numerical fit.

The practical use of the results in this section are illustrated next on data from the Slovenian public opinion survey.

6.3.3 Analysis of the Slovenian Public Opinion Survey Data

Let us illustrate the ideas developed in the second part of this chapter by means of 4 models from the BRD family, fitted to the independence and attendance outcomes from the Slovenian public opinion survey data. We select models BRD1, BRD2, BRD7, and BRD9. As already noted before, model BRD1 assumes missingness to be MCAR, whereas all others are of the MNAR type. Model BRD2 has 7 free parameters, and hence does not saturate the observed data degrees of freedom, while models BRD7 and BRD9 saturate the 8 data degrees of freedom. Each of the four models is doubled up with its MAR counterpart.

Apart from the raw data, Table 6.6 presents the fit to the observed and the hypothetical complete data for each of the models and its MAR counterpart. The fits of models BRD7, BRD9, and their MAR counterparts to the observed data, coincide with the observed data. As the theory states, every MNAR model and its MAR counterpart produce exactly the same fit to the observed data, which is therefore also seen for BRD1 and BRD2. However, while models BRD1 and BRD1(MAR) coincide in their fit to the hypothetical complete data, this is not the case for the other three models. The reason is clear: since model BRD1 belongs to the MAR family from the start, its counterpart BRD1(MAR) will not produce any difference, but merely copies the fit of BRD1 to the unobserved data, given the observed ones. Finally, while BRD7 and BRD9 produce a different fit to the complete data, BRD7(MAR) and BRD9(MAR) coincide.

Table 6.6: *Slovenian public opinion survey. Analysis is restricted to the independence and attendance questions. The observed data are shown, as well as the fit of models BRD1, BRD2, BRD7, and BRD9, and their MAR counterparts, to the observed data and to the hypothetical complete data. The contingency tables' rows (columns) correspond to 'yes'* vs. *'no' on the independence (attendance) question.*

Observed data &
Fit of BRD7, BRD7(MAR), BRD9, and BRD9(MAR) to incomplete data

1439	78	159
16	16	32

144	54	136

Fit of BRD1 and BRD1(MAR) to incomplete data

1381.6	101.7	182.9
24.2	41.4	8.1

179.7	18.3	136.0

Fit of BRD2 and BRD2(MAR) to incomplete data

1402.2	108.9	159.0
15.6	22.3	32.0

181.2	16.8	136.0

Fit of BRD1 and BRD1(MAR) to complete data

1381.6	101.7	170.4	12.5	176.6	13.0	121.3	9.0
24.2	41.4	3.0	5.1	3.1	5.3	2.1	3.6

Fit of BRD2 to complete data

1402.2	108.9	147.5	11.5	179.2	13.9	105.0	8.2
15.6	22.3	13.2	18.8	2.0	2.9	9.4	13.4

Fit of BRD2(MAR) to complete data

1402.2	108.9	147.7	11.3	177.9	12.5	121.2	9.3
15.6	22.3	13.3	18.7	3.3	4.3	2.3	3.2

Fit of BRD7 to complete data

1439	78	3.2	155.8	142.4	44.8	0.4	112.5
16	16	0.0	32.0	1.6	9.2	0.0	23.1

Fit of BRD9 to complete data

1439	78	150.8	8.2	142.4	44.8	66.8	21.0
16	16	16.0	16.0	1.6	9.2	7.1	41.1

Fit of BRD7(MAR) and BRD9(MAR) to complete data

1439	78	148.1	10.9	141.5	38.4	121.3	9.0
16	18	11.8	20.2	2.5	15.6	2.1	3.6

Table 6.7: *Slovenian public opinion survey. Summaries on each of the Models BRD1–BRD9 are presented as in Table 6.5, added with a column labelled '$\widehat{\theta}_{MAR}$', which displays the estimate of the MAR bodyguard, that is, the model corresponding to the given one, with the same fit to the observed data, but with missing data mechanism of the MAR type.*

Model	Structure	d.f.	loglik	$\widehat{\theta}$	C.I.	$\widehat{\theta}_{\text{MAR}}$
BRD1	(α, β)	6	-2495.29	0.892	[0.878;0.906]	0.8920
BRD2	(α, β_j)	7	-2467.43	0.884	[0.869;0.900]	0.8915
BRD3	(α_k, β)	7	-2463.10	0.881	[0.866;0.897]	0.8915
BRD4	(α, β_k)	7	-2467.43	0.765	[0.674;0.856]	0.8915
BRD5	(α_j, β)	7	-2463.10	0.844	[0.806;0.882]	0.8915
BRD6	(α_j, β_j)	8	-2431.06	0.819	[0.788;0.849]	0.8919
BRD7	(α_k, β_k)	8	-2431.06	0.764	[0.697;0.832]	0.8919
BRD8	(α_j, β_k)	8	-2431.06	0.741	[0.657;0.826]	0.8919
BRD9	(α_k, β_j)	8	-2431.06	0.867	[0.851;0.884]	0.8919

This is because the fits of BRD7 and BRD9 coincide with respect to their fit to the observed data, and indeed, due to their saturation, coincide with the observed data as such. This fit is the sole basis for the models' MAR extensions. It is noteworthy that, while BRD7, BRD9, and BRD7(MAR)≡BRD9(MAR) all saturate the observed data degrees of freedom, their complete-data fits are dramatically different.

Let us return to the implications of our results for the primary estimand θ, the proportion of people voting YES by simultaneously being in favor of independence and deciding to take part in the vote.

As mentioned in Section 6.1.5, the likelihood ratio tests to assess the model fit are in favour of the saturated BRD6–BRD9 models, which might give the impression that the MAR-based BRD1 estimate is not preferable. However, studying the $\widehat{\theta}_{\text{MAR}}$ values from each of the models BRD1(MAR)–BRD9(MAR), as displayed in the last column of Table 6.7, it is clear that this value is remarkably stable and hence a value of $\widehat{\theta} = 0.892$, based on the four bodyguards BRD6(MAR)–BRD9(MAR), is a sensible choice after all. Thus, a main contribution resulting from considering the bodyguards

in this particular example, is the provision of a solid basis for the MAR-based estimate. Obviously, since models BRD6(MAR)–BRD9(MAR) are exactly the same and exhibit a perfect fit, the corresponding probabilities $\widehat{\theta}_{\text{MAR}}$ are exactly equal too. In this particular case, even though BRD2(MAR)–BRD5(MAR) differ among each other, the probability of being in favor of independence and attending the plebiscite is constant across these four models. This is a mere coincidence, since all three other cell probabilities are different, but only slightly so. For example, the probability of being in favour of independence combined with not attending ranges over 0.066–0.0685 across these four models.

We have made the following two-stage use of models BRD6(MAR)–BRD9(MAR). At the first stage, in a conventional way, the fully saturated model is selected as the only adequate description of the observed data. At the second stage, these models are transformed into their MAR counterpart, from which inferences are drawn. As such, the MAR counterpart usefully supplements the original models BRD6–BRD9 and provide one further, important scenario to model the incomplete data. In principle, the same exercise can be conducted when the additional secession variable would be used.

6.4 Conclusion

In the first part of this chapter, we have given an outline of several existing MNAR models within the selection model framework. In particular, a more detailed overview is provided of the Diggle-Kenward model for continuous incomplete longitudinal data on the one hand Diggle and Kenward (1994), and of the BRD model family for two binary outcomes prone to non-monotone missingness on the other hand Baker, Rosenberger and DerSimonian (1992). Both models are illustrated through application to data from the second depression trial and the Slovenian public opinion survey, respectively. Next, a brief overview of pattern-mixture models has been given.

Further, in this chapter, we have shown that every MNAR model, fitted to a set of incomplete data, can be replaced by an MAR version which produces exactly the same fit to the observed data. There are in particular two important implications of this. First, unless one puts a priori belief in the posited MNAR model, it is not possible to use the fit of an MNAR model for or against MAR. Second, one can fit a versatile MNAR model, to ensure a good fit to the observed data, and then use the MAR version for data analysis or for sensitivity analysis.

A re-analysis of the Slovenian public opinion survey data has shown that, while a set of MNAR models produces a widely varying range of conclusions about the proportion of people who are jointly in favor of independence and plan to attend the plebiscite, the corresponding MAR models produce a very narrow range of estimates, which in addition all lie close to the outcome of the plebiscite. This provides evidence for the claim, also made in Rubin, Stern and Vehovar (1995), that choosing an MAR model as one's main route of analysis is a sensible one.

The determination of the MAR version of an MNAR model is straightforward in the case of dropout, since the ACMV restrictions, established by Molenberghs *et al.* (1998b) and translated in a computational scheme by Thijs *et al.* (2002), provides a convenient algorithm. In the case of non-monotone missingness, the marginal density of the outcomes is needed. This is straightforward when the model fitted is of the SeM type. When a PMM is fitted, the marginal density follows from a weighted sum over the pattern-specific measurement models.

While the result of Theorem 1 is general, we have focused in this chapter on SeM and PMM formulations. It is worth re-emphasizing that also the SPM is covered without any problem. In this case, the likelihood is expressed as

$$L = \prod_i \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{\theta}, \boldsymbol{b}_i) f(\boldsymbol{r}_i | \boldsymbol{\psi}, \boldsymbol{b}_i) d\boldsymbol{y}_i^m, \tag{6.58}$$

with $\boldsymbol{b}_i$ the shared parameter, often taking the form of random effects. To apply our result, $f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \widehat{\boldsymbol{\theta}}, \boldsymbol{b}_i)$ needs to be integrated over the shared parameter. The model as a whole needs to be used to produce the fit to the observed data, and then (6.13) is used to extend the observed-data fit to complete-data MAR version.

7

Sensitivity Analysis

In Section 3.1.3, we already indicated that the possibility of an underlying MNAR missingness mechanism cannot be ruled out as such. Consequently, in Chapter 6, we have given an overview of existing selection models valid under MNAR. However, while such models seem to be the proper answer to the need for more flexible models, criticisms have been formulated first and foremost by a variety of discussants to Diggle and Kenward (1994), such as Laird (1994), Little (1994b), and Rubin (1994). They claim the conclusions from such models are sensitive to model-based assumptions which cannot be checked from the data under analysis. The sensitivity of MNAR selection models was illustrated by Verbeke and Molenberghs (2000, Ch. 17), who showed that, in the context of a clinical trial in onychomycosis, excluding a small amount of measurement error, drastically changes the likelihood ratio test statistics for the MAR null hypothesis. The growing amount of modelling tools for selection models (Heckman, 1976; Diggle and Kenward, 1994) requires the understanding of such sensitivities (Glynn, Laird and Rubin, 1986), as well as tools to deal with it (Draper, 1995; Vach and Blettner, 1995; Copas and Li, 1997).

The nature of sensitivity of MNAR models originates from the fact that such an MNAR model is not fully verifiable from the data, rendering the formal distinction between MNAR and MAR missingness hard or even impossible, unless one is prepared to accept the posited MNAR model in an unquestionable way. In the previous chapter

we have shown that each MNAR model has got a corresponding MAR bodyguard, reproducing the same fit to the observed data. Additionally, this proves that one can never test the MNAR *versus* MAR hypothesis. This underscores the great sensitivity of inferences based on MNAR models to posited and unverifiable model assumptions. As a consequence, a primary (definite) analysis should not be based on a single MNAR model. A further consequence is that rather than either forgetting about or blindly shifting to an MNAR framework, the optimal place for MNAR analyses is within a sensitivity analysis context. Such analyses can be used to assess the sensitivity of inferences resulting from posited models.

In this chapter, we formulate a definition of sensitivity analyses and sketch its main strands in Section 7.1. After a short review of sensitivity analysis using the global influence approach in Section 7.2, Section 7.3 is devoted to a popular sensitivity tool based on local influence (Cook, 1986), which is applied both to the Diggle-Kenward model (Section 7.3.2) and the BRD model family (Section 7.3.3). In the latter, the local influence approach of Jansen *et al.* (2003) is extended by basing its terminology on cell counts rather than parameters, as well as by perturbing the cell probabilities rather than the model parameters. Finally, Section 7.4 is devoted to the sensitivity analyses of the second depression trial data which is based on the local influence approach (Section 7.4.1), and of the Slovenian public opinion survey data, for which several sensitivity assessments are considered (Section 7.4.2). The contribution of both analyses can be found in Shen *et al.* (2006) and Beunckens *et al.* (2007c), respectively, the latter being joint work with Cristina Sotto.

7.1 Concepts of Sensitivity Analysis

We will use the working definition that a sensitivity analysis is one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools, such as diagnostic measures. This informal definition encompasses a wide variety of useful approaches. The simplest procedure is to fit a selected number of (non-random) models, which are all deemed plausible, or in which a preferred (primary) analysis is supplemented with a number of variations. The extent to which conclusions (inferences) are stable across such ranges provides an indication about the belief that can be put into them. Variations to a basic model can be constructed in different ways.

The most obvious strategy is to consider various dependencies of the missing data process on the outcomes and/or on covariates, as was done in Section 6.1.2 and 6.1.5

for the second depression trial and the Slovenian public opinion survey, respectively. Alternatively, the distributional assumptions of the models can be changed, a route followed by, e.g., Kenward (1998) and Molenberghs, Kenward and Goetghebeur (2001a). A review of the interval-of-ignorance based sensitivity analysis of the Slovenian public opinion survey proposed by Molenberghs, Kenward and Goetghebeur (2001a) is given in Section 7.4.2. Related to this, we can assess how an MNAR model, or a collection of MNAR models, differs from the set of models with equal fit to the observed data but that are of an MAR nature, as we proposed in the previous chapter.

Additionally, a sensitivity analysis can also be performed on the level of individual observations instead of on the level of the models. In that case, interest is directed towards finding those individuals who drive the conclusions towards one or more MNAR models. Therefore, the influence of every individual separately will be explored. Two techniques exist, that is, global influence and local influence. The global influence methodology, also known as the case-deletion method (Cook and Weisberg, 1982), is introduced by Cook (1979, 1986) in linear regression, and by Thijs, Molenberghs and Verbeke (2000) and Molenberghs *et al.* (2003) in linear mixed models. We will give a review of the global influence technique in Section 7.2.

Further, several authors have advocated using local influence tools, in which one considers the impact that one or a few influential subjects might have on the model parameters, based on the specific influence assessment methodology that has been developed over the years (Cook, 1986). Applications of local influence analysis to the Diggle-Kenward model Diggle and Kenward (1994) can be found in Thijs *et al.* (2000), Verbeke *et al.* (2001b), and Molenberghs *et al.* (2001b). Similar ideas for the context of categorical longitudinal data have been developed in Van Steen *et al.* (2001) and Jansen *et al.* (2003). In particular, Van Steen *et al.* (2001) adapted the local influence ideas to the model of Molenberghs, Kenward and Lesaffre (1997) for monotone repeated ordinal data, whereas Jansen *et al.* (2003) applied the local influence tool to the family of BRD models (Baker, Rosenberger and DerSimonian, 1992) for two binary outcomes prone to non-monotone missingness. Hens *et al.* (2005) proposed kernel weighted influence measures. Local influence is the topic of Section 7.3.

7.2 Global Influence as a Sensitivity Tool

Let us give a short review of global influence, one of the tools to perform a sensitivity analysis with an eye on individual observations, starting from case deletion. The methodology is based on the difference in log-likelihood between the model fitted to

the entire data set on the one hand, and the data set minus one subject on the other hand. One might also consider, as we do here, the reverse operation of adding single case. Denoting by $\ell_i(\boldsymbol{\phi})$ the contribution of the ith individual to the log-likelihood, where $\boldsymbol{\phi}$ is the s-dimensional vector of unknown parameters of the particular model, the complete log-likelihood is

$$\ell(\boldsymbol{\phi}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\phi}). \tag{7.1}$$

Further, denote by

$$\ell_{(\pm i)}(\boldsymbol{\phi}) \tag{7.2}$$

the log-likelihood function, where the contribution of the ith subject has been removed $(-i)$ or added $(+i)$. Cook's distances (CD) are based on measuring the discrepancy between either the maximized log-likelihoods (7.1) and (7.2) or (subsets of) the estimated parameter vectors $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\phi}}_{(\pm i)}$, with obvious notation. Precisely, we can consider

$$CD_{1i}(\boldsymbol{\phi}) = 2\left[\widehat{\ell}(\boldsymbol{\phi}) - \widehat{\ell}_{(\pm i)}(\boldsymbol{\phi})\right], \tag{7.3}$$

or

$$CD_{2i}(\boldsymbol{\phi}) = 2(\widehat{\boldsymbol{\phi}} - \widehat{\boldsymbol{\phi}}_{(\pm i)})' \, \ddot{L}^{-1} \, (\widehat{\boldsymbol{\phi}} - \widehat{\boldsymbol{\phi}}_{(\pm i)}), \tag{7.4}$$

with $\ddot{L}$ the matrix of second-order derivatives of $\ell(\boldsymbol{\phi})$, with respect to $\boldsymbol{\phi}$, evaluated at $\widehat{\boldsymbol{\phi}}$.

7.3 Local Influence as a Sensitivity Tool

A drawback of global influence is that the specific cause of the influence is hard to retrieve since, by deleting or adding a subject, all types of influence stemming from it are lumped together. Local influence, studying the effect of infinitesimally small model perturbations around a given null model, is more suitable for this purpose.

The original goal of local influence methods for sensitivity analysis was detection of observations with a high impact on the conclusions *due to their aberrant missingness mechanism*. A motivating scenario for this was one where most missing measurements might be MAR, with a few being MNAR. However, in most successful applications, where a seemingly MNAR mechanism turned out to be MAR or even MCAR after removing the influential subjects identified upon the use of local influence, the situation turned out to be more complex than anticipated. The influential subjects often are influential for other than missingness related features. For example, in the mastitis dataset of Molenberghs *et al.* (2001b), the three influential cows had complete data but were identified by an extreme increase between the measurements at

two subsequent years. Jansen *et al.* (2006b) concluded that local influence tools in the incomplete data context are useful, not to detect individuals that drop out non-randomly, but rather to identify anomalous subjects that seemingly lead to MNAR. A careful study of such subjects, combined with appropriate treatment (e.g., correction of errors, removal, etc.), can lead to an appropriate level of confidence in the originally proposed, perhaps MAR, primary analysis. Identifying and further studying the subjects that drive the missing data conclusions may shed light on, for example, trial conduct, differential effect of therapy in sub-classes of subjects, etc.

The main ideas of local influence are discussed in Section 7.3.1. Afterwards, we consider the application of the local influence tool to the Diggle-Kenward model for repeated continuous outcomes in Section 7.3.2, as well as to the family of BRD models for two binary outcomes in Section 7.3.3.

7.3.1 Concepts of Local Influence

Let us first review the key concepts of local influence (Cook, 1986). As before, $\ell(\boldsymbol{\phi})$ represents the log-likelihood function of the posited null model. Further, we denote the log-likelihood function corresponding to the perturbed model, in which the null model is nested, by $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\phi}|\omega_i)$, in which $\ell_i(\boldsymbol{\phi}|\omega_i)$ is the contribution of the ith individual, and where $\boldsymbol{\phi}$=($\boldsymbol{\theta}$,$\boldsymbol{\psi}$) is the s-dimensional vector, grouping, respectively, the parameters of the measurement and dropout models, but not including the $N \times 1$ vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)'$ of weights defining the perturbation. Assume that $\boldsymbol{\omega}$ belongs to an open subset Ω of $\mathbb{R}^N$. For $\boldsymbol{\omega}$ equal to $\boldsymbol{\omega}_o = (0, 0, \ldots, 0)'$, $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$ is the log-likelihood corresponding to the simpler of the two models.

Let $\widehat{\boldsymbol{\phi}}$ be the maximum likelihood estimator for $\boldsymbol{\phi}$, obtained by maximizing $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$, and let $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ denote the maximum likelihood estimator for $\boldsymbol{\phi}$ under $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$. The local influence approach compares $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ with $\widehat{\boldsymbol{\phi}}$. Similar values will indicate that the parameter estimates are robust with respect to perturbations in the direction of the extended model. Cook (1986) proposed to measure the distance between $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\phi}}$ by the so-called likelihood displacement, defined as $LD(\boldsymbol{\omega}) = 2[\ell(\widehat{\boldsymbol{\phi}}|\boldsymbol{\omega}_o) - \ell(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}_o)]$. This takes into account the variability of $\widehat{\boldsymbol{\phi}}$. Indeed, $LD(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$ is strongly curved at $\widehat{\boldsymbol{\phi}}$, which means that $\boldsymbol{\phi}$ is estimated with high precision, and small otherwise. Therefore, a graph of $LD(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ contains essential information on the influence perturbations. It is useful to view this graph as the geometric surface formed by values of the $N + 1$ dimensional vector $\zeta(\boldsymbol{\omega}) = (\boldsymbol{\omega}', LD(\boldsymbol{\omega}))'$ as $\boldsymbol{\omega}$ varies throughout Ω. Since this so-called *influence graph* (Lesaffre and Verbeke, 1998) can only be depicted when $N = 2$, Cook (1986) proposed to consider local influence, that

is, at the normal curvatures $C_{\boldsymbol{h}}$ of $\zeta(\boldsymbol{\omega})$ in $\boldsymbol{\omega}_o$, in the direction of some N-dimensional vector $\boldsymbol{h}$ of unit length. Let $\boldsymbol{\Delta}_i$ be the s-dimensional vector defined by

$$\boldsymbol{\Delta}_i = \left. \frac{\partial^2 \ell_i(\boldsymbol{\phi}|\omega_i)}{\partial \omega_i \partial \boldsymbol{\phi}} \right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}},\omega_i=0}, \tag{7.5}$$

and define $\boldsymbol{\Delta}$ as the $(s \times N)$ matrix with $\boldsymbol{\Delta}_i$ as its ith column. Let $\ddot{L}$ denote the $(s \times s)$ matrix of second order derivatives of $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$ with respect to $\boldsymbol{\phi}$, also evaluated at $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$. Cook (1986) has then shown that $C_{\boldsymbol{h}}$ can easily be calculated by

$$C_{\boldsymbol{h}} = 2 \left| \boldsymbol{h}' \, \boldsymbol{\Delta}' \, (\ddot{L})^{-1} \, \boldsymbol{\Delta} \, \boldsymbol{h} \right|, \tag{7.6}$$

for any direction $\boldsymbol{h}$. One evident choice is the vector $\boldsymbol{h}_i$ containing one in the ith position and zero elsewhere, and corresponding to the perturbation of the ith subject only, thereby reflecting the influence of allowing the ith subject to drop out in a more general fashion than the others. The corresponding local influence measure, denoted by C_i, then becomes $C_i = 2 \left| \boldsymbol{\Delta}_i' \, (\ddot{L})^{-1} \, \boldsymbol{\Delta}_i \right|$. Another important direction is the direction $\boldsymbol{h}_{\max}$ of maximal normal curvature $C_{\max}$. It shows how to perturb the model to obtain the largest local changes in the likelihood displacement. It is readily seen that $C_{\max}$ is the largest eigenvalue of $-2\boldsymbol{\Delta}'(\ddot{L})^{-1}\boldsymbol{\Delta}$, with $\boldsymbol{h}_{\max}$ the corresponding eigenvector. Calculation of local influence measures reduces to evaluation of $\boldsymbol{\Delta}$ and $\ddot{L}$ and a convenient computational scheme can be used whenever a program is available to fit the full alternative model, since it then suffices to compute the second derivative at $(\widehat{\boldsymbol{\phi}}, \omega_i = 0)$, for each observation separately, from which the $\boldsymbol{\Delta}_i = (\boldsymbol{\phi}, \omega_i)$ subvector is selected.

It should be noted that $C_{\boldsymbol{h}}$ is a measure of the local influence on the log-likelihood function, that is, quantifying the effect of perturbations in terms of the displacement in the log-likelihood. At times, however, it might be more meaningful to assess the influence that infinitesimal changes may have on a particular function of the parameters, rather than on the log-likelihood itself. In the case of contingency tables, for instance, one might be more interested in the impact of perturbations on the predicted cell counts, $Z_{r_1 r_2, j_1 j_2}$, which are functions of the parameter vector $\boldsymbol{\phi}$. If we denote a particular function of the model parameters by $Z(\boldsymbol{\phi})$, we can consider the local influence of the perturbations around the posited null model in terms of the difference between this function evaluated in $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\phi}}$, that is $\hat{Z}(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}})$ and $\hat{Z}(\widehat{\boldsymbol{\phi}})$. Analogous to Cook's reasoning, we propose to measure this discrepancy by

$$D(\boldsymbol{\omega}) = 2[\hat{Z}(\widehat{\boldsymbol{\phi}}) - \hat{Z}(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}})].$$

The influence graph of $D(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ contains again essential information on the influence perturbations. Therefore, the local influence is depicted by the normal

curvatures of such influence graphs at $\boldsymbol{\omega}_0$, in the direction of some vector of unit length. In a similar way as Cook, we can show that the expression for $C_{\boldsymbol{h}}$ is now generalized to

$$C_{\boldsymbol{h}} = 2 \left| h' \, \boldsymbol{\Delta}' \, (\ddot{\boldsymbol{L}})^{-1} \, \ddot{\boldsymbol{Z}} \, (\ddot{\boldsymbol{L}})^{-1} \, \boldsymbol{\Delta} \, h \right|, \tag{7.7}$$

with $\|h\| = 1$, $\boldsymbol{\Delta}$ and $\ddot{\boldsymbol{L}}$ as before, and $\ddot{\boldsymbol{Z}}$ the $(s \times s)$ matrix of second order derivatives of $Z(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ and evaluated at $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$. It can easily be seen that the expression (7.7) reduces to (7.6) when the function of interest, $Z(\boldsymbol{\phi})$, is the log-likelihood $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$ itself. Whereas (7.6) quantifies influence in terms of the displacement in the log-likelihood function, (7.7) describes influence through the displacement in the particular function of interest.

Note that since the resulting influence diagnostics can in many cases be expressed analytically, they often allow for a decomposition into interpretable components, thus yielding additional insight. For instance when a subset $\boldsymbol{\phi}_1$ of $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \boldsymbol{\phi}_2')'$ is of special interest, a similar approach can be used, replacing the log-likelihood by the profile log-likelihood for $\boldsymbol{\phi}_1$, and the methods discussed above for the full parameter vector directly carry over (Lesaffre and Verbeke, 1998).

7.3.2 Applied to the Diggle-Kenward Model

Verbeke *et al.* (2001b), Thijs, Molenberghs and Verbeke (2000), Molenberghs *et al.* (2001b), and Jansen *et al.* (2006b) investigated sensitivity of estimation of quantities of interest, such as treatment effect, growth parameters, or the dropout model parameters, with respect to the dropout model assumptions considering the full selection Diggle-Kenward model, discussed in Section 6.1.1. To this end, they considered the following perturbed version of dropout model (6.4):

$$\text{logit}\left[P(D_i = j \mid D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})\right] \;=\; \psi_0 + \psi_1 y_{i,j-1} + \omega_i y_{ij}. \tag{7.8}$$

where the ω_i are local, individual-specific perturbations around a null model. They should not be confused with subject-specific parameters. The null model will be the MAR model, corresponding to setting $\psi_2 = 0$ in (6.4).

Using this proposal, one can study the impact on key model features, induced by small perturbations in the direction, or seemingly so, of MNAR. This can practically be done by constructing local influence measures as shown in Section 7.3.1. When small perturbations in a specific ω_i lead to relatively large differences in the model parameters, this suggests that the subject is likely to drive key conclusions. For example, if such a subject would drive the model towards MNAR, then the conditional

expectations of the unobserved measurements, given the observed ones, may deviate substantially from the ones under an MAR mechanism (Kenward, 1998).

Some caution is needed when interpreting local influence. Even though we may be tempted to conclude that an influential subject drops out non-randomly, this conclusion is misguided since we are not aiming to detect (groups of) subjects that drop out non-randomly but rather subjects that have a considerable impact on the dropout and measurement model parameters. In other words, such subjects drive the sensitivity of the analysis to missing data assumptions (Jansen *et al.*, 2006b).

Let us now apply the local influence sensitivity tool to the Diggle-Kenward model. All derivations and calculations here are valid in the general case and can be implemented in statistical software. Details on the implementation of this approach in the SAS software using IML can be found in Section 11.6. While fully generally valid, for clarity of exposition, we present our calculations for the specific case of three measurements.

As described in Section 6.1.1, the Diggle-Kenward model combines a multivariate normal model for the measurement process and a logistic regression model for the dropout process. We denote $P(D_i = j|D_i \geq j, \boldsymbol{h}_{ij}, y_{ij})$, that is, the conditional probability for dropout at occasion j, given that the subject was still observed at the previous occasion, which is allowed to depend on the history $\underline{\boldsymbol{h}}_{ij}$ and the possibly unobserved current outcome y_{ij}, by $g(\underline{\boldsymbol{h}}_{ij}, y_{ij})$. The parameter dependencies are suppressed for notational ease. In this case, the marginal probability of dropout at each occasion, as given in (6.3) can be rewritten as

$$
\begin{aligned}
f(d_i|\boldsymbol{y}_i) &= P(D_i = d_i|\boldsymbol{y}_i) \qquad (7.9)\\
&= \begin{cases}
\displaystyle\prod_{j=2}^{n}[1 - g(\boldsymbol{h}_{ij}, y_{ij})] & \text{for a completer } (d_i = n+1),\\
g(\boldsymbol{h}_{id_i}, y_{id_i}) \displaystyle\prod_{j=2}^{d_i-1}[1 - g(\boldsymbol{h}_{ij}, y_{ij})] & \text{for a dropout } (d_i \leq n),
\end{cases}
\end{aligned}
$$

When denoting $\ell_i(\boldsymbol{\phi}|\omega_i)$ by $\ell_{i\omega}$, the log-likelihood contribution of a complete sequence is given by

$$\ell_{i\omega} = \ln f(\boldsymbol{y}_i) + \sum_{j=2}^{n} \ln[1 - g(\boldsymbol{h}_{ij}, y_{ij})],$$

in which the density $f(\boldsymbol{y}_i)$ is multivariate normal, following from the linear mixed model. The contribution from an incomplete sequence is more complicated.

Its log-likelihood term is

$$\begin{aligned}\ell_{i\omega} &= \ln f(y_{i1},\ldots,y_{i,d_i-1}) + \sum_{j=2}^{d_i-1} \ln[1-g(\boldsymbol{h}_{ij},y_{ij})] \\ &\quad + \ln \int f(y_{id_i}|y_{i1},\ldots,y_{i,d_i-1}) g(\boldsymbol{h}_{id_i},y_{id_i}) dy_{id_i}.\end{aligned}$$

Further details can be found in Verbeke *et al.* (2001b). We need expressions for $\boldsymbol{\Delta}$ and $\ddot{L}$. Straightforward derivation shows that the columns $\boldsymbol{\Delta}_i$ of $\boldsymbol{\Delta}$ are given by

$$\left.\frac{\partial^2 \ell_{i\omega}}{\partial\boldsymbol{\theta}\partial\omega_i}\right|_{\omega_i=0} = \mathbf{0}, \tag{7.10}$$

$$\left.\frac{\partial^2 \ell_{i\omega}}{\partial\boldsymbol{\psi}\partial\omega_i}\right|_{\omega_i=0} = -\sum_{j=2}^{n} \boldsymbol{h}_{ij} y_{ij} g(\boldsymbol{h}_{ij})[1-g(\boldsymbol{h}_{ij})], \tag{7.11}$$

for complete sequences (no drop out) and by

$$\left.\frac{\partial^2 \ell_{i\omega}}{\partial\boldsymbol{\theta}\partial\omega_i}\right|_{\omega_i=0} = [1-g(\boldsymbol{h}_{id_i})]\frac{\partial\lambda(y_{id_i}|\boldsymbol{h}_{id_i})}{\partial\boldsymbol{\theta}}, \tag{7.12}$$

$$\begin{aligned}\left.\frac{\partial^2 \ell_{i\omega}}{\partial\boldsymbol{\psi}\partial\omega_i}\right|_{\omega_i=0} &= -\sum_{j=2}^{d_i-1} \boldsymbol{h}_{ij} y_{ij} g(\boldsymbol{h}_{ij})[1-g(\boldsymbol{h}_{ij})] \\ &\quad -\boldsymbol{h}_{id_i}\lambda(y_{id_i}|\boldsymbol{h}_{id_i}) g(\boldsymbol{h}_{id_i})[1-g(\boldsymbol{h}_{id_i})],\end{aligned} \tag{7.13}$$

for incomplete sequences. All above expressions are evaluated at $\widehat{\boldsymbol{\phi}}$, and $g(\boldsymbol{h}_{ij}) = g(\boldsymbol{h}_{ij},y_{ij})|_{\omega_i=0}$, is the MAR version of the dropout model. In (7.12), we make use of the conditional mean

$$\lambda(y_{id_i}|\boldsymbol{h}_{id_i}) = \lambda(y_{id_i}) + V_{i,21}V_{i,11}^{-1}[\boldsymbol{h}_{id_i} - \lambda(\boldsymbol{h}_{id_i})]. \tag{7.14}$$

The variance matrices follow from partitioning the responses as $(y_{i1},\ldots,y_{i,d_i-1}|y_{id_i})'$.

Let $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ indicate the subvector of the fixed-effect and covariance parameters respectively, within the vector $\boldsymbol{\theta}$ of the measurement model parameters. The derivatives of (7.14) with respect to these measurement model parameters are given by

$$\begin{aligned}\frac{\partial\lambda(y_{id_i}|\boldsymbol{h}_{id_i})}{\partial\boldsymbol{\beta}} &= \boldsymbol{x}_{id_i} - V_{i,21}V_{i,11}^{-1}X_{i,(d_i-1)}, \\ \frac{\partial\lambda(y_{id_i}|\boldsymbol{h}_{id_i})}{\partial\boldsymbol{\alpha}} &= \left[\frac{\partial V_{i,21}}{\partial\boldsymbol{\alpha}} - V_{i,21}V_{i,11}^{-1}\frac{\partial V_{i,11}}{\partial\boldsymbol{\alpha}}\right] V_{i,11}^{-1}[\boldsymbol{h}_{id_i} - \lambda(\boldsymbol{h}_{id_i})]\end{aligned}$$

where $\boldsymbol{x}'_{id_i}$ is the d_ith row of X_i, and where $X_{i,(d_i-1)}$ indicates the first (d_i-1) rows X_i.

In practice, the measurement model parameters $\boldsymbol{\theta}$ are often of primary interest. Since $\ddot{L}$ is block-diagonal with blocks $\ddot{L}(\boldsymbol{\theta})$ and $\ddot{L}(\boldsymbol{\psi})$, we have that for any unit vector $\boldsymbol{h}$, $C_{\boldsymbol{h}}$ equals $C_{\boldsymbol{h}}(\boldsymbol{\theta}) + C_{\boldsymbol{h}}(\boldsymbol{\psi})$, with

$$C_{\boldsymbol{h}}(\boldsymbol{\theta}) = -2\boldsymbol{h}' \left[\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\theta} \partial \omega_i}\bigg|_{\omega_i=0} \right]' \ddot{L}^{-1}(\boldsymbol{\theta}) \left[\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\theta} \partial \omega_i}\bigg|_{\omega_i=0} \right] \boldsymbol{h} \tag{7.15}$$

$$C_{\boldsymbol{h}}(\boldsymbol{\psi}) = -2\boldsymbol{h}' \left[\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\psi} \partial \omega_i}\bigg|_{\omega_i=0} \right]' \ddot{L}^{-1}(\boldsymbol{\psi}) \left[\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\psi} \partial \omega_i}\bigg|_{\omega_i=0} \right] \boldsymbol{h}, \tag{7.16}$$

evaluated at $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$.

A Special Case of Three Measurements

We will now consider the special but insightful case of three measurement occasions, using the three-dimensional version of (6.3), where V_i follows a heterogeneous first-order autoregressive structure:

$$V_i = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}.$$

Recall that we assume the dropout model to be of the form (6.4), that is,

$$\text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})\right] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}.$$

Since there are three measurements (Y_{i1}, Y_{i2}, Y_{i3}), three different situations can arise: (1) all three measurements are available (a completer), (2) only the first two measurements are available and the subject drops out after the second time point, or (3) only the first measurement is available, which means the subject drops out after the first measurement. In the first case, the components of the columns $\boldsymbol{\Delta}_i$ of $\boldsymbol{\Delta}$ are given by (7.10) and (7.11), and using (7.8) we get

$$g(\boldsymbol{h}_{ij}) = g(\boldsymbol{h}_{ij}, y_{ij})|_{\omega_i=0} = \frac{\exp(\psi_0 + \psi_1 y_{i,j-1})}{1 + \exp(\psi_0 + \psi_1 y_{i,j-1})}.$$

In the case of dropout, the components of the columns $\boldsymbol{\Delta}_i$ of $\boldsymbol{\Delta}$ are given by (7.12) and (7.13), which means we need $V_{i,11}$, $V_{i,21}$, and their derivatives with respect to the four variance components σ_1, σ_2, σ_3, and ρ.

Now, to get expressions for $\boldsymbol{h}_{id_i}$, $\lambda(y_{id_i})$, $\lambda(\boldsymbol{h}_{id_i})$, $V_{i,11}$, $V_{i,21}$, and their derivatives, we distinguish between dropout after the first measurement ($d_i = 2$) and dropout after the second one ($d_i = 3$).

When $d_i = 2$, we have $\boldsymbol{h}_{id_i} = y_{i1}$, and since the mean of the measurement model is $\boldsymbol{X}_i\boldsymbol{\beta}$, $\lambda(y_{id_i})$ equals the second value of $\boldsymbol{X}_i\boldsymbol{\beta}$, whereas $\lambda(\boldsymbol{h}_{id_i})$ equals the first value of $\boldsymbol{X}_i\boldsymbol{\beta}$. Further, $V_{i,11} = \sigma_1^2$, and $V_{i,21} = \rho\sigma_1\sigma_2$, and thus the derivatives are

$$\frac{\partial V_{i,11}}{\partial \sigma_1} = 2\sigma_1, \frac{\partial V_{i,11}}{\partial \sigma_2} = \frac{\partial V_{i,11}}{\partial \sigma_3} = \frac{\partial V_{i,11}}{\partial \rho} = 0,$$
$$\frac{\partial V_{i,21}}{\partial \sigma_1} = \rho\sigma_1, \frac{\partial V_{i,21}}{\partial \sigma_2} = \rho\sigma_2, \frac{\partial V_{i,21}}{\partial \sigma_3} = 0, \frac{\partial V_{i,21}}{\partial \rho} = \sigma_1\sigma_2.$$

Next, in case $d_i = 3$, we have $\boldsymbol{h}_{id_i} = (y_{i1}, y_{i2})'$, $\lambda(y_{id_i})$ equals the third value of $\boldsymbol{X}_i\boldsymbol{\beta}$, whereas $\lambda(\boldsymbol{h}_{id_i})$ is the vector of the first and second values of $\boldsymbol{X}_i\boldsymbol{\beta}$. Further,

$$V_{i,11} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad \text{and} \quad V_{i,21} = \begin{pmatrix} \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 \end{pmatrix},$$

and thus the derivatives are written as

$$\frac{\partial V_{i,11}}{\partial \sigma_1} = \begin{pmatrix} 2\sigma_1 & \rho\sigma_2 \\ \rho\sigma_2 & 0 \end{pmatrix}, \frac{\partial V_{i,11}}{\partial \sigma_2} = \begin{pmatrix} 0 & \rho\sigma_1 \\ \rho\sigma_1 & 2\sigma_2 \end{pmatrix}, \frac{\partial V_{i,11}}{\partial \sigma_3} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\frac{\partial V_{i,11}}{\partial \rho} = \begin{pmatrix} 0 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & 0 \end{pmatrix}, \frac{\partial V_{i,21}}{\partial \rho} = \begin{pmatrix} 2\rho\sigma_1\sigma_3 & \sigma_2\sigma_3 \end{pmatrix}.$$

$$\frac{\partial V_{i,21}}{\partial \sigma_1} = \begin{pmatrix} \rho^2\sigma_3 & 0 \end{pmatrix}, \frac{\partial V_{i,21}}{\partial \sigma_2} = \begin{pmatrix} 0 & \rho\sigma_3 \end{pmatrix}, \frac{\partial V_{i,21}}{\partial \sigma_3} = \begin{pmatrix} \rho^2\sigma_1 & \rho\sigma_2 \end{pmatrix},$$

Using all of this information, we can easily derive expressions (7.12) and (7.13), that is, the components of the columns $\boldsymbol{\Delta}_i$ of $\boldsymbol{\Delta}$.

7.3.3 Applied to the BRD Model Family

Jansen *et al.* (2003) developed a local influence approach for binary data, subject to non-monotone missingness, based on the BRD model family. These authors focus on perturbations of a given BRD model in the direction of an alternative model with one additional parameter. First, we provide a review of this technique, whereafter we additionally consider perturbations in the observed cell probabilities, rather than the parameters of the model.

Perturbation in Parameters: One BRD Model vs. Another

In line with Jansen *et al.* (2003), we consider perturbations of a given BRD model in the direction of another BRD model with one or more parameters in which the

first model is nested, implying that perturbations lie along the edges of Figure 6.1. For such a nested pair, the simpler of the two models equates two parameters from the more complex one. For example, BRD4 includes the parameter $\beta_{.j_2}, (j_2 = 1, 2)$, whereas for BRD1, only $\beta_{..}$ is included. For the influence analysis, ω_i is then included as a contrast between two such parameters; for the perturbation of BRD1 in the direction of BRD4, one considers $\beta_{..}$ and $\beta_{..} + \omega_i$. The vector of all ω_i's defines the direction in which such a perturbation is considered.

To illustrate this approach, we begin by first defining the log-likelihood for the BRD family of models. We have

$$\begin{aligned} \ell(\boldsymbol{\phi}|\boldsymbol{\omega}) &= \sum_{j_1,j_2} Z_{11,j_1j_2} \ln \pi_{11,j_1j_2} + \sum_{j_1} Z_{10,j_1+} \ln \pi_{10,j_1+} \\ &+ \sum_{j_2} Z_{01,+j_2} \ln \pi_{01,+j_2} + Z_{00,++} \ln \pi_{00,++}, \end{aligned} \tag{7.17}$$

where $\pi_{r_1r_2,j_1j_2} = p_{j_1j_2}\, q_{r_1r_2|j_1j_2}$, with $p_{j_1j_2}$ and $q_{r_1r_2|j_1j_2}$ as in (6.6) and (6.7).

Distinction among the 9 BRD models occurs in expression (6.7) describing $q_{r_1r_2|j_1j_2}$. For instance, for BRD4 with $(\alpha_{..}, \beta_{.j_2})$, this expression yields:

$$q_{r_1r_2|j_11} = \frac{\exp\{\alpha_{..}(1-r_1) + \beta_{..}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..}) + \exp(\beta_{..}) + \exp(\alpha_{..} + \beta_{..} + \gamma)},$$

$$q_{r_1r_2|j_12} = \frac{\exp\{\alpha_{..}(1-r_1) + (\beta_{..} + \omega_i)(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..}) + \exp(\beta_{..} + \omega_i) + \exp(\alpha_{..} + \beta_{..} + \omega_i + \gamma)}.$$

Note that, for $\omega_i = 0$, the two previous expressions are equivalent and BRD4 reduces to the simpler BRD1. For this pair of nested models, BRD4 contains one more parameter compared to BRD1, this extra parameter being the distinguishing feature between both models. That is, under the more complicated model BRD4, the extra parameter ω_i defines a difference between the dropout probabilities above, while under the simpler (null) model BRD1, the two expressions reduce to a single dropout probability. Similar motivations hold for the other pairs of nested BRD models. Given now the fully-defined log-likelihood, one can proceed with deriving local influence measures (7.6) and (7.7).

Note that the influence analysis focuses on the missingness model, rather than on the measurement model parameters. This may be seen as slightly odd, since often scientific interest focuses on the measurement model parameters. However, it has been documented (Rubin, 1994; Kenward, 1998; Verbeke *et al.*, 2001b) that the missingness model parameters are often the most sensitive ones to take up all kinds of misspecification and influential features. These may then, in turn, impact

conclusions coming from the measurement model parameters, such as time evolution, or combinations thereof, such as covariate effects for certain groups of responders.

Perturbation in Cell Probabilities

Another route to studying local influence is to add an infinitesimally small value to the cell probabilities. Such an approach leads to the following expression for the log-likelihood:

$$\begin{aligned}\ell(\boldsymbol{\phi}|\boldsymbol{\omega}) &= \sum_{j_1,j_2} (Z_{11,j_1j_2} + N\omega_{11,j_1j_2}) \ln \pi_{11,j_1j_2} \\ &+ \sum_{j_1} (Z_{10,j_1+} + N\omega_{10,j_1+}) \ln \pi_{10,j_1+} \\ &+ \sum_{j_2} (Z_{01,+j_2} + N\omega_{01,+j_2}) \ln \pi_{01,+j_2} \\ &+ (Z_{00,++} + N\omega_{00,++}) \ln \pi_{00,++},\end{aligned}$$

with $\pi_{r_1r_2,j_1j_2}$ as before. It is important to note that the previously described approach of local influence differs from the approach proposed here, since now the perturbation is done directly in the observed cell probabilities, rather than the parameters of the model. This implies that we are perturbing the cells one at a time and observing which one brings about the largest changes, in likelihood or in the predicted cell counts, within a given BRD model. Consequently, although influence measures are computed in the same fashion, a difference in interpretation is warranted. A peak in the influence curve now represents the particular observed cell at which a probability perturbation causes substantial displacement in either the log-likelihood or in the predicted cell counts.

Computation of local influence measures (7.6) and (7.7) is straightforward once the log-likelihood, $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$, is clearly defined. Let us now show the calculations of the necessary derivatives.

Derivatives of the Log-Likelihood Function

In order to compute the local influence measures (7.6) and (7.7), we need to calculate the second order partial derivatives of $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$ with respect to the elements of $\boldsymbol{\omega}$ and $\boldsymbol{\phi}$, that is,

$$\frac{\partial^2 \ell_\omega}{\partial \phi_i \partial \omega_{r_1r_2,j_1j_2}}.$$

Let us assume no covariates in the measurement model, resulting in three measurement parameters p_{11}, p_{12}, and p_{21}.

First, we take derivatives with respect to $\boldsymbol{\omega}$, after which we differentiate this with respect to $\boldsymbol{\phi}$, resulting in

$$\frac{\partial \ell_\omega}{\partial \omega_{r_1r_2,j_1j_2}} = N \ln \pi_{r_1r_2,j_1j_2} \qquad \text{and} \qquad \frac{\partial^2 \ell_\omega}{\partial \phi_i \partial \omega_{r_1r_2,j_1j_2}} = \frac{N}{\pi_{r_1r_2,j_1j_2}} \frac{\partial \pi_{r_1r_2,j_1j_2}}{\partial \phi_i},$$

and for the different missingness mechanisms (r_1r_2) this becomes

$$(11) \quad : \quad \frac{\partial^2 \ell_\omega}{\partial \boldsymbol{\phi} \partial \omega_{11,j_1j_2}} = \frac{N}{\pi_{11,j_1j_2}} \frac{\partial \pi_{11,j_1j_2}}{\partial \boldsymbol{\phi}},$$

$$(10) \quad : \quad \frac{\partial^2 \ell_\omega}{\partial \boldsymbol{\phi} \partial \omega_{10,j+}} = \frac{N}{\pi_{10,j+}} \left(\frac{\partial \pi_{10,j_1 1}}{\partial \boldsymbol{\phi}} + \frac{\partial \pi_{10,j_1 2}}{\partial \boldsymbol{\phi}} \right),$$

$$(01) \quad : \quad \frac{\partial^2 \ell_\omega}{\partial \boldsymbol{\phi} \partial \omega_{01,+k}} = \frac{N}{\pi_{01,+k}} \left(\frac{\partial \pi_{01,1 j_2}}{\partial \boldsymbol{\phi}} + \frac{\partial \pi_{01,2 j_2}}{\partial \boldsymbol{\phi}} \right),$$

$$(00) \quad : \quad \frac{\partial^2 \ell_\omega}{\partial \boldsymbol{\phi} \partial \omega_{00,++}} = \frac{N}{\pi_{00,++}} \left(\frac{\partial \pi_{00,11}}{\partial \boldsymbol{\phi}} + \frac{\partial \pi_{00,12}}{\partial \boldsymbol{\phi}} + \frac{\partial \pi_{00,21}}{\partial \boldsymbol{\phi}} + \frac{\partial \pi_{00,22}}{\partial \boldsymbol{\phi}} \right),$$

respectively. The next step is to calculate the partial derivatives of $\pi_{r_1r_2,j_1j_2}$ with respect to the elements of $\boldsymbol{\phi}$:

$$(a) \quad : \quad \frac{\partial \pi_{r_1r_2,j_1j_2}}{\partial p_{11}} = \begin{cases} q_{r_1r_2|j_1j_2} & \text{if } (j_1,j_2) = (1,1) \\ -q_{r_1r_2|j_1j_2} & \text{if } (j_1,j_2) = (2,2) \\ 0 & \text{if } (j_1,j_2) = (1,2) \text{ or } (2,1) \end{cases}$$

$$(b) \quad : \quad \frac{\partial \pi_{r_1r_2,j_1j_2}}{\partial p_{12}} = \begin{cases} q_{r_1r_2|j_1j_2} & \text{if } (j_1,j_2) = (1,2) \\ -q_{r_1r_2|j_1j_2} & \text{if } (j_1,j_2) = (2,2) \\ 0 & \text{if } (j_1,j_2) = (1,1) \text{ or } (2,1) \end{cases}$$

$$(c) \quad : \quad \frac{\partial \pi_{r_1r_2,j_1j_2}}{\partial p_{21}} = \begin{cases} q_{r_1r_2|j_1j_2} & \text{if } (j_1,j_2) = (2,1) \\ -q_{r_1r_2|j_1j_2} & \text{if } (j_1,j_2) = (2,2) \\ 0 & \text{if } (j_1,j_2) = (1,1) \text{ or } (1,2) \end{cases}$$

$$(d) \quad : \quad \frac{\partial \pi_{r_1r_2,j_1j_2}}{\partial \alpha_{j_1'j_2'}} = p_{j_1j_2} \cdot \frac{\partial q_{r_1r_2|j_1j_2}}{\partial \alpha_{j_1'j_2'}}$$

$$(e) \quad : \quad \frac{\partial \pi_{r_1r_2,j_1j_2}}{\partial \beta_{j_1'j_2'}} = p_{j_1j_2} \cdot \frac{\partial q_{r_1r_2|j_1j_2}}{\partial \beta_{j_1'j_2'}}$$

$$(f) \quad : \quad \frac{\partial \pi_{r_1r_2,j_1j_2}}{\partial \gamma} = p_{j_1j_2} \cdot \frac{\partial q_{r_1r_2|j_1j_2}}{\partial \gamma}$$

Finally, in previous derivatives $(d) - (f)$, we need expressions for the derivatives of $q_{r_1r_2|j_1j_2}$ with respect to $\alpha_{j_1'j_2'}$, $\beta_{j_1'j_2'}$ and γ. First note that for all missingness patterns, the derivatives with respect to $\alpha_{j_1'j_2'}$ and $\beta_{j_1'j_2'}$ equal zero if $(j_1', j_2') \neq (j_1, j_2)$. Further, we will show the computations for the pattern of completers, that is, $(r_1r_2) = (11)$, as well as for the incomplete pattern for which the first outcome is observed, and the second one is not, that is, $(r_1r_2) = (10)$. Calculations for the remaining two incomplete patterns are analogous.

$$(11) \quad : \quad \begin{cases} \dfrac{\partial q_{11|j_1j_2}}{\partial \alpha_{j_1j_2}} &= -q_{11|j_1j_2}\, q_{0+|j_1j_2} \\ \dfrac{\partial q_{11|j_1j_2}}{\partial \beta_{j_1j_2}} &= -q_{11|j_1j_2}\, q_{+0|j_1j_2} \\ \dfrac{\partial q_{11|j_1j_2}}{\partial \gamma} &= -q_{11|j_1j_2}\, q_{00|j_1j_2} \end{cases}$$

$$(10) \quad : \quad \begin{cases} \dfrac{\partial q_{10|j_1j_2}}{\partial \alpha_{j_1j_2}} &= -q_{10|j_1j_2}\, q_{0+|j_1j_2} \\ \dfrac{\partial q_{10|j_1j_2}}{\partial \beta_{j_1j_2}} &= -q_{10|j_1j_2}\, (1 - q_{+0|j_1j_2}) \\ \dfrac{\partial q_{10|j_1j_2}}{\partial \gamma} &= -q_{10|j_1j_2}\, q_{00|j_1j_2} \end{cases}$$

7.4 Examples

In this section, we will apply the local influence tool to two of the example data sets. First, the continuous $HAMD_{17}$ depression score from the second depression trial is considered, and thus the local influence approach applied to the Diggle-Kenward model in Section 7.3.2 is used. Secondly, we perform a sensitivity analysis of two (binary) questions raised in the Slovenian public opinion survey. Accordingly, we will use the local influence approach as discussed in Section 7.3.3.

7.4.1 Sensitivity Analysis of the Second Depression Trial Data

In Section 6.1.2, we analysed the Second Depression Trial data using the Diggle-Kenward model, for which the missingness assumptions varied from MCAR, to MAR, and MNAR. By considering such different dependencies of the missingness mechanism on the outcomes, we took a first step on the route of sensitivity analysis for the second depression trial.

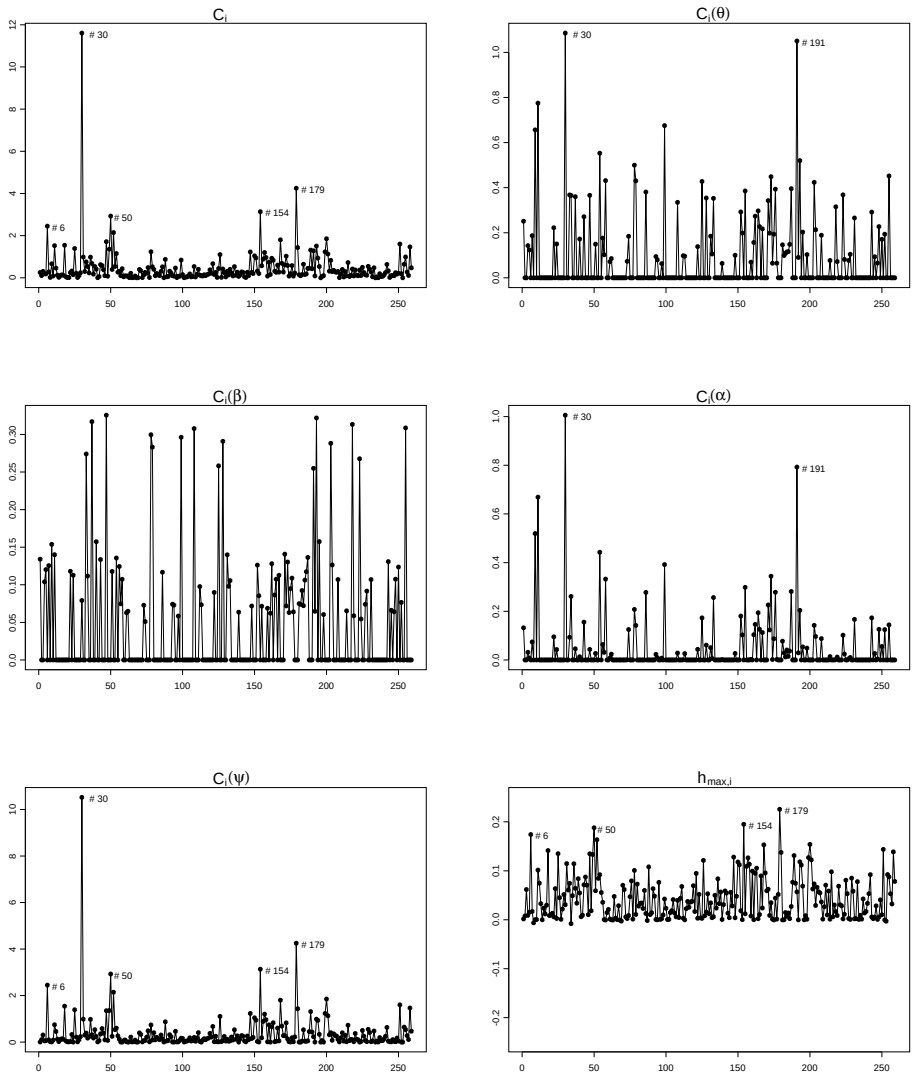


Figure 7.1: *Second depression trial. Index plots of* C_i, $C_i(\boldsymbol{\theta})$, $C_i(\boldsymbol{\alpha})$, $C_i(\boldsymbol{\beta})$, $C_i(\boldsymbol{\psi})$ *and of the components of the direction* $h_{\max,i}$ *of maximal curvature.*

In this section, we will perform a sensitivity analysis at the level of the individuals by switching to local influence approach, as discussed in Section 7.3.2 for the Diggle-Kenward model. We will calculate the following normal curvatures in the direction of the unit vector $\boldsymbol{h}_i$ containing one in the ith position and zero elsewhere: C_i, $C_i(\boldsymbol{\beta})$, $C_i(\boldsymbol{\alpha})$, $C_i(\boldsymbol{\theta})$, and $C_h(\boldsymbol{\psi})$, as well as the normal curvature in the direction of $\boldsymbol{h}_{\max}$ of maximal normal curvature $C_{\max}$.

Figure 7.1 displays overall C_i and influences for subvectors $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\psi}$. In addition, the direction $\boldsymbol{h}_{\max}$, corresponding to maximal local influence, is given. The main emphasis should be put on the relative magnitudes. It is observed that patients $\#6, \#30, \#50, \#154,$ and $\#179$ have larger C_i values compared to other patients, which means they can be considered influential. Among these, patient #30 clearly shows the largest C_i. Virtually the same picture holds for $C_i(\boldsymbol{\psi})$. Turning attention to the influence on the measurement model, we see that for $C_i(\boldsymbol{\beta})$, there are no relatively high peaks, whereas $C_i(\boldsymbol{\alpha})$ again reveals a considerable peak for patient #30, and for patient #191. Note that patient #191 does not have a high peak for the overall C_i. This is due to the fact that the scale for $C_i(\boldsymbol{\alpha})$ is relatively small, comparing to the overall C_i. Nevertheless, these patients can still be considered influential. Finally, the direction of maximum curvature does not really highlight any influential patients, although the four influential completers seem to have the highest values.

In Figure 7.2, the individual profiles of the influential observations are highlighted. Let us now take a closer look at these cases. Patients #30 and #191 dropped out of the study, whereas the others completed the 9-week study. Further, patients #30

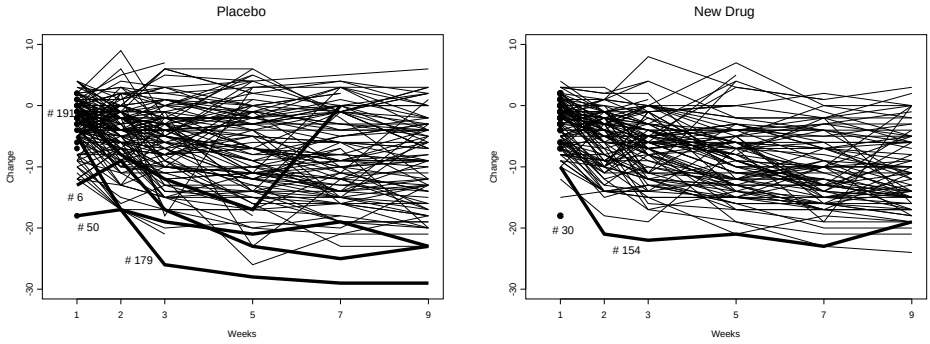


Figure 7.2: *Second depression trial. Individual profiles for both treatment arms, with influential subjects highlighted.*

and #154 belong to the new drug group, while patients #6, #50, #179, and #191 were on placebo.

As we can see from Figure 7.2, patient #30 dropped out after the first post-baseline measurement occasion. This patient achieved the largest change in $HAMD_{17}$ at week 1, its score went down from 24 at baseline to 6 at week 1. The second incomplete influential subject, patient #191, dropped out at the second last measurement occasion, resulting in five observed measurements. It is clear from Figure 7.2 that its $HAMD_{17}$ was decreasing up to the fourth measurement occasion, meaning the patient was improving. However, at the last observed measurement occasion, its value became 0 again, so its $HAMD_{17}$ score was again the same as the one at baseline indicating a marked worsening from the previous visit. Finally, all the other influential patients had a big improvement within the first three weeks, which remained more or less constant afterwards.

Let us provide some clarification for these influential patients. First, the large influence of patient #30 is caused by achieving the biggest reduction in $HAMD_{17}$ score in the new drug group at week 1. Based on the dropout model under the MAR assumption, the dropout probability at week 2 for this subject was very small but the subject dropped out nevertheless. Hence the subject had a large influence on $\boldsymbol{\theta}$ from (7.12). For the same reason, the values of $\boldsymbol{h}_{id_i}$ and $\lambda(y_{id_i}|\boldsymbol{h}_{id_i})$ are large in (7.13), resulting in a large value of $C_i(\boldsymbol{\psi})$. Next, since patients $\#6, \#50, \#154$, and #179 are completers, all their influences are placed on $\boldsymbol{\psi}$. Indeed, it immediately follows from (7.10) and (7.12) that direct influence on $\boldsymbol{\theta}$ only arises from those measurement occasions at which dropout occurs. Their relatively large influence on $\boldsymbol{\psi}$ is due to two facts. First, their profiles are relatively higher in magnitude than others, and hence y_{ij} and $\boldsymbol{h}_{ij}$ in (7.11) are large. Second, since all of these patients are completers, (7.11) contains a maximal number of large terms. Finally, the relatively large influence of $C_i(\boldsymbol{\alpha})$ for patient #191 is due to the large residual $h_{id_i} - \lambda(h_{id_i})$. This can be explained by the fact that the observed change in $HAMD_{17}$ score from baseline at the last observed visit, week 7, was zero, which is distant from the group mean at that time point.

It is interesting to consider an analysis without these influential observations. Therefore, we applied the selection model on three subsets of the data. To get the first subset, patient #30 was removed, since this is overall the most influential one. In the second subset of the data, patients #30 and #191, which seemed to be influencing the measurement model the most, were removed. Finally, all the six influential patients mentioned above were removed, resulting in the third subset. Result of these analyses are shown in Table 7.1. Let us compare the results of the MAR and MNAR analyses.

Table 7.1: *Second depression trial. Parameter estimates (standard errors) assuming ignorability, as well as explicitly modeling the missing data mechanism under MCAR, MAR, and MNAR assumptions, after removing subject* #30*, subjects* #30 *and* #191*, and subjects* #6, #30, #50, #154, #179, *and* #191*.*

Removed Subjects	**#30**				(**#30**, **#191**)				(**#6**, **#30**, **#50**, **#154**, **#179**, **#191**)			
	MAR		MNAR		MAR		MNAR		MAR		MNAR	
Parameters	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Mean Parameters												
β_0 : *intercept*	6.74	(1.46)	6.79	(1.44)	6.70	(1.46)	6.75	(1.46)	5.41	(1.41)	5.47	(1.41)
β_1 : *baseline*	-0.35	(0.07)	-0.35	(0.07)	-0.35	(0.07)	-0.35	(0.07)	-0.29	(0.07)	-0.29	(0.07)
β_2 : *treatment*	-0.47	(0.65)	-0.47	(0.64)	-0.51	(0.64)	-0.51	(0.66)	-0.40	(0.63)	-0.40	(0.64)
β_3 : *time*	-2.41	(0.30)	-2.50	(0.30)	-2.40	(0.29)	-2.49	(0.30)	-2.36	(0.29)	-2.45	(0.30)
β_4 : $time^2$	0.14	(0.03)	0.15	(0.03)	0.14	(0.03)	0.15	(0.03)	0.14	(0.03)	0.14	(0.03)
β_5 : $time \times treatment$	0.60	(0.40)	0.61	(0.40)	0.64	(0.40)	0.64	(0.40)	0.65	(0.39)	0.65	(0.40)
β_6 : $time^2 \times treatment$	-0.03	(0.04)	-0.04	(0.04)	-0.04	(0.04)	-0.04	(0.04)	-0.04	(0.04)	-0.04	(0.04)
Variance Parameters												
σ_1 : *std at time* 1	3.93	(0.17)	3.92	(0.17)	3.95	(0.17)	3.94	(0.17)	3.76	(0.16)	3.75	(0.16)
σ_2 : *std at time* 2	5.23	(0.23)	5.19	(0.23)	5.25	(0.24)	5.22	(0.23)	5.05	(0.23)	5.02	(0.23)
σ_3 : *std at time* 3	5.91	(0.26)	5.88	(0.26)	5.93	(0.27)	5.90	(0.26)	5.71	(0.26)	5.69	(0.26)
σ_4 : *std at time* 4	6.46	(0.29)	6.51	(0.30)	6.41	(0.28)	6.46	(0.29)	6.25	(0.28)	6.31	(0.29)
σ_5 : *std at time* 5	6.19	(0.27)	6.15	(0.27)	6.15	(0.27)	6.11	(0.27)	5.95	(0.27)	5.92	(0.26)
σ_6 : *std at time* 6	6.27	(0.29)	6.24	(0.29)	6.26	(0.29)	6.23	(0.29)	6.09	(0.29)	6.07	(0.29)
common correlation ρ	0.72	(0.02)	0.72	(0.02)	0.72	(0.02)	0.72	(0.02)	0.71	(0.02)	0.70	(0.02)
Missing Data Parameters												
ψ_0	-2.20	(0.14)	-2.43	(0.27)	-2.22	(0.14)	-2.44	(0.27)	-2.23	(0.15)	-2.47	(0.28)
ψ_1	-0.05	(0.02)	0.12	(0.06)	0.05	(0.02)	0.11	(0.05)	0.05	(0.02)	0.11	(0.06)
ψ_2			-0.08	(0.06)			-0.07	(0.06)			-0.08	(0.06)
-2 log-likelihood	7919.8		7918.5		7875.2		7873.9		7701.6		7700.2	
difference at endpoint (p-value)	2.15 (.0197)		2.14 (.0198)		2.07 (.0241)		2.07 (.0237)		2.40 (.0082)		2.39 (.0083)	

The largest C_i is observed for patient #30. Its relatively large influence is caused by this patient's big improvement, that is, big drop in $HAMD_{17}$ score, just before dropout. By removing patient #30, the estimate of ψ_1 changed from 0.11 to 0.12. Formulating the dropout model in terms of the increment $y_{ij} - y_{i,j-1}$ and the previous measurement $y_{i,j-1}$, the coefficient for the increments does not change, while the coefficient for $y_{i,j-1}$ increased from 0.03 to 0.04. Since the coefficient for the current measurement y_{ij} does not change by removing patient #30, there is not much influence on the likelihood ratio test for MAR against MNAR: $G^2 = 1.3$ compared to $G^2 = 1.5$. The parameter for the main treatment effect decreases, as well as the difference between the new drug and placebo at week 9, resulting in a slightly increased p value. This holds for both MAR and MNAR. This can be explained by the patient's big improvement before dropout, and membership to the new drug group.

By removing patients #30 and #191, the parameter for the *treatment*-by-*time* interaction under MAR changed from 0.59 to 0.64. Also the estimate for the interaction between *treatment* and $time^2$ changed slightly from -0.03 to -0.04. The change in these estimates is due to the unusual individual profile for patient #191. As observed before, the $HAMD_{17}$ score decreased during the first four post-baseline visits and suddenly went back to the level at baseline at the penultimate visit. Figure 7.3 shows the fitted mean profiles of the change in $HAMD_{17}$, both for placebo and the new drug group, for all subjects and for the subset with patients #30 and #191 removed. We used the mean baseline value to calculate the mean values. For the new drug group, the coefficients for *time* and $time^2$ remain the same for all subjects and for the subset, the two mean profiles are parallel and the difference between the fitted mean profile of all subjects versus the subset of removing patients #30 and #191 is relatively small. In the placebo group, the mean profile becomes steeper by removing the two patients. The estimate for treatment effect dropped from -0.35 to -0.51, mainly due to the big improvement on patient #30. Overall, the difference at endpoint decreased from 2.19 to 2.07, and the p-value for this difference increased from 0.0176 to 0.0241. A similar pattern is seen for the MNAR analysis.

It is noted that there is a small impact on the likelihood ratio test for MAR against MNAR by removing patients #30 and #191. The value of G^2 changed from 1.5 to 2.07. This is partially due to a weaker incremental component in the dropout model. The coefficient for the increments $y_{i,j} - y_{i,j-1}$ changes from -0.08 to -0.07, and the coefficient of the previous measurement increases from 0.03 to 0.04.

Finally, we perform the same analyses on the third subset with patients #6, #30,#50,#154,#179 and #191 removed. Again, we observe for the MAR analysis that the parameter for the interaction between time and treatment changed to

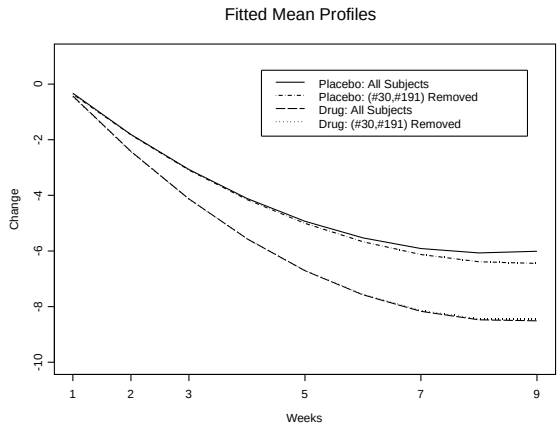


Figure 7.3: *Second depression trial. Fitted mean profiles, both for placebo group and treatment group, for all subjects and for the subset resulting from removing subjects* #30 *and* #191*. Mean baseline value is used to calculate the means.*

0.65, mainly due to the unusual profile of patient #191. There is an increase of 0.31 in the difference between the new drug and placebo at the endpoint. Also, the p-value for this difference decreases from 0.0176 to 0.0082. A similar pattern is found for the MNAR analysis. This change in difference at the endpoint could be due to the fact that the profiles for patient #6, #50 and #179 are relatively low in the placebo group. The estimates of the parameters in the dropout model do not change much, and the deviance for MAR against MNAR is nearly the same.

To conclude, the following observations can be made. Most of the influential subjects were completers and they have a considerable impact on the dropout and/or measurement model parameters due to their unusual individual profiles. With respect to the primary analysis, the mean change in $HAMD_{17}$ is significantly different between the new drug group and the placebo group under the different specifications of the missingness mechanism. The result still holds after removing the influential subjects. Further, if the likelihood ratio test of MAR against MNAR would follow a standard χ^2_1-distribution, we would fail to reject the null hypothesis, which leads us to the MAR assumption. However, the test of MAR against MNAR is non-standard and it cannot be used as such (Rotnitzky *et al.*, 2000; Jansen *et al.*, 2006b).

Using the local influence method, the most influential subject was patient #30. Belonging to the treatment group, this patient had the unusual profile of a very big

improvement, but still dropped out after the first visit. To better understand the influence of this patient, the demographic information for this patient was investigated. This patient was in his/her first Major Depressive Disorder (MDD) episode, when s/he was enrolled. The patient dropped out of the study after week 1, based on his/her own decision and claimed that the symptoms of depression were caused by high carbon monoxide levels in his/her house. Given this information, it is unlikely such a patient provides meaningful information regarding the risks and benefits of the investigational treatment, and it is probably best to consider the merits of the drug after excluding this observation.

And while excluding this patient had little effect on interpretations of the treatment effect in this well-powered confirmatory clinical trial, which is very useful information in and of itself, it is also useful to consider the benefits of sensitivity analyses in a proof of concept setting where the sample sizes are much smaller. For example, if there had been only 40 subjects per arm, excluding this one subject would have had a much bigger impact. Knowing how strongly results depend on one or a few subjects, or on specific assumptions, could potentially improve decisions on whether to continue or discontinue development of an intervention.

7.4.2 Sensitivity Analysis of the Slovenian Public Opinion Survey Data

Both local influence approaches displayed in Section 7.3.3 are applied to the Slovenian public opinion survey data, and juxtaposed with the global influence analysis (Section 7.2), the interval-of-ignorance based sensitivity analysis of Molenberghs, Kenward and Goetghebeur (2001a), and the computation of the so-called MNAR bodyguard to the model considered (Section 6.3).

Interval of Ignorance

It is useful to distinguish between two types of statistical uncertainty. The first, statistical imprecision, is due to finite sampling. The Slovenian public opinion survey included not all Slovenians but only 2074 respondents. However, even if all would have been included, there would have been residual uncertainty because some fail to report at least one answer. This second source of uncertainty, due to incompleteness, is called statistical ignorance. The 16 complete-cell probabilities are as in Table 6.2(a), thus producing 15 complete data degrees of freedom. Similarly, the 9 observed cells can be represented as in Table 6.2(b), which is directly comparable with the observed data structure. Molenberghs, Kenward and Goetghebeur (2001a), Kenward, Goetghebeur

and Molenberghs (2001) and Vansteelandt *et al.* (2006) combined both concepts into *statistical uncertainty* as following.

A sample with underlying theoretical distribution as shown in Table 6.2(b) will produce empirical proportions representing the π's with error. This results in imprecision, which is usually captured by way of such quantities as standard errors and confidence intervals. This first source of imprecision disappears as the sample size tends to infinity and the estimators are consistent. What remains is ignorance regarding the redistribution of all but the first four π's over the missing outcomes value. This leaves ignorance regarding any probability in which at least one of the first or second indices is equal to 0, and hence regarding any derived parameter of scientific interest. For such a parameter, θ say, a region of possible values, which is consistent with Table 6.2(b), is called a *region of ignorance*. Analogously an observed incomplete table leaves ignorance regarding the would-be observed complete table, which in turn leaves imprecision regarding the true complete probabilities. The region of estimators for θ consistent with the observed data provides an estimated region of ignorance. The $(1-\alpha)100\%$ *region of uncertainty* is a larger region in the spirit of a confidence region, designed to capture the combined effects of imprecision and ignorance. Various ways for constructing regions of ignorance and regions of uncertainty are conceivable. For a single parameter, the regions obviously become intervals.

In standard statistical practice, ignorance is hidden in the consideration of a single identified model. As seen in Section 6.1.5, BRD6–BRD9 saturated the observed degrees of freedom. These models cannot be distinguished in terms of their fit to observed data alone. However, they can produce substantially different inferences, as exemplified in Table 6.5. To obtain a measure of ignorance, Molenberghs, Kenward and Goetghebeur (2001a), Kenward, Goetghebeur and Molenberghs (2001) and Vansteelandt *et al.* (2006) consider models that would be identified if the data were complete, and fit them to the observed, incomplete data, thereby producing a range of estimates rather than a point estimate. These authors use the non-identifiability to delineate the range of inferences consistent with the observed data, that is, to capture ignorance. Maximization of the likelihood function is a natural approach. To manage overspecification of the likelihood, they consider a minimal set of parameters, called *sensitivity parameters*, conditional upon which the others, the *estimable parameters* are identified. Each value of the sensitivity parameter will produce an estimate for the estimable parameter, of which the union produces the estimated region of ignorance.

For a bivariate binary outcome with non-monotone missingness any model within the BRD family with 9 or more parameters is non-identifiable. For the Slovenian public opinion survey, Molenberghs, Kenward and Goetghebeur (2001a) consider three

Table 7.2: *The Slovenian public opinion survey. Intervals of ignorance and intervals of uncertainty for the proportion θ (confidence interval) attending the plebiscite following from fitting.*

			$\widehat{\theta}$	
Model	d.f.	loglik	II	IU
Model 10	9	-2431.06	[0.762;0.893]	[0.744;0.907]
Model 11	9	-2431.06	[0.766;0.883]	[0.715;0.920]
Model 12	10	-2431.06	[0.694;0.905]	

such overspecified models, among which two (model 10 and 11) contain one sensitivity parameter, whereas model 12 includes two. Model 10 is defined as $(\alpha_{j_2}, \beta_{j_1 j_2})$ with

$$\beta_{j_1 j_2} = \beta_0 + \beta_{j_1} + \beta_{j_2}, \tag{7.18}$$

thereby considering an additive decomposition for missingness on the independence question, while Model 11 assumes $(\alpha_{j_1 j_2}, \beta_{j_1})$ and uses an additive decomposition of the missingness parameter on the attendance question, that is,

$$\alpha_{j_1 j_2} = \alpha_0 + \alpha_{j_1} + \alpha_{j_2}. \tag{7.19}$$

Finally, Model 12 is defined as $(\alpha_{j_1 j_2}, \beta_{j_1 j_2})$, a combination of both (7.18) and (7.19).

Molenberghs, Kenward and Goetghebeur (2001a) provide a table with the estimated intervals of ignorance and intervals of uncertainty. Recall there appeared to be a small computational error, and therefore the corrected results are shown in Table 7.2. Further, a graphical representation of the YES votes is given in Figure 7.4.

Model 10 shows an interval of ignorance which is very close to $[0.741, 0.892]$, the range produced by the models BRD1–BRD9, while Model 11 is somewhat sharper and just fails to cover the plebiscite value. However, it should be noted that the corresponding intervals of uncertainty contain the true value.

Interestingly, Model 12 virtually coincides with the non-parametric range, that is, the pessimistic-optimistic interval, even though it does not saturate the complete data degrees of freedom. To do so, not 2 but in fact 7 sensitivity parameters would have to be included. Thus, it appears that a relatively simple sensitivity analysis is sufficient to increase the insight in the information provided by the incomplete data about the proportion of valid YES votes.

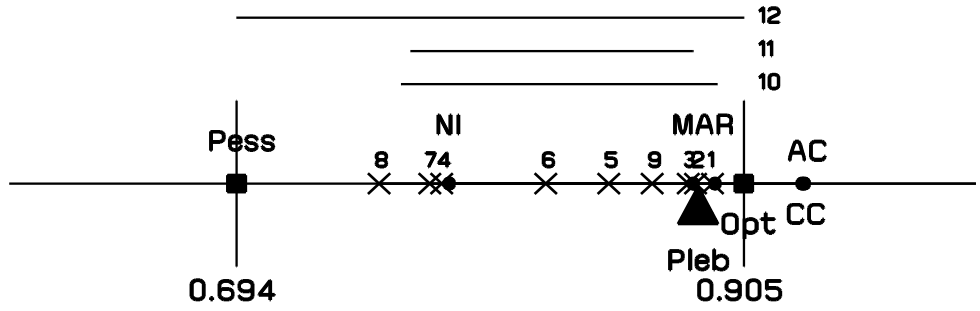


Figure 7.4: *The Slovenian public opinion survey. Relative position for the estimates of "proportion of YES votes", based on the models considered in Rubin, Stern and Vehovar (1995) and on the BRD Models. The vertical lines indicate the nonparametric pessimistic-optimistic bounds. (Pess: pessimistic boundary; Opt: optimistic boundary; MAR: Rubin* et al *'s MAR model; NI: Rubin* et al *'s MNAR model; AC: available cases; CC: complete cases; Pleb: plebiscite outcome. Numbers refer to the BRD models. Intervals of ignorance (Models 10–12) are represented by horizontal bars.)*

An MAR Bodyguard for an MNAR Model

In Section 6.3 and in Molenberghs *et al.* (2007), we showed that, strictly speaking, the correctness of the alternative model can only be verified in as far as it fits the *observed* data. Thus, evidence for or against MNAR can only be provided within a particular, predefined parametric family, the plausibility of which cannot be verified in empirical terms alone. This implies that an overall (omnibus) assessment of MAR *versus* MNAR is not possible, since every MNAR model can be doubled up with a uniquely defined MAR counterpart, producing exactly the same fit as the original MNAR model, in the sense that it produces exactly the same predictions to the observed data (e.g., fitted counts in an incomplete contingency table) as the original MNAR model, and depending on exactly the same parameter vector. While this so-called MAR bodyguard generally does not belong to a conventional parametric family, its existence has important ramifications. We have illustrated the use of the MAR bodyguard by means of Section 6.3.3.

Global Influence

Performing a global influence analysis on data with categorical outcomes is less time consuming than on data with continuous outcomes, since the data can then be organized into cells, as in Table 6.3. Thus, instead of removing subjects on a one by

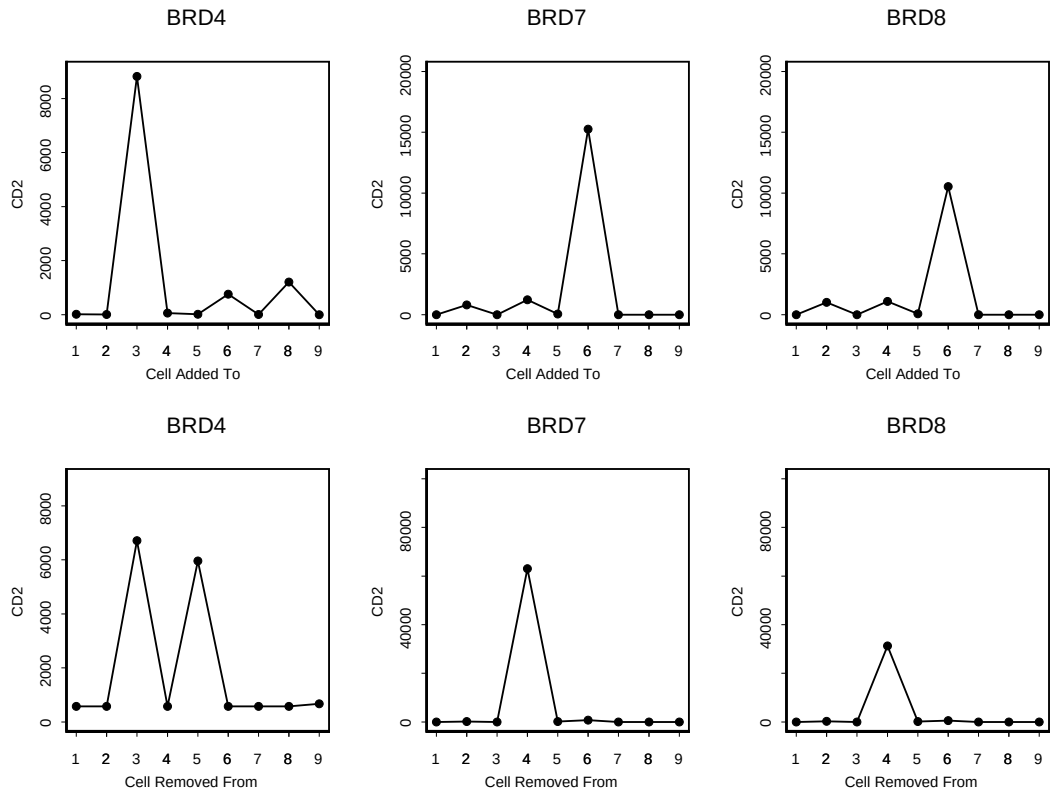


Figure 7.5: *Slovenian public opinion survey. Global influence analysis for BRD4, BRD7 and BRD8. Cook's distance measure, CD_{2i}, is evaluated when an observation is added to a specific cell (first row) and when an observation is deleted from a specific cell (second row).*

one basis, we only need to remove one subject per cell and per covariate level, in case covariates are considered too.

Figure 7.5 shows a selection of the results for the global influence analysis on the Slovenian public pinion survey data. Results of the Cook's distance measure CD_{2i} for BRD4, 7, and 8 are presented. Observe that, for BRD4, adding a single observation to cell #3 has a large influence on the parameters, as well as deletion from either cells #3 or #5. Cell #3 represents subjects with a NO on the attendance question and a YES on the independence question. An addition or removal of one such respondent can largely affect the parameters of BRD4. Similarly, exclusion of a single respondent with a YES on the attendance question but a missing response

on the independence question (cell #5), also influences BRD4's model parameters, though to a lesser extent.

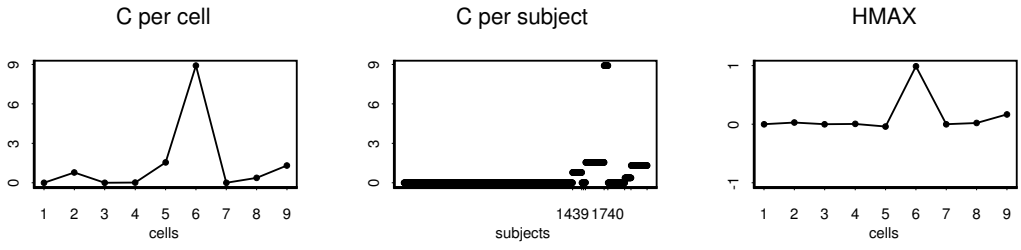
For models BRD7–8, an additional observation in cell #6 or a deletion from cell #4 leads to significant influence on these models' parameters. Thus, adding a subject with a NO for attendance and a missing independence response, or excluding a respondent with NO on both questions, yields changes in the model parameters of BRD7–8. These findings hint on the influential nature of subjects with a NO on the attendance question, which is likely related with this group's sparseness.

For all other models, Cook's distance measure CD_{2i} was approximately zero for all cells, indicating no substantial influence when adding or removing a single case from a particular cell.

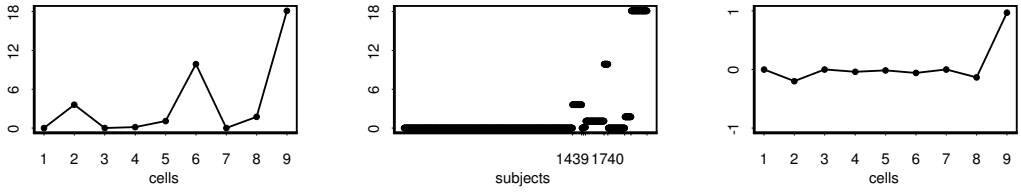
Local Influence by Perturbing Parameters: One BRD Model vs. Another

Turning to the local influence tool for the Slovenian public opinion survey data, we will first consider perturbations of a given BRD model in the direction of another BRD model in which the null model is nested, implying that we consider perturbations along the edges of Figure 6.1. We will consider local influence measures on both the likelihood displacement (7.6) and the predicted cell counts (7.7) for different model pairs. Although 12 model nestings are possible (Figure 6.1), we focus on the model pairs BRD1 *vs.* BRD4, BRD3 *vs.* BRD7, and BRD4 *vs.* BRD7. The rationale for these choices, in addition to conciseness, is that in these 3 model pairs substantial influence was seen when considering local influence on the likelihood displacement. In addition, for the local influence on the predicted cell counts, discussed in the next section, these three model pairs are indicative for the various features that were seen across all 12 comparisons.

Figure 7.6 shows, for the 3 comparisons considered, the influence measures C_i, plotted against the ith observed cell, as well as against each subject within that cell, and $\boldsymbol{h}_{\max}$ against the ith observed cell. A peak in the graph at a particular cell indicates that the corresponding cell drives the data towards the more complex, rather than the simpler model. For the comparison of BRD1 *vs.* BRD4, a peak is observed at cell #6, for both C_i and $\boldsymbol{h}_{\max}$, implying that respondents in this cell drive the data more towards BRD4 $(\alpha_{\cdot\cdot}, \beta_{\cdot j_2})$ rather than BRD1 $(\alpha_{\cdot\cdot}, \beta_{\cdot\cdot})$. That is, subjects with a NO on the attendance question and a missing value on the independence question are influential when perturbing the model such that missingness in the independence question depends on the corresponding unobserved answer (BRD4) rather than being constant (BRD1).



(a) *BRD1 vs BRD4*



(b) *BRD3 vs. BRD7*

(c) *BRD4 vs. BRD7*

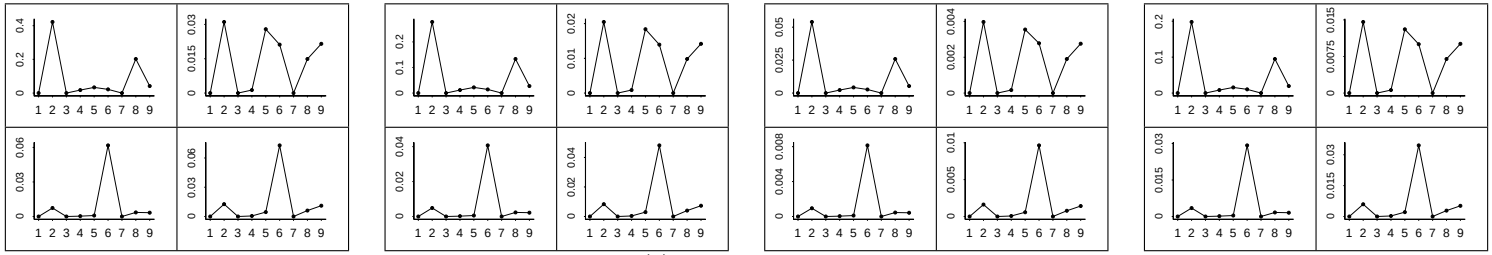
Figure 7.6: *Slovenian public opinion survey. Local influence analysis on parameters for model pairs (a) BRD1 vs. BRD4, (b) BRD3 vs. BRD7, and (c) BRD4 vs. BRD7. First column shows the local influence measure C_i at the ith observed cell; the second column shows the same measure but plotted for each of the subjects within the ith observed cell; and, the third column shows $\boldsymbol{h}_{max}$ for the ith observed cell.*

For BRD3 *vs.* BRD7, a peak is observed at cell #9, subjects with a missing response on both questions, implying that such subjects drive the data in the direction of BRD7 $(\alpha_{.j_2}, \beta_{.j_2})$ rather than BRD3 $(\alpha_{.j_2}, \beta_{..})$. That is, missingness in the independence question is driven to depend on the corresponding unobserved answer by subjects with missing responses on both questions, and, also slightly by those with a NO on attendance and a missing value on independence (cell #6). Finally, it is primarily subjects with missing responses on both questions (cell #9) that seem to push the data towards BRD7 $(\alpha_{.j_2}, \beta_{.j_2})$ from BRD4 $(\alpha_{..}, \beta_{.j_2})$. These subjects, along with those that have a YES on attendance and a missing value on independence (cell

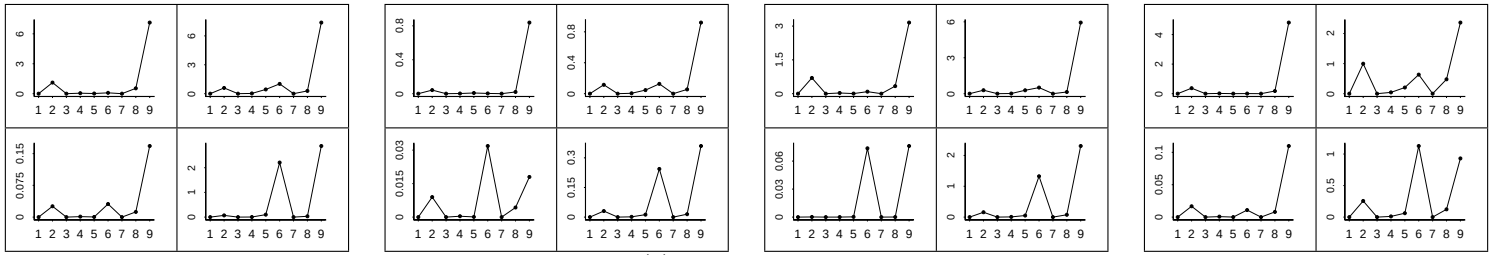
#5), make the missingness in the attendance question depend on the response of the independence question.

We now turn to the results of the local influence analysis on the fitted cell counts. Graphs of the local influence measure (7.7) on the predicted cell counts are presented in Figure 7.7, with the graphs for the 16 predicted cell counts arranged in their respective positions as in Table 6.2(a). From the first panel, for model pair BRD1 *vs.* BRD4, we observe that the influence graphs show similar shapes, albeit with differing magnitudes, for a particular cell (j_1, j_2), across the four missingness patterns. For instance, the influence curves for $Z_{r_1 r_2,11}$ (upper left corners) for $(r_1, r_2) = (1,1), (1,0), (0,1), (0,0)$ have more or less identical shapes. Occurrences of peaks at particular cells are thus common across the missingness patterns, yielding more or less a clear result for each cell (j_1, j_2). For $(j_1, j_2) = (1,1)$, it is cell #2 that shows influence, and also slightly cell #8. Respondents with either a YES or a missing value on attendance and a NO on independence therefore drive the predicted cell count $Z_{r_1 r_2,11}$ towards a model in which the missingness in the independence question depends on its value (BRD4). For $(j_1, j_2) = (1,2)$, cells #2 and #5, as well as #6 and #9, stand out. Cells #2 and #5 denote, respectively, respondents having YES on attendance/NO on independence, and YES on attendance/missing value on independence, and these respondents make the predicted cell count $Z_{r_1 r_2,11}$ seem to have come more from BRD4 rather than from BRD1. For $(j_1, j_2) = (2,1)$ and $(j_1, j_2) = (2,2)$, similar curves are obtained across the four missingness patterns, with a clear peak at cell #6, implying that the "NO-on-attendance/missingness-on-independence" responses perturb predicted cell counts $Z_{r_1 r_2,21}$ and $Z_{r_1 r_2,22}$ in the direction of a model in which the missingness in the independence question is dependent on its value, rather than on one in which missingness in the independence question is constant.

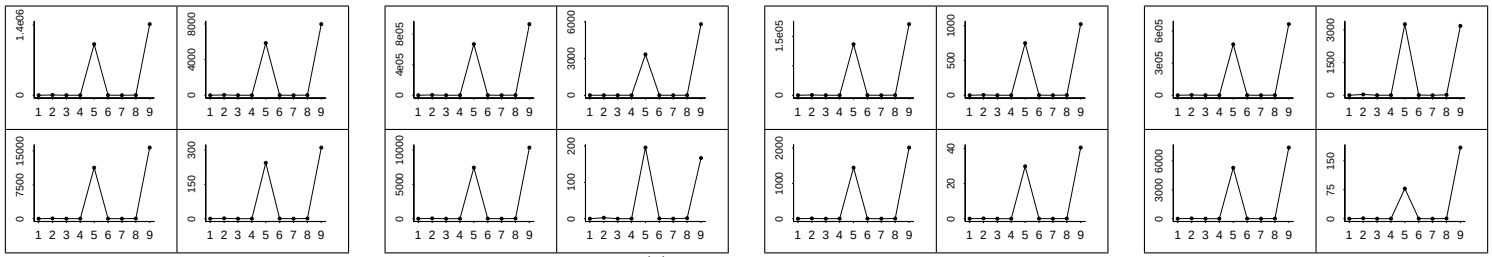
The resulting patterns for the comparison of BRD3 against BRD7 (Figure 7.7b) differs from what was observed for BRD1 *vs.* BRD4. Whereas for the latter, influence curves for a particular cell (j_1, j_2) remained the same across the missingness patterns, for BRD3 *vs.* BRD7, variations now arise across these missingness patterns, leading to a less clear-cut overall picture. For $(j_1, j_2) = (1,1)$ and $(j_1, j_2) = (1,2)$, that is, top row of the 4 sets of tables, although relative peaks are observed at the same positions across the 4 sets of tables, the degree of the peak varies across the missingness patterns, causing some to appear more like a peak and some less so. This is further complicated by what can be observed for $(j_1, j_2) = (2,1)$ and $(j_1, j_2) = (2,2)$, that is, the bottom row of the tables, for which curve shapes vary across the missingness patterns. We proceed to look at the results for $(j_1, j_2) = (1,1)$ and $(j_1, j_2) = (1,2)$.



(a) *BRD1 vs. BRD4*



(b) *BRD3 vs. BRD7*



(c) *BRD4 vs. BRD7*

Figure 7.7: *Slovenian public opinion survey. Local influence analysis on the predicted cell counts for model pairs (a) BRD1 vs. BRD4, (b) BRD3 vs. BRD7, and (c) BRD4 vs. BRD7. Plots show C_i values for each of the 16 predicted cell counts (in their respective positions as in Table 6.2) against the 9 observed cells (as labelled in Table 6.3).*

Across the 4 missingness patterns, the predicted cell counts $Z_{r_1r_2,11}$ and $Z_{r_1r_2,12}$ are primarily influenced by subjects with both responses missing, and slightly by those having a YES on attendance/NO on independence. For cell $(j_1, j_2) = (2, 1)$, similar graphs are obtained for $(r_1, r_2) = (1, 1)$ and $(r_1, r_2) = (0, 0)$, that is, the completers and double non-responders, respectively, with a peak at cell #9. It is therefore subjects with both responses missing that influence cell counts $Z_{11,21}$ and $Z_{00,21}$, in the direction of a model in which missingness in the independence question depends on its value. For the other two missingness patterns, $(r_1, r_2) = (1, 0)$ and $(r_1, r_2) = (0, 1)$, referring to subjects with a single nonresponse, peaks occur at cells #6 and #9. Thus, subjects with a NO on attendance/missingness of independence and those with both responses missing have an influence on predicted cell counts $Z_{10,21}$ and $Z_{01,21}$. These same subjects also influence the predicted cell counts $Z_{r_1r_2,22}$, since we observe similarly shaped influence curves across the missingness patterns for cell position $(j_1, j_2) = (2, 2)$, with peaks either at cell #9 or cell #6.

Whereas the comparison of BRD3 *vs.* BRD7 presents the most variable influence graphs, BRD4 *vs.* BRD7 shows the most consistent ones. All 16 influence curves exhibit a single shape, although of varying magnitudes, implying that influence on any predicted cell count is coming from a common source, regardless of the missingness pattern. Here, we see a clear peak at cells #9 and #5, similar to what was observed for this model pair when considering influence on the likelihood displacement. Subjects with missing responses on both questions and those with YES on attendance/missingness on independence, have an influence that drives any predicted cell count towards a model where the missingness in the attendance question depends on the response of the independence question.

Local Influence by Perturbing Cell Probabilities

Next, we apply the second local influence approach discussed in Section 7.3.3 to the Slovenian public opinion survey data, studying the effect of infinitesimally small perturbations in the cell probabilities. We first derive influence measures on the likelihood displacement; these are graphed in Figure 7.8. For most BRD models, it seems small perturbations in the probabilities of cells #3 and/or #4 has a large influence. That is, if we slightly alter the probabilities with which the "NO-on-attendance/YES-on-independence" or the NO/NO respondents occur, we can expect substantial likelihood displacement. Also notable is the influence of changes in cell #6 for BRD8, implying that under this model, changing the probability of the NO/missingness category slightly causes displacement in the likelihood. These observations suggest that

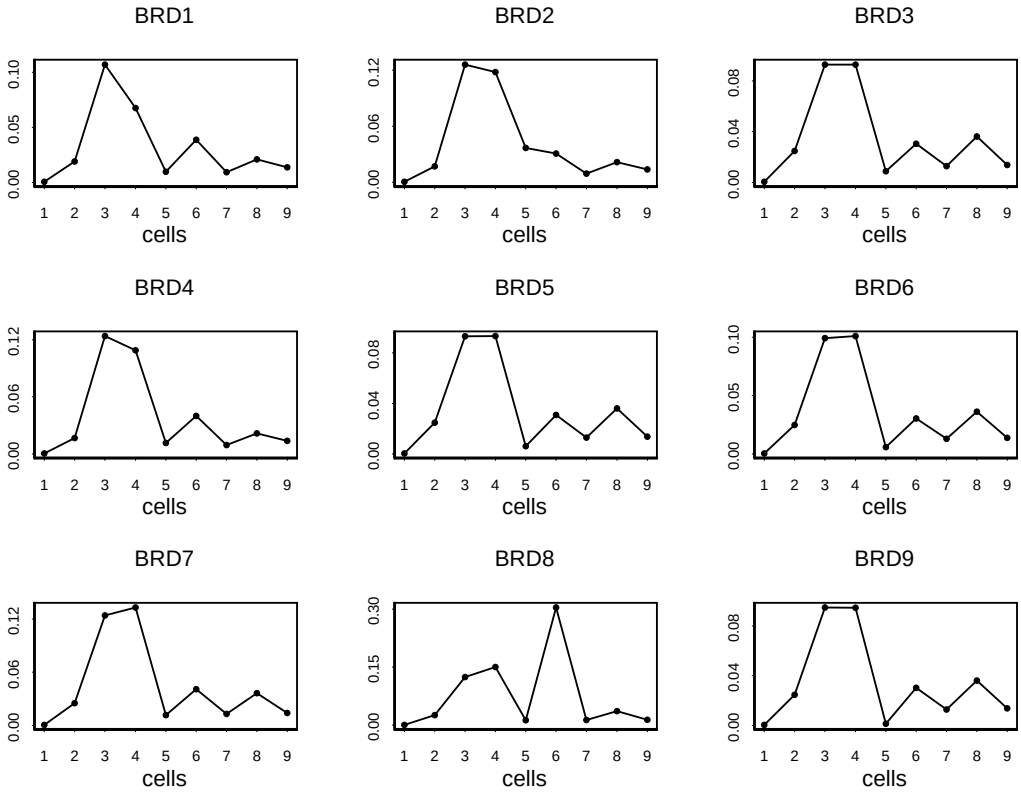


Figure 7.8: *Slovenian public opinion survey. Local influence analysis on the log-likelihood for the 9 BRD models. Plots show C_i values against the 9 observed cells for each BRD model.*

the most influential cells for virtually all BRD models are the completers answering NO on attendance, likely attributable to the small counts in these cells, while for BRD8, it is those subjects answering NO on attendance and unobserved response on independence that are influential.

Table 7.3 provides a summary of the results of the local influence analysis on the predicted cell counts when perturbing a particular cell probability. No particular influence can be seen for any BRD model when perturbing probabilities of cells #1 and #2, as might be expected since the observed cell counts in these cells are large, and thus infinitesimal changes in their respective cell probabilities may not have a large impact. We can also see that small perturbations in cell probability 3 seem to affect only the predicted cell counts in the top row ($j_1 = 1$, YES on attendance) under

Table 7.3: *Slovenian Public Opinion Survey. Local influence analysis on cell counts when perturbing each of the 9 observed cell probabilities. Entries in boxes denote the BRD model number for which influence is largest when the particular cell probability is perturbed.*

Adding ω to cell	Z_{11,j_1j_2}	Z_{10,j_1j_2}	Z_{01,j_1j_2}	Z_{00,j_1j_2}
1				
2				3
3	5,6 / 5,6 / 1–3,9	5,6 / 5,6 / 2,3,9	5,6 / 5,6 / 1,2,3,6	5 / 5 / 1,2,3,6 / 6
4	4,6,7,9 / 4–7 / 4–7,9 / 2–9	4–7,9 / 4–8 / 4–7 / 2–9	4–7 / 4–8 / 4–7,9 / 2–8	4–7,9 / 3–9 / 4–7,9 / 2–5,7,9
5	2	1–3 / 1,2		
6	8 / 8 / 8 / 1,8	8 / 8 / 1,8 / 8	8 / 8	8 / 8 / 8
7	3,9	9	1–3,9	9
8	1–3,9	9 / 3,9	1–3,9 / 1,9	9
9	1,2			1–3 / 1,2 / 1 / 1

BRD5 and/or BRD6, while such changes impact the cell position $(j_1, j_2) = (2, 1)$ (NO on attendance/YES on independence) under BRD's 1,2,3, 6 and/or 9.

Perhaps the most striking result that can be observed from Table 7.3 is that for perturbations in cell probability 4 (NO/NO respondents), which yields influence on *all* 16 predicted cell counts in most of the higher-numbered BRD models 4 to 9. Also of particular interest are the results for perturbations in cell #6, indicating that it is primarily under BRD8 where a large influence is observed in the most of the predicted cell counts. Finally, we note that changes in the probability of the doubly missing category (cell #9) affects only the predicted cell counts of this missingness pattern and only under BRD's 1,2 and/or 3.

Comparison of Sensitivity Analyses

As mentioned before, a first family of sensitivity analyses is based on considering a variety of models. First, simple analyses have been augmented with a non-parametric interval in Section 6.1.5, providing absolute bounds for the proportion of people in favor of independence, which is expressed through at the same time possessing and expressing an opinion in favor of independence. Further, a nonparametric interval was supplemented with a finite collection of identifiable models from the BRD family in Section 6.1.5. In addition, for each of these models, the corresponding MAR bodyguards were calculated in Section 6.3.3; these are models with the same fit to the observed data than their corresponding original models, but with missingness of the MAR type. Additionally, in this section, infinite collections resulting from overspecified models are considered, providing intervals of ignorance and intervals of uncertainty. Whereas the nonparametric range is [0.694,0.905], with the parametric ranges subsets thereof, the MAR models center around 0.89, close to the actual plebiscite values.

A second family of sensitivity analyses studies influence of observations on the model's conclusions, expressed through either parameters or cell counts. In this section, we considered both global influence, that is, case deletion, and local influence, based on infinitesimal perturbation. It was found that perturbing some, but not all, small counts can have an extremely large effect on the conclusions, often through partially observed or unobserved cells. Such influences can strongly affect conclusions about estimands such as the one considered here. Indeed, the proportion of people attending the plebiscite and at the same time being in favor of independence is made up of adding up the (1,1) cell across all missingness patterns, and hence depends on how a model distributes partially observed counts over the cells.

7.5 Concluding Remarks

In many longitudinal settings which are prone to missingness, the assumption of an MAR missingness mechanism is a reasonable starting point. However, MNAR can never be completely ruled out. Understanding how the results depend on the specification of the missingness mechanism will be very helpful in understanding the data. Since the models for non-random dropout rest on strong and untestable assumptions, the optimal place for the MNAR analyses is within a sensitivity analysis framework.

In this chapter, we have given a definition of sensitivity analysis and presented a variety of sensitivity analyses tools. Further, we performed a sensitivity analysis both on the second depression trial data and on the Slovenian public opinion survey data.

Recall that sensitivity analyses can be based on either considering a variety of models, or on studying influences of observations on the model's conclusions. The latter can be achieved through the local influence tool, which is used to depict anomalous subjects that lead to a seemingly MNAR mechanism. Although the original idea behind the use of local influence methods was to detect subjects that drop out non-randomly, several authors (Verbeke *et al.*, 2001b; Jansen *et al.*, 2006b) have shown that the influential subjects often are influential for other than missingness-related features. Jansen *et al.* (2006b) assert that "a subject that drives the conclusion towards MNAR may be doing so, not only because its true data generating mechanism is of an MNAR type, but also for a wide variety of other reasons, such as an unusual mean profile or autocorrelation structure". In this chapter, we have given a detailed discussion of this approach, adopting it both on the Diggle-Kenward model (Section 7.3.2) and the BRD model family (Section 7.3.3).

A careful study of influential subjects, combined with knowing how the results of primary interest change by removing such subjects and by altering assumptions regarding missing data, can lead to a better understanding of the nature of clinical trial data. Sensitivity analyses can also help develop an appropriate level of confidence in the originally proposed primary analysis and help develop alternative analyses in which more confidence can be placed by the researchers. Thus, when an analysis with and without influential subjects yields essentially the same conclusion about, for example, the trial's treatment effect, one will place more confidence in the conclusions than when no sensitivity analysis had been conducted. However, when there is a difference between both, careful scrutiny ought to follow and one may decide to remove all or some of the influential subjects.

In conclusion, we believe it is important to conduct a sensitivity analysis and in particular that such an analysis that combines insight from considering families of

models on the one hand and from studying influence is able to paint a relatively complete picture. This allows one to put a perspective on the conclusions that can confidently be reached about an estimand based on an incomplete set of data, something that considering a single model arguably never can.

8

A Latent-Class Mixture Model for Incomplete Longitudinal Gaussian Data

In Section 3.1.2 the three main modeling frameworks - selection, pattern-mixture, and shared-parameter - were introduced. Note that so far the focus has been on selection models. However, it is possible to formulate models that combine aspects of the three families: indeed Molenberghs *et al.* (1998b) place all three in one overall framework. In this chapter, we propose a so-called latent-class mixture model, an example of such a combination, using latent classes.

Besides the fact that this model provides a flexible modeling tool in its own right, the method's use lies predominantly within the sensitivity analysis context. Such a sensitivity analysis is clearly useful when the more elaborate model modifies the results from the simpler alternative. However, even when it confirms earlier results, it will typically increase confidence in the conclusions reached.

The latent-class mixture model is introduced in Section 8.1. The corresponding likelihood function and associated methods of estimation are discussed in Section 8.2. In Section 8.3 we explore how the method can be used as a device for classifying subjects into latent groups. Using simulations, some insight into its performance is

provided in Section 8.4. Finally, in Section 8.5, the methodology is illustrated using the first depression trial data.

8.1 Latent-Class Mixture Models

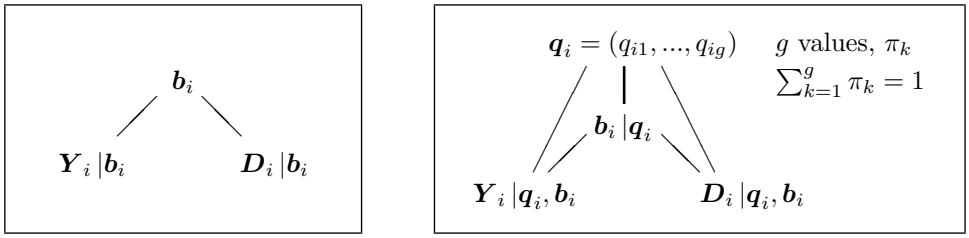
We propose a *latent-class mixture model*, bringing together features of the selection, pattern-mixture, and shared-parameter model frameworks. Precisely, information from the location and evolution of the response profiles, a selection model concept, and from the dropout patterns, a pattern-mixture idea, is used simultaneously to define latent groups and variables, a shared-parameter feature. This approach has a number of appealing features. First, it allows for using the information in a more symmetric and therefore more elegant way. Second, apart from providing a more flexible modeling tool, the new framework is ideally suited to be used as a sensitivity analysis instrument. Third, a strong added advantage over existing methods is that we now will be able to classify subjects into latent groups. While this has to be done with due caution, it can enhance substantive knowledge and generate hypotheses for further research. Fourth, while computational burden evidently increases, fitting the proposed method is remarkably stable and falls within acceptable time limits for applications of the type considered here and for simulations reported.

As before, let the random variable Y_{ij} denote the response of interest, for the ith individual, with measurements planned at times t_{ij}, $i = 1, \ldots, N$, and $j = 1, \ldots, n$. We group the outcomes into the vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})'$. We also define the dropout indicator to have its usual meaning (Section 3.1.1). Recall the shared-parameter model factorization (3.3):

$$f(\boldsymbol{y}_i, d_i|\boldsymbol{b}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) f(d_i|\boldsymbol{b}_i, \boldsymbol{\psi}),$$

in which we suppress explicit reference to the covariates. The above factorization presumes the existence of a random-effects vector $\boldsymbol{b}_i$, conditional upon which the measurement and dropout processes are independent. This particular shared-parameter model can be represented as in Figure 8.1(a).

We propose an extension to this model that captures possible heterogeneity between the subjects, which is not measured through a covariate, but rather through a latent variable. This extended model is called a *latent-class mixture model*, and a representation of it is shown in Figure 8.1(b). Next to one or more so-called shared parameters, $\boldsymbol{b}_i$, the model contains a latent variable, $\boldsymbol{Q}_i$, dividing the population in g subgroups. This latent variable is a vector of group indicators $\boldsymbol{Q}_i = (Q_{i1}, \ldots, Q_{ig})$,



(a) (b)

Figure 8.1: *Representation of (a) shared-parameter models and (b) their extension to latent-class mixture models.*

defined as $Q_{ik} = 1$, if subject i belongs to group k, and 0 otherwise. The measurement process as well as the dropout process depend on this latent variable, not only directly, but also through the subject-specific effects $\boldsymbol{b}_i$. The distribution of $\boldsymbol{Q}_i$ is multinomial and defined by $P(Q_{ik} = 1) = \pi_k$, where k ranges from 1 to g and π_k denotes the group or component probability. Note that the component probabilities obey $\sum_{k=1}^{g} \pi_k = 1$. In what follows, π_k will also be called the prior probability of an observation belonging to the kth component of the mixture.

The measurement process is specified by means of a so-called heterogeneity linear mixed model, originally proposed by Verbeke and Lesaffre (1996) and also described by Verbeke and Molenberghs (2000, Chapter 12). The model is given by

$$\boldsymbol{Y}_i | q_{ik} = 1, \boldsymbol{b}_i \sim N(\boldsymbol{X}_i \boldsymbol{\beta}_k + \boldsymbol{Z}_i \boldsymbol{b}_i, \boldsymbol{\Sigma}_i^{(k)}),$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices, $\boldsymbol{\beta}_k$ are fixed effects, possibly depending on the group components, and $\boldsymbol{b}_i$ denote the shared parameters, following a mixture of g normal distributions with mean vectors $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{D}_k$, that is,

$$\boldsymbol{b}_i | q_{ik} = 1 \sim N(\boldsymbol{\mu}_k, \boldsymbol{D}_k),$$

and therefore

$$\boldsymbol{b}_i \sim \sum_{k=1}^{g} \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{D}_k).$$

The measurement error terms $\boldsymbol{\varepsilon}_i$ follow a normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_i^{(k)}$ and are independent of the shared parameters.

The mean and the variance of $\boldsymbol{Y}_i$ can be derived as

$$\begin{aligned} E(\boldsymbol{Y}_i) &= \boldsymbol{X}_i \sum_{k=1}^{g} \pi_k \boldsymbol{\beta}_k + \boldsymbol{Z}_i \sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k, \\ \text{Var}(\boldsymbol{Y}_i) &= \boldsymbol{Z}_i' \left[\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k^2 - \left(\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k \right)^2 + \sum_{k=1}^{g} \pi_k \boldsymbol{D}_k \right] \boldsymbol{Z}_i + \sum_{k=1}^{g} \pi_k \boldsymbol{\Sigma}_i^{(k)}. \end{aligned}$$

Further, we have to assume that the shared effects are 'calibrated', that is, $\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k = \boldsymbol{0}$, then the latter expressions for the mean and variance simplify to:

$$\begin{aligned} E(\boldsymbol{Y}_i) &= \boldsymbol{X}_i \sum_{k=1}^{g} \pi_k \boldsymbol{\beta}_k, \\ \text{Var}(\boldsymbol{Y}_i) &= \boldsymbol{Z}_i' \left[\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k^2 + \sum_{k=1}^{g} \pi_k \boldsymbol{D}_k \right] \boldsymbol{Z}_i + \sum_{k=1}^{g} \pi_k \boldsymbol{\Sigma}_i^{(k)}. \end{aligned}$$

The dropout model is specified consistently with (6.3) and (6.4), but now the shared parameter $\boldsymbol{b}_i$ and the latent class membership indicators q_{ik} are part of the model:

$$g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik}) = P(D_i = j | D_i \geq j, \boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik} = 1),$$

where $\boldsymbol{w}_{ij}$ is a vector containing all relevant covariates. An obvious choice is to further assume that

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \boldsymbol{w}_{ij} \boldsymbol{\gamma}_k + \boldsymbol{\lambda} \boldsymbol{b}_i.$$

The joint likelihood of the measurement and dropout processes takes the form:

$$\begin{aligned} f(\boldsymbol{y}_i, d_i) &= \sum_{k=1}^{g} P(q_{ik} = 1) f(\boldsymbol{y}_i, d_i | q_{ik} = 1) \\ &= \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i, d_i | q_{ik} = 1, \boldsymbol{b}_i) f_k(\boldsymbol{b}_i) d\boldsymbol{b}_i \\ &= \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i | q_{ik} = 1, \boldsymbol{b}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) f(d_i | q_{ik} = 1, \boldsymbol{b}_i, \boldsymbol{w}_i) f_k(\boldsymbol{b}_i) d\boldsymbol{b}_i, \end{aligned} \tag{8.1}$$

where $f(\boldsymbol{y}_i|q_{ik}=1,\boldsymbol{b}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i)$ is the density function of the normal distribution $N(\boldsymbol{X}_i\boldsymbol{\beta}_k+\boldsymbol{Z}_i\boldsymbol{b}_i,\boldsymbol{\Sigma}_i^{(k)})$, $f_k(\boldsymbol{b}_i)$ is the density function of $N(\boldsymbol{\mu}_k,\boldsymbol{D}_k)$, and

$$f(d_i|q_{ik}=1,\boldsymbol{b}_i,\boldsymbol{w}_i) \tag{8.2}$$

$$= \begin{cases} \displaystyle\prod_{j=2}^{n}[1-g_{ij}(\boldsymbol{w}_{ij},\boldsymbol{b}_i,q_{ik})] & \text{for a completer } (d_i=n+1), \\ g_{id_i}(\boldsymbol{w}_{id_i},\boldsymbol{b}_i,q_{ik})\displaystyle\prod_{j=2}^{d_i-1}[1-g_{ij}(\boldsymbol{w}_{ij},\boldsymbol{b}_i,q_{ik})] & \text{for a dropout } (d_i\leq n). \end{cases}$$

The latter equation is the latent-class mixture analogue of (6.3). Note that the dropout model can depend, not only on the outcomes, but also on relevant covariates such as treatment allocation, time, gender, age, etc. Not all models that can be formulated in this way are identified, so restrictions are needed. We return to this issue in Section 8.2.

Whereas selection models and pattern-mixture models derive from two different factorizations of the joint density of the measurement and dropout processes, the latent-class mixture model is based on assuming an additional latent structure. The selection model lends itself naturally to formulate such concepts as MAR and ignorability, even though they can be considered in the pattern-mixture framework as well (Molenberghs, Michiels, Kenward and Diggle, 1998b; Kenward, Molenberghs and Thijs, 2003). In the pattern-mixture model, the observed dropout patterns are taken into account when modeling the measurement process. The latent-class mixture models modify this idea by grouping the subjects by means of a latent variable, thereby accounting for inter-group differences both in terms of their dropout pattern as well as their measurement profiles.

8.2 Likelihood Function and Estimation

Estimation of the unknown parameters in the latent-class mixture model can be based on the maximum likelihood principle. The likelihood function of the latent-class mixture model is formulated in Section 8.2.1. Since it would be very cumbersome to maximize this likelihood function analytically, the EM algorithm (Dempster, Laird and Rubin, 1977) is proposed as it is a practical tool for maximum likelihood estimation in the case of finite mixtures (Redner and Walker, 1984). In Section 8.2.2 an outline is provided of how the likelihood can be maximized using the EM algorithm.

8.2.1 The Likelihood Function

Let $\boldsymbol{\pi}$ be the vector of component probabilities $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_g)$ and group all other unknown parameters of the measurement process in the vector $\boldsymbol{\theta}$, of the dropout process in $\boldsymbol{\psi}$, and of the mixture distribution in $\boldsymbol{\alpha}$. If $\boldsymbol{\sigma}$ denotes the vector of covariance parameters of all $\boldsymbol{\Sigma}_i^{(k)}$, $\boldsymbol{\delta}$ the covariance parameters of all $\boldsymbol{D}_k$, $\boldsymbol{\mu}' = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g)$, and $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_g)$, then $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma})$, $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\lambda})$ and $\boldsymbol{\alpha} = (\boldsymbol{\mu}, \boldsymbol{\delta})$. Denote by $\boldsymbol{\Omega}$ the vector containing all unknown parameters in the model, that is $\boldsymbol{\Omega}' = (\boldsymbol{\pi}', \boldsymbol{\theta}', \boldsymbol{\psi}', \boldsymbol{\alpha}')$. Estimation and inference for the $\boldsymbol{\Omega}$ will now be based on the observed data likelihood, $L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$, obtained by integrating out the unobserved data from the joint distribution of measurement and dropout process and expressed by:

$$
\begin{aligned}
& L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}) \\
&= \prod_{i=1}^{N} f(\boldsymbol{y}_i^o, d_i|\boldsymbol{\Omega}) \\
&= \prod_{i=1}^{N} \int f(\boldsymbol{y}_i, d_i|\boldsymbol{\Omega})\, d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{N} \int \left\{ \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1) f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i \right\} d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{N} \sum_{k=1}^{g} \pi_k \int \left\{ \int f(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1)\, d\boldsymbol{y}_i^m \right\} f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i \\
&= \prod_{i=1}^{N} \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i^o|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1) f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i, \qquad (8.3)
\end{aligned}
$$

where $\boldsymbol{y}^{o'} = (\boldsymbol{y}_1^o, \ldots, \boldsymbol{y}_N^o)$ is the vector containing all observed response values and $\boldsymbol{d} = (d_1, \ldots, d_N)$ is the vector of all values of the dropout indicator.

Note that this likelihood function is invariant under the $g!$ possible permutations of the parameters corresponding to each of the g mixture components. To overcome this, the constraint suggested by Aitkin and Rubin (1985), $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_g$, is imposed.

Identifiability is a delicate issue. Böhning (1999) shows that a mixture of two normals with simultaneously different means and different variances is not identifiable. Such problems arise from latency, now occurring through latent classes, random effects, and missingness. In line with Böhning (1999) and McLachlan and Peel (2000), one could consider several variations to the target model. The likelihood values, parameter estimates, and information matrices can then be studied in view of identifiability. All of this cautions against certain uncritical uses of the model.

For example, one should not place a blind belief in the number of components resulting when applying the model; this would be a license for mischief. Rather, the model can play a role as a sensitivity analysis tool. The allocation of subjects to latent groups can help formulate hypotheses, which then have to be checked against substantive scientific knowledge and/or in follow up studies. On the other hand, when focusing on certain inferences, such as testing for treatment effect, the number of components, for example, may be less essential.

The log-likelihood function corresponding to likelihood function (8.3) is

$$\begin{aligned}&\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})\\ &= \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i^o|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1) f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i \right\}. \quad (8.4)\end{aligned}$$

To maximize (8.4) with respect to $\boldsymbol{\Omega}$, the Estimation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) will be employed. The EM algorithm is a numerical iterative procedure, designed for maximum likelihood estimation in situations with missing data. Here, the underlying latent variable $\boldsymbol{Q}_i$, representing component membership, will be considered missing. Thus, the response vector $\boldsymbol{Y}_i^o$ and the dropout indicator D_i, together with the (unobserved) population indicators $\boldsymbol{Q}_i$ can be seen as the so-called *augmented data*, whereas vectors $\boldsymbol{Y}_i^o$ and D_i constitute the observed data.

Clearly, the likelihood function $L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$ corresponds to the incomplete data. Since the joint density of $\boldsymbol{Y}_i^o$, D_i and $\boldsymbol{Q}_i$ equals

$$\begin{aligned}&f_i(\boldsymbol{y}_i^o, d_i, Q_{i1}=q_{i1}, \ldots, Q_{ig}=q_{ig})\\ &= f_i(\boldsymbol{y}_i^o, d_i|Q_{i1}=q_{i1}, \ldots, Q_{ig}=q_{ig}) \cdot P(Q_{i1}=q_{i1}, \ldots, Q_{ig}=q_{ig})\\ &= \left\{ \prod_{k=1}^{g} [f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})]^{q_{ik}} \right\} \cdot \left\{ \prod_{k=1}^{g} \pi_k^{q_{ik}} \right\}\\ &= \prod_{k=1}^{g} [\pi_k f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})]^{q_{ik}},\end{aligned}$$

the joint likelihood $L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ of the augmented data, that is, the likelihood function that would have been obtained if the values $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{ig})'$ of the population indicators $\boldsymbol{Q}_i$ had been observed, would be

$$L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q}) = \prod_{i=1}^{N} \prod_{k=1}^{g} [\pi_k f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})]^{q_{ik}}, \quad (8.5)$$

with $\boldsymbol{q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n)'$ the vector of all hypothetically observed population indicators. The log-likelihood function corresponding to the likelihood function (8.5) takes the form

$$\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q}) = \sum_{i=1}^{N} \sum_{k=1}^{g} q_{ik} \left\{ \ln \pi_k + \ln f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) \right\}. \tag{8.6}$$

8.2.2 Estimation Using The EM Algorithm

Maximizing $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ would be analytically and computationally easier than maximizing the log-likelihood $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$. However, the estimates obtained from maximizing $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ with respect to $\boldsymbol{\Omega}$, will depend on the unobserved indicators $\boldsymbol{q}$. Therefore, the EM algorithm is advisable, since then we will maximize the expected value of $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ with respect to $\boldsymbol{\Omega}$, where the expectation is taken over all unobserved $\boldsymbol{q}$, that is, $E[\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{Q})|\boldsymbol{y}, \boldsymbol{d}]$. This conditional expectation of $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ given $\boldsymbol{y}^o$ and $\boldsymbol{d}$, is calculated within the expectation (E) step of each iteration of the EM algorithm. In the maximization (M) step the expected log-likelihood function obtained from the E step is then maximized. Denote by $\mathcal{O}$ the expected log-likelihood function and call this the objective function. The EM algorithm is an iterative procedure, that is, it starts from an initial value $\boldsymbol{\Omega}^{(0)}$ for $\boldsymbol{\Omega}$, and then constructs a series of estimates $\boldsymbol{\Omega}^{(t)}$, which converges to the maximum likelihood estimator $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$. Initial values can be obtained from considering separate models for the measurement and dropout processes. Given $\boldsymbol{\Omega}^{(t)}$, the current estimate for $\boldsymbol{\Omega}$, the updated estimate $\boldsymbol{\Omega}^{(t+1)}$ is obtained through one iteration of the EM algorithm, that is, through one E step and one M step. Iteration then continuous until convergence is attained, that is, until

$$\left| \ell(\boldsymbol{\Omega}^{(t+1)}|\boldsymbol{y}^o, \boldsymbol{d}) - \ell(\boldsymbol{\Omega}^{(t)}|\boldsymbol{y}^o, \boldsymbol{d}) \right| < \varepsilon^*,$$

for some small, pre-specified $\varepsilon^* > 0$. More details are provided in the following.

The E Step

Let us describe the iteration step $t + 1$, in which the estimate is updated to $\boldsymbol{\Omega}^{(t+1)}$, using the obtained estimate from iteration step t, $\boldsymbol{\Omega}^{(t)}$. The E step consists of the calculation of the conditional expectation of $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$, given $\boldsymbol{y}^o$ and $\boldsymbol{d}$, which is

given by

$$\begin{aligned}\mathcal{O}(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)}) &= E\left[\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{Q})\,\middle|\,\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right]\\ &= E\left[\sum_{i=1}^{N}\sum_{k=1}^{g} Q_{ik}\left\{\ln\pi_k + \ln f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})\right\}\,\middle|\,\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right]\\ &= \sum_{i=1}^{N}\sum_{k=1}^{g} E\left[Q_{ik}\,\middle|\,\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right]\left\{\ln\pi_k + \ln f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})\right\}.\end{aligned}$$

Therefore, we need to calculate $E\left[Q_{ik}\,\middle|\,\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right]$:

$$\begin{aligned}E\left[Q_{ik}\,\middle|\,\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right] &= P\left(Q_{ik}=1|\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right)\\ &= \left.\frac{f_i\left(\boldsymbol{y}_i^o,d_i|Q_{ik}=1\right)P(Q_{ik}=1)}{f_i(\boldsymbol{y}_i^o,d_i)}\right|_{\boldsymbol{\Omega}^{(t)}}\\ &= \left.\frac{\pi_k f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})}{\sum_{k=1}^{g}\pi_k f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})}\right|_{\boldsymbol{\Omega}^{(t)}} = \pi_{ik}(\boldsymbol{\Omega}^{(t)}),\end{aligned}$$

where $\pi_{ik}(\boldsymbol{\Omega}^{(t)})$ is the posterior probability for the ith subject belonging to the kth component of the mixture. This means the E step reduces to the calculation of posterior probabilities $\pi_{ik}(\boldsymbol{\Omega}^{(t)})$, for $i=1,\ldots,N$ and $k=1,\ldots,g$. This requires calculation of $f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})$, and consequently integration over the unknown mixture component membership, which is done numerically using Gauss-Legendre quadrature.

The M Step

The updated estimate $\boldsymbol{\Omega}^{(t+1)}$ is now obtained from maximizing $\mathcal{O}(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)})$ with respect to $\boldsymbol{\Omega}$. From the E step we know that $\mathcal{O}$ equals

$$\begin{aligned}\mathcal{O}(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)}) &= \sum_{i=1}^{N}\sum_{k=1}^{g}\pi_{ik}(\boldsymbol{\Omega}^{(t)})\left\{\ln\pi_k + \ln f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})\right\}\\ &= \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{g}\pi_{ik}(\boldsymbol{\Omega}^{(t)})\ln\pi_k}_{=\mathcal{O}_1(\boldsymbol{\pi}|\boldsymbol{\Omega}^{(t)})} + \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{g}\pi_{ik}(\boldsymbol{\Omega}^{(t)})\ln f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})}_{=\mathcal{O}_2(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha}|\boldsymbol{\Omega}^{(t)})}\\ &= \mathcal{O}_1(\boldsymbol{\pi}|\boldsymbol{\Omega}^{(t)}) + \mathcal{O}_2(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\Omega}^{(t)}).\end{aligned} \tag{8.7}$$

The first term in (8.7) depends only on $\boldsymbol{\pi}$, whereas the second one depends only on $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\alpha}$. Hence, to find the maximum of the $\mathcal{O}$ function with respect to $\boldsymbol{\Omega}' = (\boldsymbol{\pi}', \boldsymbol{\theta}', \boldsymbol{\psi}', \boldsymbol{\alpha}')$, we can maximize both terms separately.

We first maximize the $\mathcal{O}$ function with respect to $\boldsymbol{\pi}$. This requires the maximization of $\mathcal{O}_1$, since $\mathcal{O}_2$ is independent of $\boldsymbol{\pi}$. Under the restriction $\sum_{k=1}^{g} \pi_k = 1$, we can rewrite $\mathcal{O}_1$ as follows

$$\mathcal{O}_1(\boldsymbol{\pi}|\boldsymbol{\Omega}^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{g-1} \pi_{ik}(\boldsymbol{\Omega}^{(t)}) \ln \pi_k + \sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)}) \ln\left(1 - \sum_{k=1}^{g-1} \pi_k\right).$$

If we now set all first-order derivatives with respect to $\pi_1, \ldots, \pi_{g-1}$ equal to zero, the updated estimate satisfies

$$\begin{aligned}
\frac{\partial \mathcal{O}_1}{\partial \pi_k} = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{N} \frac{\pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\pi_k^{(t+1)}} - \sum_{i=1}^{N} \frac{\pi_{ig}(\boldsymbol{\Omega}^{(t)})}{1 - \sum_{k=1}^{g-1} \pi_k^{(t+1)}} = 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{N} \frac{\pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\pi_k^{(t+1)}} = \sum_{i=1}^{N} \frac{\pi_{ig}(\boldsymbol{\Omega}^{(t)})}{\pi_g^{(t+1)}} \\
&\Leftrightarrow \quad \frac{\pi_k^{(t+1)}}{\pi_g^{(t+1)}} = \frac{\sum_{i=1}^{N} \pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})}. \qquad (8.8)
\end{aligned}$$

This, in turn, implies that

$$\begin{aligned}
1 = \sum_{k=1}^{g} \pi_k^{(t+1)} &= \sum_{k=1}^{g} \frac{\pi_g^{(t+1)} \sum_{i=1}^{N} \pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})} \\
&= \frac{\pi_g^{(t+1)} \sum_{i=1}^{N} \overbrace{\sum_{k=1}^{g} \pi_{ik}(\boldsymbol{\Omega}^{(t)})}^{=1}}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})} = \frac{N\, \pi_g^{(t+1)}}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})},
\end{aligned}$$

and hence

$$\pi_g^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)}). \qquad (8.9)$$

From (8.8) and (8.9) it follows that the updated estimates $\pi_k^{(t+1)}$, $k = 1, \ldots, g$, are given by

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \pi_{ik}(\boldsymbol{\Omega}^{(t)}),$$

that is, the updated mixture component probabilities are equal to the average posterior probabilities.

Next, to maximize of the $\mathcal{O}$ function with respect to $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\alpha}$, it suffices to maximize

$$\mathcal{O}_2(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}|\boldsymbol{\Omega}^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{g} \pi_{ik}(\boldsymbol{\Omega}^{(t)}) \ln f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})$$

with respect to these parameters. However, in general, this cannot be done analytically. Therefore, a classical numerical maximization procedure such as Newton-Raphson is needed. Note that in such cases, the EM algorithm is doubly iterative, which can have a non-negligible impact on the computation time.

Some Remarks Regarding the EM Algorithm

It has been shown (Rubin, 1987) that an iteration within the EM algorithm always increases the value of the likelihood function $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$, under mild regularity conditions, that is,

$$\ell(\boldsymbol{\Omega}^{(t+1)}|\boldsymbol{y}^o, \boldsymbol{d}) > \ell(\boldsymbol{\Omega}^{(t)}|\boldsymbol{y}^o, \boldsymbol{d}) \quad \text{for all } t.$$

This is called the monotonicity property of the EM algorithm, which guarantees convergence of the iterative procedure, provided a finite maximum exists. However, this convergence can be painfully slow, particularly with poorly selected starting values. Apart from the local maxima resulting from the non-identifiability problem, there may be local maxima yielding different likelihood values (Böhning, 1999). This suggests the use of multiple sets of starting values. If regions exist where the likelihood is flat, it is said the likelihood has a *ridge*. The EM algorithm is capable of converging to some particular point on such a ridge, which is not the case for many other, more classical, maximization algorithms.

8.3 Classification

Upon fitting the latent-class mixture model to an incomplete set of repeated measurements, one is in a position to classify the study subjects examined into the various mixture components of the fitted model, that is, into the population's latent subgroups. Through the structure of the latent-class mixture model, the subdivision of the population in latent groups depends on the number of observed measurements: on the dropout indicator or pattern, as well as on the values of the observed response measurements. Hence, the classification of subjects into different latent groups can be useful to assess the coherence between the dropout process and the measurement

process. In certain cases such latent groups can have a biological or otherwise substantive meaning. For instance, subjects of one group could have higher response values and drop out earlier in the study, whereas subjects of another group have lower values but remain longer in the study.

The decision to which component of the mixture, or, equivalently, to which subgroup of the population, a specific subject is most likely to belong will be based on posterior probabilities. Recall that $P(Q_{ik} = 1) = \pi_k$, thus the component probabilities π_k, $k = 1, \ldots, g$, express how likely the ith subject is to belong to group k without using information from the outcomes, $\boldsymbol{y}_i^o$, and dropout pattern, d_i. For this reason, the component probabilities are often called *prior* probabilities.

The *posterior* probability for subject i to belong to the kth group is given by

$$\begin{aligned} \pi_{ik} = P(Q_{ik} = 1 | \boldsymbol{y}_i^o, d_i) &= \left. \frac{f_i\left(\boldsymbol{y}_i^o, d_i | Q_{ik} = 1\right) P(Q_{ik} = 1)}{f_i(\boldsymbol{y}_i^o, d_i)} \right|_{\widehat{\boldsymbol{\Omega}}} \\ &= \left. \frac{\pi_k f_{ik}(\boldsymbol{y}_i^o, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})}{\sum_{k=1}^{g} \pi_k f_{ik}(\boldsymbol{y}_i^o, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})} \right|_{\widehat{\boldsymbol{\Omega}}}, \end{aligned}$$

where $\widehat{\boldsymbol{\Omega}}$ is the vector of parameter estimates resulting from the EM algorithm. This expresses how likely it is that the ith subject is to belong to group k, taking into account the observed response $\boldsymbol{y}_i^o$ as well as the dropout indicator d_i of that subject. Using these posterior probabilities, we can apply the following classification rule

$$\text{Classify subject } i \text{ into component } k \iff \pi_{ik} = \max_j \{\pi_{ij}\},$$

assigning subject i to the component to which it is most likely to belong.

However, we do need to be cautious with the resulting classification into latent subgroups because, for a particular subject i, the vector of posterior probabilities is given by $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ig})$ with $\sum_{k=1}^{g} \pi_{ik} = 1$. For a good comfort level, one of these posterior probabilities for subject i would be close to 1, in which case the classification of this subject is obvious and likely to be correct. However, another scenario is that two or more posterior probabilities are almost equal, of which one is the maximum of all posterior probabilities for that particular subject. For example, suppose we have $g = 2$ latent subgroups and subject i has posterior probabilities $(\pi_{i1}, \pi_{i2}) = (0.55, 0.45)$. In this case subject i would be allocated into group 1 using the classification rule, but this should be done with low confidence. This makes classification nearly random and misclassification is likely to occur. Perhaps it is safer to assert that this subject lies between both groups, in this sense being an outlier, or rather an 'in-lier'. Therefore, rather than merely considering the classification of sub-

jects into the latent subgroups, it is instructive to inspect the posterior probabilities in full.

A separate issue is the (prespecified) number of g of latent groups. It is hard to choose g with great confidence purely on a priori grounds and therefore it is advisable to explore the stability of the conclusions, by way of additional sensitivity analysis, by varying g across a range.

8.4 Simulation Study

An advantage of the latent-class mixture model is its flexible structure, which potentially makes the model a helpful analysis tool for incomplete longitudinal data. However, as already seen in Section 8.2.2, the estimation of the model parameters is based on a doubly iterative method, which we might expect to be computationally intensive. To assess whether this disadvantage counterbalances the advantage of model flexibility, and to assess performance, we conduct a simulation study. We first describe in Section 8.4.1 a simplification of the latent-class mixture model which is used in the following simulation study as well as later in the application in Section 8.5. Following this, the design and results of the simulation study are displayed in Sections 8.4.2 and 8.4.3, respectively.

8.4.1 A Simplification of the Latent-Class Mixture Model

In what follows, we assume equal covariance matrices for the different mixture components, $\boldsymbol{D}_1 = \ldots = \boldsymbol{D}_g = \boldsymbol{D}$, as well as equal residual covariance matrices, $\boldsymbol{\Sigma}_i^{(1)} = \ldots = \boldsymbol{\Sigma}_i^{(g)} = \boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}_n$, which leads to $\boldsymbol{Y}_i | q_{ik} = 1, \boldsymbol{b}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \sigma^2\boldsymbol{I}_n)$, with $\boldsymbol{b}_i \sim \sum_{k=1}^{g} \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{D})$.

Furthermore, we simplify the general latent-class mixture model in two steps. First, it is assumed that there is only one subject-specific effect b_i, a shared intercept, influencing the measurement process, not the dropout process. Second, the measurement process is assumed to depend on the latent variable, not in a direct way, but only through the shared intercept. The model is depicted in Figure 8.2.

8.4.2 Design of the Simulation Study

The simulation study is structured as follows. Two-hundred and fifty datasets are simulated, each containing measurements and covariate information of 100 subjects. The latent variable in the model is assumed to split the subjects into two latent

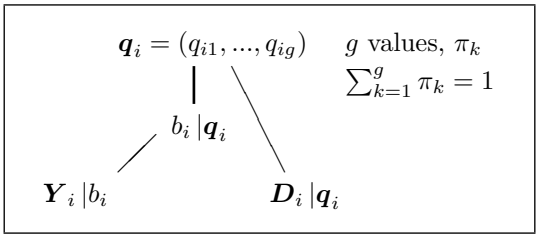


Figure 8.2: *A simplification of the latent-class mixture model.*

subgroups with component probabilities $\pi_1 = 0.6$ and $\pi_2 = 1 - \pi_1 = 0.4$, respectively. There are five measurement occasions and the outcome follows a linear trend over time with intercept $\beta_0 = 9.4$ and slope $\beta_1 = 2.25$. The shared intercept follows a mixture of two normal distributions with different means for both latent groups: $\mu_1 = -4.4$ and $\mu_2 = -\frac{\pi_1 \mu_1}{\pi_2} = 6.6$. In line with Section 8.4.1, the variances of these two normal distributions are set equal and denoted by d^2. The measurement error variance is σ^2.

Four different settings will be considered, based on varying d^2 and σ^2. In the first setting both variance parameters are chosen to be relatively small, $d = 2.0$ and $\sigma = 0.25$. While only the measurement error variance is increased in the second setting, $\sigma = 0.75$, both variance parameters are increased in the third setting, $d = 3.5$ and $\sigma = 1.00$. Up to the third setting, the chosen parameters result in a bimodal, well-separated mixture distribution. Since this might improve estimation of the parameters, we consider a fourth, unimodal setting with $d = 6$ and $\sigma = 2$.

Finally, in the dropout model, the logistic regression is based on an intercept only, which differs for both latent classes, namely, $\gamma_1 = -2.5$ and $\gamma_2 = -1.25$, respectively, with corresponding probabilities 0.73 and 0.45 of completing the study.

The latent-class mixture model can now be formulated as follows. For subject $i = 1, \ldots, 100$, belonging to latent group $k = 1, 2$, the measurement at time $j = 1, \ldots, 5$ is modelled by

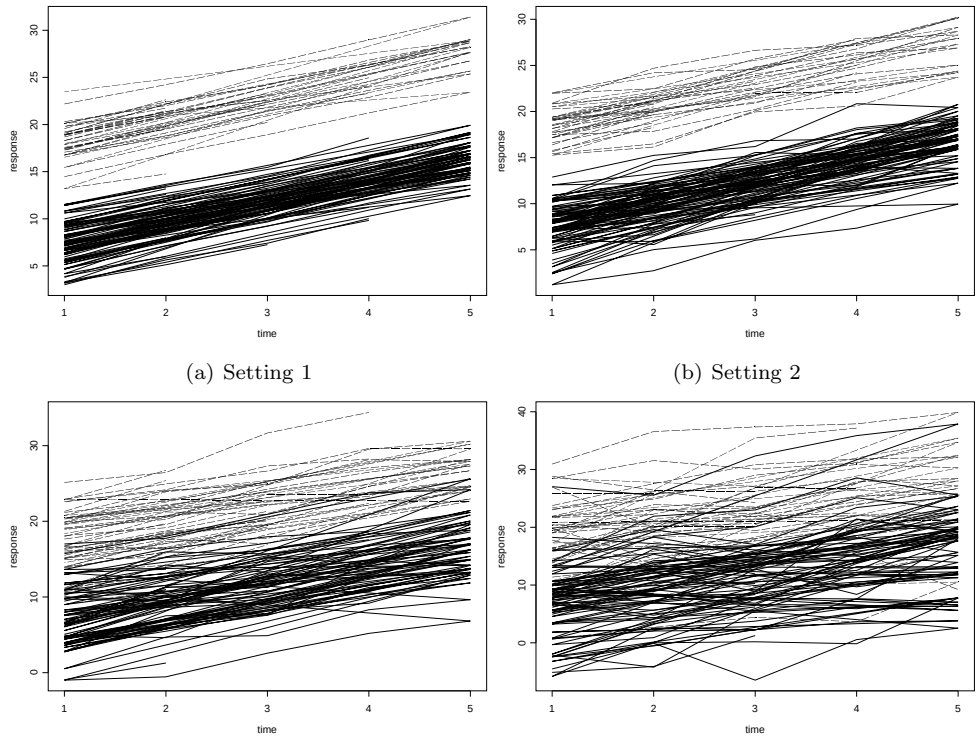
$$Y_{ij} = \beta_0 + \beta_1 \, time_j + b_i + \boldsymbol{\varepsilon}_{ij}^{(k)}, \tag{8.10}$$

with

$$b_i \sim \pi_1 N\left(\mu_1, d^2\right) + \pi_2 N\left(\mu_2, d^2\right) \quad \text{and} \quad \boldsymbol{\varepsilon}_i^{(k)} \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_5\right). \tag{8.11}$$

Furthermore , the dropout model is expressed as

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_k. \tag{8.12}$$



(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Figure 8.3: *Simulation study. Individual profiles for one dataset randomly chosen out of 250 simulated datasets, for each of the three simulation settings. Solid lines correspond to subjects from the first latent group, dashed lines to subjects from the second one.*

8.4.3 Results of the Simulation Study

To get a better feel for the four simulation settings, a dataset was selected randomly from the 250 simulated datasets for each setting. Figure 8.3 shows the individual profiles of these datasets. Table 8.1 contains the results of the simulation study. As well as comparing the mean estimates and true values of the parameters through the bias, we also consider the mean squared error (MSE), simultaneously involving bias and precision.

We discuss the four simulation settings in turn. For the first, Figure 8.3(a) shows a clear distinction between both groups, which owes to the small variance, d^2, of the mixture distribution, relative to the systematic difference between the group means,

Table 8.1: *Simulation study. Results of the simulation study: mean and true value, bias, and mean squared error (MSE) of the parameters, under the three simulations settings.*

Setting 1					**Setting 2**				
Effect	Mean	True	Bias	MSE	Effect	Mean	True	Bias	MSE
Measurement Model					*Measurement Model*				
β_0	9.37	9.40	-2.84×10^{-2}	8.07×10^{-4}	β_0	9.34	9.40	-5.75×10^{-2}	3.31×10^{-3}
β_1	2.25	2.25	1.30×10^{-4}	1.68×10^{-8}	β_1	2.25	2.25	7.56×10^{-4}	5.72×10^{-7}
σ	0.25	0.25	-2.49×10^{-4}	6.18×10^{-8}	σ	0.75	0.75	6.27×10^{-4}	3.93×10^{-7}
μ_1	-4.39	-4.40	1.31×10^{-2}	1.73×10^{-8}	μ_1	-4.36	-4.40	4.48×10^{-2}	2.00×10^{-3}
d	1.98	2.00	-1.70×10^{-2}	2.89×10^{-4}	d	1.97	2.00	-2.53×10^{-2}	6.38×10^{-4}
π_1	0.60	0.60	4.60×10^{-4}	2.12×10^{-7}	π_1	0.60	0.60	4.22×10^{-3}	1.79×10^{-5}
Dropout Model					*Dropout Model*				
γ_1	-2.52	-2.50	-2.28×10^{-2}	5.19×10^{-4}	γ_1	-2.51	-2.50	-1.26×10^{-2}	1.58×10^{-4}
γ_2	-1.26	-1.25	-1.23×10^{-2}	1.53×10^{-4}	γ_2	-1.27	-1.25	-2.30×10^{-2}	5.27×10^{-4}
Setting 3					**Setting 4**				
Effect	Mean	True	Bias	MSE	Effect	Mean	True	Bias	MSE
Measurement Model					*Measurement Model*				
β_0	9.44	9.40	3.83×10^{-2}	1.46×10^{-3}	β_0	9.59	9.40	1.92×10^{-1}	3.70×10^{-3}
β_1	2.25	2.25	1.91×10^{-4}	3.66×10^{-8}	β_1	2.24	2.25	-1.44×10^{-2}	2.06×10^{-4}
σ	0.99	1.00	-5.45×10^{-3}	2.06×10^{-5}	σ	2.01	2.00	6.07×10^{-3}	3.69×10^{-5}
μ_1	-4.69	-4.40	-2.86×10^{-1}	8.18×10^{-2}	μ_1	-4.84	-4.40	-4.39×10^{-1}	1.93×10^{-1}
d	3.43	3.50	-7.00×10^{-2}	4.90×10^{-3}	d	6.02	6.00	2.03×10^{-2}	4.10×10^{-4}
π_1	0.57	0.60	3.36×10^{-2}	1.13×10^{-3}	π_1	0.52	0.60	-8.06×10^{-2}	6.50×10^{-3}
Dropout Model					*Dropout Model*				
γ_1	-2.61	-2.50	-1.07×10^{-1}	1.14×10^{-2}	γ_1	-2.97	-2.50	-4.73×10^{-1}	2.23×10^{-1}
γ_2	-1.27	-1.25	-2.04×10^{-2}	4.17×10^{-4}	γ_2	-1.29	-1.25	-3.89×10^{-2}	1.51×10^{-3}

$\mu_1 - \mu_2$. Furthermore, the small measurement error variance, σ^2, ensures the within-subject variability to be small, resulting in almost straight individual profiles. From Table 8.1, the mean estimates of the parameters are close to the true values, with biases of the order 10^{-2} or less. Together with small MSE values, the magnitude of which does not exceed 10^{-4}, this indicates the fit of the latent-class mixture model is very close to the simulated data. This was expected due to earlier observations.

Increasing the measurement error variance in the second simulation setting leads to an increased within-subject variability. The discrepancy between both latent groups is still present (Figure 8.3(b)). The bias increases slightly, but remains of the same order. For the MSE values, we observe a small increase, but its magnitude does not exceed 10^{-3}. We can therefore conclude the model fits the data well, even with a larger within-subject variability.

In the penultimate simulation setting, not only the measurement error variance is increased, but also the variance in the mixture components. In Figure 8.3(c), we observe that on top of the larger within-subject variability, the gap between both latent groups now vanishes. The discrepancy between the groups seems to have vanished, and profiles appear to be homogeneous. We consider the results in Table 8.1 to examine influences on the model fit. For some of the parameters, the mean estimates deviate little from the true value. However, bias and MSE values remain small, the order of magnitude not exceeding 10^{-1} and 10^{-3}, respectively. Thus, here too, the latent-class mixture model does fit the simulated data well.

Finally, in the last simulation setting, in which even larger values for both variance parameters result in simulated data following an unimodal mixture distribution, profiles again seem to be homogeneous (Figure 8.3(d)). Remarkably, even in this setting, bias and MSE values remain small, both with order of magnitude below 10^{-1}.

In all four simulation settings, we ascertain small bias and MSE values. This is true, not only for the model parameters, but also for derived quantities, such as the treatment effect at the last time and the area under the curve. Thus from the four simulation settings we conclude that the latent-class mixture model does fit well. Remark that this applies even when the mixture distribution is unimodal. The equivalent statement for a real application is that the fit allegedly will be good in most cases where the researcher has decent insight into the true mean structure. Computation time increased from about 30 minutes for fitting the latent-class mixture model to a simulated dataset of the first setting, to a bit over two hours for fitting one of the later settings. Thus, fitting the latent-class mixture model is not unreasonable in terms of computation time, perhaps against initial expectation.

8.5 Analysis of the First Depression Trial

We apply the latent-class mixture model to the first depression trial, introduced in Section 2.2. In the two subsequent sections, a latent-class mixture model is fitted to the depression trial and a sensitivity analysis performed. The latter establishes the latent-class mixture model as a viable sensitivity tool.

8.5.1 Formulating a Latent-Class Mixture Model

A latent-class mixture model is fitted to the data from the first depression trial, assuming the patients can be split into g latent subgroups.

The mean structure is determined based on an exploratory analysis. As a result, the heterogeneity linear mixed model for the change in $HAMD_{17}$ score includes as fixed effects an intercept, the treatment variable, the baseline $HAMD_{17}$ score, the linear and quadratic time variable, and the interaction between treatment and time. The parameter values for these fixed effects are assumed to be equal across the g latent subgroups. The measurement error terms are assumed to be independent and to follow a normal distribution with mean 0 and variance σ^2. A shared intercept is included in the measurement model, which follows a mixture of g normal distributions with different means, $\mu_1, \ldots, \mu_g$ respectively, but with equal variance d^2.

The dropout process is modelled based on a logistic regression, which includes and intercept and slope that can differ between latent subgroups ($\gamma_{0,1}, \ldots, \gamma_{0,g}$ corresponding to the intercept, and $\gamma_{1,1}, \ldots, \gamma_{1,g}$ corresponding to the slope).

At first, the latent-class mixture model has essentially the same structure as the one used in the simulation study in Section 8.4, based on (8.10)–(8.12), with the addition of covariates described above. Afterwards, we extend the model by adding the shared intercept to the dropout model as well, meaning the dropout model changes from

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_{0,k} + \gamma_{1,k}\, t_j \tag{8.13}$$

to

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_{0,k} + \gamma_{1,k}\, t_j + \lambda\, b_i, \tag{8.14}$$

where t_j is the jth visit.

An overview of the models considered is given in Table 8.2. The models are fitted using GAUSS code, which is outlined in Section 11.7. Since assessing the number of components by a classical likelihood ratio test is not valid in the mixture model framework (McLachlan and Peel, 2000), we calculated the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) for all models.

Table 8.2: *First depression trial. Information criteria AIC and BIC, for models with dropout model (11.3) or (11.4), and* $g = 1, 2, 3$.

Model	Dropout Model	g	# Par	-2ℓ	AIC	BIC
1	$\gamma_{0,k} + \gamma_{1,k}\, t_j$	1	10	4676.07	4696.08	4727.44
2	$\gamma_{0,k} + \gamma_{1,k}\, t_j$	2	14	4662.37	**4690.37**	4734.27
3	$\gamma_{0,k} + \gamma_{1,k}\, t_j$	3	18	4662.03	4698.03	4754.48
4	$\gamma_{0,k} + \gamma_{1,k}\, t_j + \lambda\, b_i$	1	11	4669.12	4691.12	**4725.61**
5	$\gamma_{0,k} + \gamma_{1,k}\, t_j + \lambda\, b_i$	2	15	4662.02	4692.02	4739.06

A model building exercise is performed starting with fitting a one-component latent-class mixture model, which comes down to a classical shared-parameter model, as well as a two-component latent-class mixture model. Next, we compare these models using the AIC and BIC criteria, and depending on the choice made by both criteria, we decide whether we fit a latent-class mixture model with three latent subgroups.

Table 8.2 shows that when assuming dropout model (11.3), AIC opts for the model with two latent subgroups (Model 2), whereas BIC gives preference to the shared-parameter model (Model 1). Further, in case of dropout model (11.4) however, both information criteria select the shared-parameter model (Model 4). Note that, since the dropout model in Model 1 does not depend on the shared intercept, the dropout model and the measurement model are independent, resulting in the MCAR assumption, whereas in Model 2, the dropout model is linked to the measurement model through the latent classes (MNAR).

Overall, the AIC criterion prefers Model 2, the 2-component latent-class mixture model with no random effect in the dropout model, whereas BIC picks Model 4, the classical shared-parameter model. Since both criteria select a different model, we first take a more detailed look at the latent-class mixture model with two components, indicated by AIC, whereas we consider the classical shared-parameter model in a sensitivity analysis in the next section.

Parameter estimates with corresponding standard errors and p-values of the two-component latent-class mixture model are shown in Table 8.3.

Once this latent-class mixture model has been fitted to the depression trial data, the posterior probabilities can be used to classify the patients into two subgroups as shown in Section 8.3. In this way, the 170 patients divide into 79 and 91 patients

Table 8.3: *First depression trial. Parameter estimates, standard errors, and p-values for the latent-class mixture model with two latent subgroups and dropout model (11.3).*

Parameter	Estimate	s.e.	p-value
Measurement Model			
β_0 : *intercept*	23.17	3.75	< 0.0001
β_1 : *treatment*	2.69	1.49	0.072
β_2 : *time*	-6.18	1.18	< 0.0001
β_3 : *time* $\times$ *treatment*	-0.52	0.24	0.028
β_4 : *baseline*	-0.42	0.07	< 0.0001
β_5 : $time^2$	0.41	0.10	< 0.0001
σ : *measurement error*	4.24	0.13	< 0.0001
Dropout Model			
$\gamma_{0,1}$: *intercept Group* 1	-8.58	3.57	0.009
$\gamma_{1,1}$: *time Group* 1	0.83	0.44	0.056
$\gamma_{0,2}$: *intercept Group* 2	-1.35	1.28	0.292
$\gamma_{1,2}$: *time Group* 1	-0.05	0.20	0.793
Shared Effects			
μ_1 : *mean shared intercept Group* 1	-3.64	0.43	< 0.0001
d : *variance shared intercept*	2.67	0.50	< 0.0001
$\pi_1 = \pi$: *prior probability Group* 1	0.48	0.10	< 0.0001
Loglikelihood		-2331.18	

classified into the first and second group, respectively. In Figure 8.4, the left panel represents the individual profiles of patients classified into the first latent group, and the right one represents the individual profiles of patients classified into the second group. Clearly, the first group corresponds to patients with lower $HAMD_{17}$ scores, which continue to decrease over time. This implies that these patients are the ones whose condition is improving. On the other hand, the second group contains patients with a higher change versus baseline compared to the patients from the first group. Their changes in $HAMD_{17}$ score fluctuate around 0, more specifically somewhere in

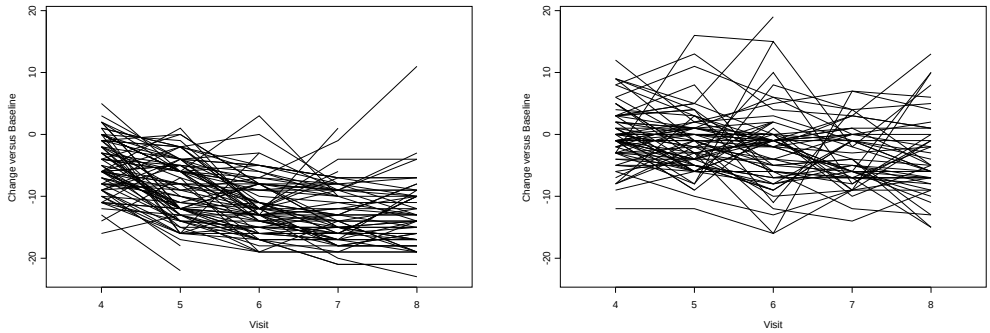


Figure 8.4: *First depression trial. Classification of the subjects of the depression trial based on a latent-class mixture model. Solid lines correspond to patients classified into first group (left panel), dashed lines to patients classified into second one (right panel).*

the region between -10 and 10. In addition, without taking into account the within-subject variability, their profiles appear more or less time-constant. A more formal comparison of both latent groups regarding their change of $HAMD_{17}$ score versus baseline confirms this association between the classification and the profile over time. Furthermore, a formal test for association of baseline values and group classification is not significant, indicating similar baseline $HAMD_{17}$ scores for patients in both groups.

Based on this difference in location of the profiles between both groups, this classification of subjects can be interpreted as being a split into acute versus chronic depression. Patients in both the acute and chronic groups enter the study with a baseline value indicating depression. However, the profiles of the patients in the acute group show recovery during the trial, whereas the depression score of patients in the chronic group remains more or less level.

Further, this difference between both latent groups is not due to treatment, since the classification of subjects in latent subgroups is independent of their treatment allocation. Indeed, the estimated odds ratio between the latent classification variable and the treatment allocation is 0.75, which was expected since the observed treatment groups are included in the mean structure of the measurement model. Moreover, when the treatment variable would be included in the dropout model, this independence would even increase.

Regarding the incompleteness of the patients in both latent groups, we notice a clear difference, which is confirmed by chi-square tests for independence, implying a

Table 8.4: *First depression trial. Classification of subjects based on the magnitude of posterior probabilities* π_{i1}.

π_{i1}	Classification	# Patients
$0.80 \rightarrow 1.00$	Clearly Group 1	61
$0.60 \rightarrow 0.80$	Group 1	8
$0.55 \rightarrow 0.60$	Doubtful, more likely Group 1	5
$0.45 \rightarrow 0.55$	Uncertain	8
$0.40 \rightarrow 0.45$	Doubtful, more likely Group 2	5
$0.20 \rightarrow 0.40$	Group 2	19
$0.00 \rightarrow 0.20$	Clearly Group 2	64

significant association between the dropout pattern and the latent classification. The first latent group mainly contains patients who complete the study, 62 in total. Of the 17 patients who drop out, merely 2 drop out at visit 6, 3 more at visit 7, and 12 patients missed the last visit only. The dropout percentage in the second latent group is larger, 48.4% compared to 21.5% in the first group, or 44 out of 91 patients. Of these incompleters, 17 drop out after the first visit, 10 more at visit 6, 11 at the penultimate visit, and 6 more at the last visit.

Finally, the latent groups can also be compared by focussing on demographic characteristics such as age, gender, and origin, yielding no association between the latent classification with either gender or origin, but a significant association with age. Consequently, patients in the acute group are younger than the patients in the chronic group, with a mean age of 38.5 and 42.4, and corresponding 95% confidence intervals $[36.1, 41.0]$ and $[40.0, 44.7]$,respectively. This is important insight, even though it may be hard to disentangle the causal relationship between age and chronicity. Even though age explains a part of the latent-class structure, it is relevant to further entertain the connection with chronicity, since this may have important implications for differentiated, more effective therapy.

However, as mentioned in Section 8.3, using this classification rule does not render insight into how strongly patients are allocated to one group rather than the other. This depends on the magnitude of the maximal posterior probability. Since the latent-class mixture model considered here only contains two latent groups, we merely need to consider one of the posterior probabilities, for example the posterior

probability that the subject belongs to group 1, π_{i1}. Based on this π_{i1}, the subjects can be classified following the guidelines of Table 8.4. If the posterior probability π_{i1} lies between 0.45 and 0.55, it is uncertain to which group the subject can be classified. Only 8 out of 170 patients in the depression trial are in this situation. For most patients, 152 or 89.4%, it is clear into which group they can be classified, since their maximal posterior probability is above 0.60. Furthermore, aforementioned association of the latent classification with the location of profiles, the dropout pattern, and patient's age as well as independence of baseline values and patient's origin and gender, is confirmed by testing the independence of these variables with the posterior probabilities, which can be viewed as continuous variables ranging from 0 to 1.

8.5.2 A Sensitivity Analysis

We now illustrate the use of the latent-class mixture model as a sensitivity analysis tool. In addition to the two-component latent-class mixture model introduced above, a classical shared-parameter model will be fitted, as well as a pattern-mixture model, and two selection models, based on the selection models introduced by Diggle and Kenward (1994); see also Section 6.1.1. All models contain the same fixed effects as in the two-component latent-class mixture model, that is, intercept, treatment, time, baseline, time^2, and treatment-by-time interaction.

The classical shared-parameter model, selected by the BIC criterion in Section 8.5.1, includes a shared intercept $b_i \sim N(0, d^2)$, conditional upon which the measurement model follows a normal distribution $\boldsymbol{Y}_i|b_i \sim N(\boldsymbol{X}_i\beta + b_i, \sigma^2\boldsymbol{I}_{n_i})$, and the dropout process is based on (11.4).

Next, in the Diggle-Kenward models, the covariance structure of the measurement process is assumed to be unstructured. The dropout model takes the conventional form (6.4). We consider an MAR version as well as the full MNAR version of the model.

Finally, a pattern-mixture model is fitted by adding pattern-specific intercepts and slopes to the same multivariate normal model as used in the Diggle-Kenward selection models. Notice that the classification function in the latent-class mixture model is a data driven approach to define groups, whereas pattern-mixture models use the assumption to define groups in function of dropout patterns.

As the main focus of the depression trial was in the treatment effect at the last visit, Table 8.5 shows the estimates, standard errors, and p-values for this effect under the five fitted models. Clearly, the p-values resulting from all five models are very similar and between around 0.07 and 0.11, yielding the same conclusion for the

Table 8.5: *First depression trial. Estimates, standard errors, and p-values for the treatment effect at visit 8, as well as the treatment-by-time interaction, for the latent-class mixture model, the shared-parameter model, the pattern-mixture model and both selection models, assuming either MAR or MNAR.*

	Treatment at Endpoint			Treatment × Time		
Model	Estimate	s.e.	p-value	Estimate	s.e.	p-value
Latent-Class Mixture Model	-1.44	0.91	0.114	-0.52	0.23	0.028
Shared-Parameter Model	-1.69	0.93	0.069	-0.50	0.24	0.035
Pattern-Mixture Model	-2.01	1.20	0.096	-0.55	0.31	0.077
MAR Selection Model	-2.17	1.25	0.082	-0.58	0.32	0.068
MNAR Selection Model	-2.16	1.24	0.081	-0.57	0.31	0.068

treatment effect at visit 8. Thus, the significance results are not sensitive to the model used, and hence more trust can be put into the conclusion. However, note that using both the two-component latent-class mixture model and the classical shared-parameter model, the standard error is reduced by 0.3 units, compared to either selection model, or pattern-mixture model, resulting in a more accurate confidence interval for the treatment effect at the last visit.

We continue by exploring the sensitivity of the treatment-by-time interaction by comparing the estimates, standard errors and p-values under the five fitted models in Table 8.5. The p-values are clearly moving around the 0.05 boundary. Whereas under the latent-class mixture model and the shared-parameter model the p-value is about 0.03, the p-value under both selection models and the pattern-mixture model is around 0.07. While one should be cautious with over-interpretation of p-values, there are contexts, such as regulated clinical trials, where strict decision rules are implemented. In such a case and when in addition the treatment by time interaction is the primary effect, the latent-class mixture model and the shared-parameter model would lead to a claim of significance, whereas this would not be justified with neither the selection models nor the pattern-mixture model.

8.6 Concluding Remarks

In this chapter, we have proposed latent-class mixture models for the analysis of longitudinal data subject to dropout. The model extends the shared-parameter model, in the sense that both the measurement and dropout processes are allowed to share a set of random effects, conditional upon which both processes are assumed to be independent. It can, at the same time, be seen as an extension of the pattern-mixture model, now with latent rather than explicitly observed groups. It uses ideas from random-effects and latent-class modeling. Therefore, it captures unobserved heterogeneity between latent subgroups of the population. The results from the simulation study underscore the fact that the flexibility of such latent-class mixture models outweighs the expected modeling complexity.

Apart from a flexible modeling technique, the proposed latent-class mixture model can be used as a sensitivity analysis instrument, and for further exploration of the latent class membership. However, when clusters are detected by classifying subjects into the latent subgroups, care has to be taken when interpreting latent classes, since in some applications they may merely be artifacts, without any substantive grounds. In others, there may be more basis for their existence. We believe, together with mental health scientists, the two-component classification in our example, refers to the natural split of the patients, regardless of which treatment they were allocated to, into the more chronic and the more acute ones. An additional word of caution is needed regarding the number of latent classes to be considered. This is a tricky but well documented problem (McLachlan and Peel, 2000). A practical way out is to consider several choices for the number of components, pick the most reasonable one, and assess whether alternative choices would substantially alter the conclusions. The latter leads us to the second additional purpose of the model, that is, the latent-class mixture model can be used as a sensitivity tool. Applying the tool to the first depression trial increased the confidence level in the conclusions reached.

An initial criticism of the latent-class mixture model would be its computational complexity. Evidently, the computational burden of the latent-class mixture does increase over non-latent-class models, but is still reasonable. For example, whereas the MNAR version of the Diggle-Kenward model takes around one hour and the one-component mixture needs about the same amount of time, the two-component mixture increases needs around one order of magnitude more. Furthermore, the performance of the algorithm is remarkably computationally stable, given sensible starting values (e.g., built from non-mixture classical models). Details on starting value selection are embedded in the software Section 11.7.

The development of this latent-class mixture model was published in Beunckens *et al.* (2007a).

9

Analysis of the ARMD Trial

In this chapter we will perform a thorough sensitivity analysis the age-related macular degeneration trial as introduced in Section 2.4. Recall that there are 240 subjects, 188 of which have a complete follow-up. Note that of the 52 subjects with incomplete follow-up, 8 exhibit a non-monotone pattern. While this does not hamper direct-likelihood analyses, it is a challenge for WGEE. One way forward is to monotonize the missingness patterns by means of multiple imputation and then conduct WGEE, or to switch to MI-GEE altogether. However, to be consistent throughout the analyses, those subjects with non-monotone profiles will not be taken into account. Further, neither will the subjects without any follow-up measurement be included in the analyses, since at least one measurement should be taken. This results in a data set of 226 of 240 subjects, or 94.17 %.

The original outcome is the visual acuity, that is the number of letters correctly read on a vision chard, which can be considered continuous. The dichotomous outcome is defined as increase or decrease in number of letters read compared with baseline. Section 9.1 and 9.2 are devoted to the analysis of the continuous and the binary outcome respectively, using first simple methods as discussed in Chapter 4, and second a number of viable candidates for a standard analysis including direct-likelihood (Chapter 4) and versions of generalized estimating equations, that is GEE, WGEE and MI-GEE (Chapters 3 and 5). Next, in Section 9.3 a sensitivity analysis based on the continuous outcomes is performed making use of models and sensitivity

tools within the selection, the pattern-mixture and the shared-parameter framework.

9.1 Simple and Direct-Likelihood Analysis of the Continuous Outcome

We consider a multivariate normal model, with unconstrained time trend under placebo, an time-specific treatment effect, and an unstructured variance covariance matrix. Let Y_{ij} be the visual acuity of subject $i = 1, \ldots, 226$, at time point $j = 1, \ldots, 4$, and T_i the treatment assignment for subject i, then the mean model takes the following form

$$E(Y_{ij}) = \beta_{j1} + \beta_{j2} T_i.$$

Thus, this longitudinal model features a full treatment by time interaction with eight mean model parameters. The direct-likelihood analysis based on all observed data is contrasted with the simple CC and LOCF analyses.

Results of these three analyses are displayed in Table 9.1. From the parameter estimates, it is clear that the treatment effects are underestimated when considering the completers only. Whereas for all observed data treatment effect at week 12 and week 52 are borderline significant, both turn insignificant when deleting subjects with missing values. For the LOCF analysis, going from week 4 to the end of the study, the underestimation of the treatment effect increases. Therefore, the effect at week 12 is borderline significant, but at week 52 it becomes insignificant. Once again, CC and LOCF miss important treatment differences, the most important one being that at week 52, the end of the study.

9.2 Analysis of the Binary Outcome

We now switch to the binary outcome, which indicates whether the number of letters correctly read at the follow-up occasion is higher or lower than the corresponding number of letters at baseline. Both marginal models and random-effects models are considered.

9.2.1 Marginal Models

In this section, a population-averaged (or marginal) model is used. In line with the previous section, we compare analyses performed on the completers only (CC), on the LOCF imputed data, as well as on the observed data. In all cases, standard GEE

Table 9.1: *Age-related macular degeneration trial. Parameter estimates (standard errors) for the linear-mixed models, fitted to the continuous outcome visual acuity on the CC and LOCF population, and on the observed data (direct-likelihood). p-values are given for treatment effect at each of the four time points.*

Effect	Parameter	CC		LOCF		Observed data	
Parameter estimates (standard errors)							
Intercept 4	β_{11}	54.47	(1.54)	54.00	(1.47)	54.00	(1.47)
Intercept 12	β_{21}	53.08	(1.66)	53.03	(1.59)	53.01	(1.60)
Intercept 24	β_{31}	49.79	(1.80)	49.35	(1.72)	49.20	(1.74)
Intercept 52	β_{41}	44.43	(1.83)	44.59	(1.74)	43.99	(1.79)
Treatment effect 4	β_{12}	-2.87	(2.28)	-3.11	(2.10)	-3.11	(2.10)
Treatment effect 12	β_{22}	-2.89	(2.46)	-4.45	(2.27)	-4.54	(2.29)
Treatment effect 24	β_{32}	-3.27	(2.66)	-3.41	(2.45)	-3.60	(2.49)
Treatment effect 52	β_{42}	-4.71	(2.70)	-3.92	(2.48)	-5.18	(2.59)
p-values							
Treatment effect 4	β_{12}		0.211		0.140		0.140
Treatment effect 12	β_{22}		0.241		0.051		0.048
Treatment effect 24	β_{32}		0.220		0.165		0.150
Treatment effect 52	β_{42}		0.083		0.115		0.046

will be considered. For the observed, partially incomplete data, GEE is supplemented with WGEE and MI-GEE. Results of the GEE analyses are reported in Table 9.2. In all cases, we use the logit link, and the model takes the form

$$\text{logit}[P(Y_{ij} = 1 \mid T_i, t_j)] = \beta_{j1} + \beta_{j2} T_i, \tag{9.1}$$

with notational conventions as before, except that Y_{ij} is the indicator for whether or not letters of vision have been lost for subject i at time j, relative to baseline.

A working exchangeable correlation matrix is considered. For the WGEE analysis,

Table 9.2: *Age-related macular degeneration trial. Parameter estimates (empirically corrected standard errors) for the marginal models: standard GEE on the CC and LOCF population, and on the observed data. In the latter case, standard GEE, WGEE and MI-GEE with imputation based on the continuous and binary outcome is used.*

				Observed data			
						Continuous	Binary
Effect	Parameter	CC	LOCF	Unweighted	WGEE	MI-GEE	MI-GEE
Parameter estimates (standard errors)							
Intercept 4	β_{11}	-1.01 (0.24)	-0.95 (0.21)	-0.95 (0.21)	-0.98 (0.44)	-0.95 (0.21)	-0.95 (0.21)
Intercept 12	β_{21}	-0.89 (0.24)	-0.99 (0.21)	-1.01 (0.22)	-1.77 (0.37)	-1.00 (0.22)	-0.98 (0.22)
Intercept 24	β_{31}	-1.13 (0.25)	-1.09 (0.22)	-1.07 (0.23)	-1.11 (0.33)	-1.05 (0.22)	-1.06 (0.25)
Intercept 52	β_{41}	-1.64 (0.29)	-1.46 (0.24)	-1.64 (0.29)	-1.72 (0.39)	-1.52 (0.26)	-1.57 (0.29)
Treatment 4	β_{12}	0.40 (0.32)	0.32 (0.29)	0.32 (0.29)	0.78 (0.66)	0.32 (0.29)	0.32 (0.29)
Treatment 12	β_{22}	0.49 (0.31)	0.59 (0.29)	0.62 (0.29)	1.83 (0.60)	0.60 (0.29)	0.58 (0.29)
Treatment 24	β_{32}	0.48 (0.33)	0.46 (0.29)	0.43 (0.30)	0.72 (0.53)	0.40 (0.30)	0.42 (0.32)
Treatment 52	β_{42}	0.40 (0.38)	0.32 (0.33)	0.40 (0.37)	0.72 (0.52)	0.33 (0.35)	0.31 (0.41)
Corr.	ρ	0.39	0.44	0.39	0.33	0.39	0.38
p-values							
Treatment 4	β_{12}	0.209	0.268	0.268	0.242	0.268	0.268
Treatment 12	β_{22}	0.113	0.040	0.034	0.003	0.037	0.048
Treatment 24	β_{32}	0.141	0.119	0.151	0.176	0.182	0.195
Treatment 52	β_{42}	0.283	0.323	0.277	0.162	0.349	0.456

Table 9.3: *Age-related macular degeneration trial. Parameter estimates (standard errors) and p-values for a logistic regression model to describe dropout.*

Effect	Parameter	Estimate (s.e.)	p-value
Intercept	ψ_0	0.13 (0.49)	0.7930
Previous outcome	ψ_1	0.04 (0.38)	0.9062
Treatment	ψ_2	-0.87 (0.37)	0.0185
Lesion level 1	ψ_{31}	-1.82 (0.49)	0.0002
Lesion level 2	ψ_{32}	-1.89 (0.52)	0.0003
Lesion level 3	ψ_{33}	-2.79 (0.72)	0.0001
Time 2	ψ_{41}	-1.73 (0.49)	0.0004
Time 3	ψ_{42}	-1.36 (0.44)	0.0019

the following weight model is assumed:

$$\begin{aligned}\text{logit}[P(D_i = j \mid D_i \geq j)] &= \psi_0 + \psi_1 y_{i,j-1} + \psi_2 T_i \\ &\quad + \psi_{3,1} L_{1i} + + \psi_{3,2} L_{2i} + \psi_{3,3} L_{3i} \\ &\quad + \psi_{4,1} I(t_j = 2) + \psi_{4,2} I(t_j = 3),\end{aligned}$$

where $y_{i,j-1}$ is the binary outcome at the previous time $t_{i,j-1} = t_{j-1}$, $L_{ki} = 1$ if the patient's eye lesion is of level $k = 1, \ldots, 4$ (since one dummy variable is redundant, only three are used), and $I(\cdot)$ is the indicator function. Parameter estimates, standard errors and p-values for the dropout model are given in Table 9.3. Covariates of importance are treatment assignment, the level of lesions at baseline, and time at which dropout occurs. For the latter covariates, there are three levels, since dropout can occur at times 2, 3, or 4. Hence, two indicator variables are included. Finally, the previous outcome does not have a significant impact, but will be kept in the model nevertheless.

When comparing parameter estimates across CC, LOCF, and observed data analyses, it is clear that LOCF has the effect of artificially increasing the correlation between measurements. The effect is mild in this case. The parameter estimates of the observed-data GEE are close to the LOCF results for earlier time points and close to CC for later time points. This is to be expected, as at the start of the study the LOCF and observed populations are virtually the same, with the same holding between CC and observed populations near the end of the study. Note also that

the treatment effect under LOCF, especially at 12 weeks and after 1 year, is biased downward in comparison to the GEE analyses. Next, to properly use the information in the missingness process, WGEE or MI-GEE can be used. Two versions of MI-GEE are considered, that is, first the continuous outcome defined by the difference in numbers of letters correctly read compared with baseline is imputed whereafter the dichotomized version is analysed, and secondly the binary outcome is imputed and analysed.

In spite of there being no strong evidence for MAR, the results between GEE and WGEE differ quite a bit. It is noteworthy that at 12 weeks, a treatment effect is observed with WGEE which is borderline with the other marginal analyses. However, as we have shown in Chapter 5, the beneficial property of unbiasedness for WGEE is merely fulfilled for very large samples. On the other hand, MI-GEE produces only a small amount of bias in small samples, which is less compared to the bias of WGEE. Moreover, MI-GEE is robustness against misspecification of the imputation and measurement model. Therefore, in real life settings such as the age-related macular degeneration trial, we would opt to use MI-GEE instead of WGEE. Both versions of MI-GEE considered here show quite similar results. The standard errors are smaller compared to the ones estimated using WGEE. The treatment effect at week 12 detected by WGEE becomes borderline again in both MI-GEE analyses. Further, also the p-values at later time points are larger compared to WGEE.

9.2.2 Random-Effects Models

Let us now turn to a random-intercepts logistic model, in spirit to (9.1):

$$\text{logit}[P(Y_{ij} = 1 \mid T_i, t_j)] = \beta_{j1} + b_i + \beta_{j2} T_i, \tag{9.2}$$

with notation as before and $b_i \sim N(0, \tau^2)$. For the model fitting, numerical integration is used. Results are shown in Table 9.4.

We observe the usual relationship between the marginal parameters of Table 9.2 and their random-effects counterparts. Note also that the random-intercepts variance is largest under LOCF, underscoring again that this method artificially increases the association between measurements on the same subject. In this case, in contrast to the marginal models, both LOCF and CC, considerably overestimate the treatment effect at certain times, in particular at 4 and 24 weeks (unlike the continuous case; see Section 9.1. In conclusion, it is clear that CC and LOCF may differ from the direct-likelihood and weighted or MI-based GEE analyses. This underscores, once again, that the latter analyses are to be considered as candidates for primary analysis.

Table 9.4: *Age-related macular degeneration trial. Parameter estimates (standard errors) for the random-intercept models: numerical-integration based fits on the CC and LOCF population, and on the observed data (direct-likelihood).*

Effect	Parameter	CC		LOCF		Direct-lik.	
Intercept 4	β_{11}	-1.73	(0.42)	-1.76	(0.40)	-1.64	(0.37)
Intercept 12	β_{21}	-1.53	(0.41)	-1.85	(0.40)	-1.75	(0.38)
Intercept 24	β_{31}	-1.93	(0.43)	-2.01	(0.41)	-1.85	(0.39)
Intercept 52	β_{41}	-2.74	(0.48)	-2.66	(0.44)	-2.76	(0.47)
Treatment 4	β_{12}	0.64	(0.54)	0.55	(0.53)	0.51	(0.49)
Treatment 12	β_{22}	0.81	(0.53)	1.04	(0.53)	1.02	(0.50)
Treatment 24	β_{32}	0.77	(0.55)	0.80	(0.53)	0.70	(0.51)
Treatment 52	β_{42}	0.60	(0.59)	0.54	(0.56)	0.61	(0.59)
Random-intercept s.d.	τ	2.19	(0.27)	2.46	(0.27)	2.21	(0.26)
Random-intercept var.	τ^2	4.80	(1.17)	6.03	(1.33)	4.90	(1.14)

9.3 Sensitivity Analysis of the Continuous Outcome

In this section, we apply MNAR-based and sensitivity methods to the continuous outcome of the age-related macular degeneration trial. In Section 9.3.1 the Diggle-Kenward selection model (Chapter 6) is fitted to the response sequences and by way of sensitivity analysis this is supplemented with local influence analysis (Chapter 7). Further, in Section 9.3.2, the focus is on pattern-mixture models, as introduced in Chapter 6. Finally, latent-class mixture models as proposed in Chapter 8 are considered in Section 9.3.3.

9.3.1 Selection Models and Local Influence

In this section, the visual acuity is first analysed using the full selection model proposed by Diggle and Kenward (1994), discussed in Section 6.1.1. Apart from modeling the three missing data mechanisms MCAR, MAR, and MNAR, explicitly, an ignorable MAR analysis is also conducted in which the model for the response measurements only was fitted. For the measurement model, the linear mixed model was used, assuming again different intercepts and treatment effects for each of the four time points,

Table 9.5: *Age-related macular degeneration trial. Parameter estimates (standard errors) assuming ignorability, as well as explicitly modeling the missing data mechanism under MCAR, MAR, and MNAR assumptions, for all data.*

All Subjects		Ignorable		MCAR		MAR		MNAR	
Effect	Parameters	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Measurement Model									
Intercept 4	β_{11}	54.00	(1.47)	54.00	(1.46)	54.00	(1.47)	54.00	(1.47)
Intercept 12	β_{21}	53.01	(1.60)	53.01	(1.59)	53.01	(1.60)	52.98	(1.60)
Intercept 24	β_{31}	49.20	(1.74)	49.20	(1.73)	49.19	(1.74)	49.06	(1.74)
Intercept 52	β_{41}	43.99	(1.79)	43.99	(1.78)	43.99	(1.79)	43.52	(1.82)
Treatment 4	β_{12}	-3.11	(2.10)	-3.11	(2.07)	-3.11	(2.09)	-3.11	(2.10)
Treatment 12	β_{22}	-4.54	(2.29)	-4.54	(2.25)	-4.54	(2.29)	-4.67	(2.29)
Treatment 24	β_{32}	-3.60	(2.49)	-3.60	(2.46)	-3.60	(2.50)	-3.80	(2.50)
Treatment 52	β_{42}	-5.18	(2.59)	-5.18	(2.57)	-5.18	(2.62)	-5.71	(2.63)
Dropout Model									
Intercept	ψ_0			-2.79	(0.17)	-1.86	(0.46)	-1.81	(0.47)
Previous	ψ_1					-0.020	(0.009)	0.016	(0.022)
Current	ψ_2							-0.042	(0.023)
-2 log-likelihood		6488.7		6782.7		6778.4		6775.9	
Treatment effect at 1 year	p-value	0.046		0.044		0.048		0.030	

next to an unstructured variance covariance matrix. In the full selection models, the dropout is modeled by (6.4). Parameter estimates and corresponding standard errors of the fixed effects of the measurement model and of the dropout model parameters are given in Table 9.5. As expected, the parameter estimates and standard errors coincide for the ignorable direct-likelihood analysis and the selection models under MCAR and MAR, except for some numerical noise.

Since main interest lies in the treatment effect at 1 year, the corresponding p-values are displayed in Table 9.5. In all four cases, this treatment effect is (borderline) significant.

Note that for the MNAR analysis, the estimates of the ψ_1 and ψ_2 parameter are more or less of the same magnitude, but with a different sign. This is in line with the argument of Molenberghs *et al.* (2001b), stating that the dropout oftentimes depends on the increment $y_{ij} - y_{i,j-1}$. This is because two subsequent measurements are usually positively correlated. By rewriting the fitted dropout model in terms of the increment,

$$\text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})\right] = -1.81 - 0.026 y_{i,j-1} - 0.042(y_{ij} - y_{i,j-1}),$$

We find that the probability of dropout increases with larger negative increments; that is, those patients who showed or would have shown a greater decrease in visual acuity from the previous visit are more likely to drop out.

Let us now switch to local influence. Figure 9.1 displays overall C_i and influences for subvectors $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\psi}$. In addition, the direction $\boldsymbol{h}_{\max}$, corresponding to maximal local influence, is given. The main emphasis should be put on the relative magnitudes. We observe that patients #10, #27, #28, #114, #139, and #154 have larger C_i values compared to other patients, which means they can be considered influential. Virtually the same picture holds for $C_i(\boldsymbol{\psi})$.

Turning attention to the influence on the measurement model, we see that for $C_i(\boldsymbol{\beta})$, there are no strikingly high peaks, whereas $C_i(\boldsymbol{\alpha})$ reveals two considerable peak for patients #68 and #185. Note that both patients fail to have a high peak for the overall C_i. This is due to the fact that the scale for $C_i(\boldsymbol{\alpha})$ is relatively small, comparing to the overall C_i. Nevertheless, these patients can still be considered influential. Finally, the direction of maximum curvature reveals the same six influential patients as the overall C_i.

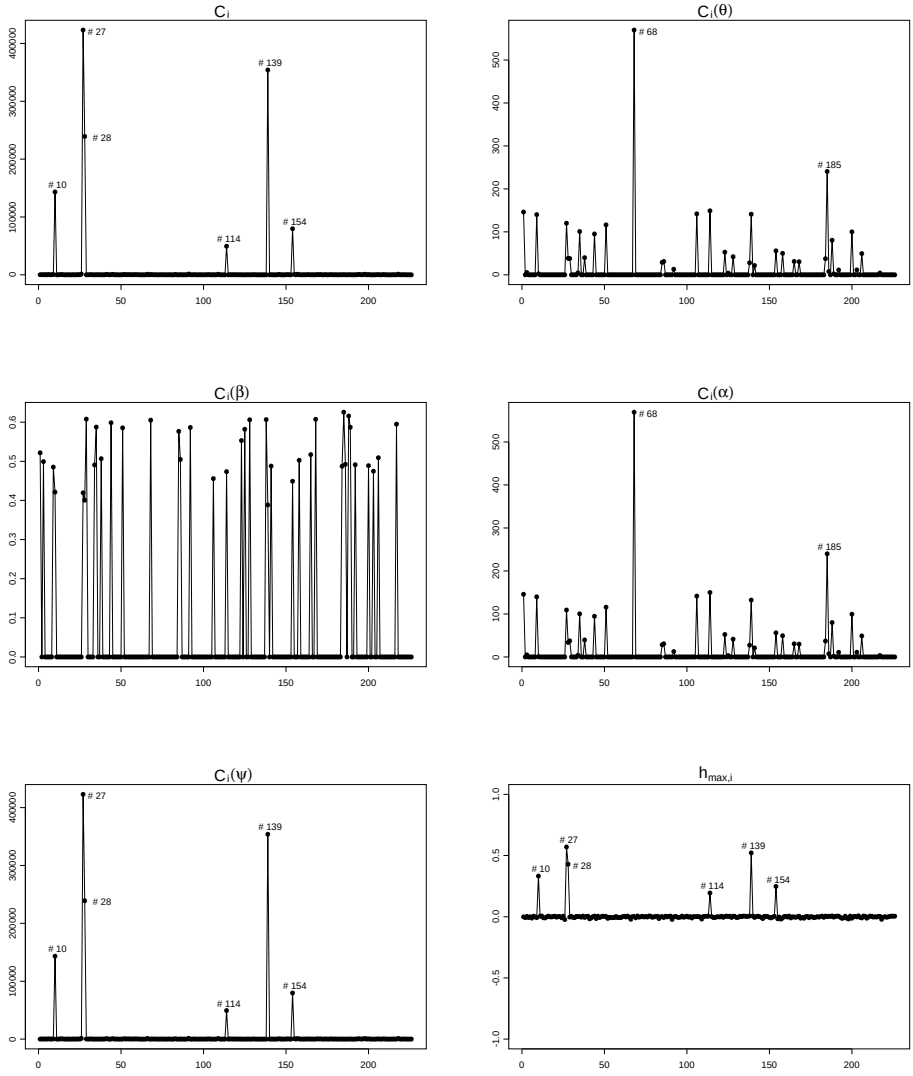


Figure 9.1: *Age-related macular degeneration trial. Index plots of* C_i, $C_i(\boldsymbol{\theta})$, $C_i(\boldsymbol{\alpha})$, $C_i(\boldsymbol{\beta})$, $C_i(\boldsymbol{\psi})$ *and of the components of the direction* $\boldsymbol{h}_{\max,i}$ *of maximal curvature.*

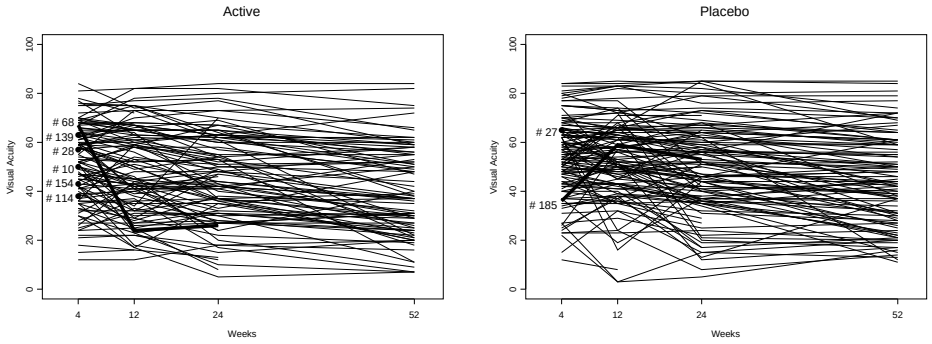


Figure 9.2: *Age-related macular degeneration trial. Individual profiles for both treatment arms, with influential subjects highlighted.*

In Figure 9.2, the individual profiles of the influential observations are highlighted. Let us take a closer look at these cases. The six patients, which are influencing the dropout model parameters, are those that drop out after the first measurement is taken at week 4. All of these patients are in the active treatment arm, except for #27. On the other hand, the two patients influential for the measurement model parameters, stay in the study up to week 24 and have no observation for the last measurement occasion at 1 year. Patient # 68 received the active treatment, and his/her visual acuity decreases substantially after week 4, thereafter staying more or less level. Opposite, patient #185 is enrolled in the the placebo treatment arm and his/her visual acuity increases after week 4, then sloping down a little after week 12.

It is interesting to consider an analysis without these influential observations. Therefore, we applied the selection model to three subsets of the data. The first subset was obtained by removing all the eight influential patients mentioned before. In the second subset of the data, patients #10, #27, #28, #114, #139, and #154 were removed, since these are overall the most influential ones. Finally, patients #68 and #185, who seemed to be influencing the measurement model the most, were removed, resulting in the third subset. Result of these analyses are shown in Table 9.6. We compare the results of the MAR and MNAR analyses.

Table 9.6: *Age-related macular degeneration trial. Parameter estimates (standard errors) explicitly modeling the missing data mechanism under MAR and MNAR assumptions, after removing the following subsets of subjects 10, 27, 28, 114, 139, 154, 68, 185 (Set 1); 10, 27, 28, 114, 139, 154 (Set 2); and 68, 185 (Set 3).*

Subjects Removed		Set 1		Set 2		Set 3	
		MAR	MNAR	MAR	MNAR	MAR	MNAR
Effect	Parameter	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)
Measurement Model							
Intercept 4	β_{11}	54.14 (1.51)	54.15 (1.49)	54.30 (1.47)	54.30 (1.46)	53.84 (1.48)	53.84 (1.47)
Intercept 12	β_{21}	53.09 (1.64)	53.06 (1.62)	53.16 (1.59)	53.13 (1.59)	52.94 (1.60)	52.91 (1.59)
Intercept 24	β_{31}	49.56 (1.77)	49.46 (1.75)	49.31 (1.74)	49.20 (1.72)	49.44 (1.73)	49.31 (1.72)
Intercept 52	β_{41}	44.40 (1.82)	43.97 (1.84)	44.00 (1.79)	43.58 (1.82)	44.38 (1.78)	43.90 (1.82)
Treatment 4	β_{12}	-3.13 (2.17)	-3.13 (2.11)	-3.28 (2.08)	-3.28 (2.06)	-2.95 (2.07)	-2.95 (2.05)
Treatment 12	β_{22}	-4.48 (2.36)	-4.63 (2.29)	-4.55 (2.26)	-4.69 (2.24)	-4.47 (2.26)	-4.60 (2.23)
Treatment 24	β_{32}	-3.80 (2.56)	-4.04 (2.49)	-3.55 (2.48)	-3.79 (2.44)	-3.85 (2.44)	-4.04 (2.42)
Treatment 52	β_{42}	-5.45 (2.66)	-6.12 (2.66)	-5.06 (2.59)	-5.72 (2.61)	-5.56 (2.55)	-6.09 (2.58)
Dropout Model							
Intercept	ψ_0	-1.90 (0.47)	-1.85 (0.49)	-1.90 (0.47)	-1.85 (0.49)	-1.85 (0.46)	-1.81 (0.47)
Previous	ψ_1	-0.019 (0.010)	0.018 (0.022)	-0.019 (0.010)	0.017 (0.022)	-0.020 (0.009)	0.017 (0.022)
Current	ψ_2		-0.044 (0.024)		-0.043 (0.024)		-0.043 (0.024)
-2 log-likelihood		6535.3	6532.7	6606.9	6604.4	6706.4	6703.8
Treatment at 1 year	p-value	0.040	0.021	0.051	0.028	0.029	0.018

After removing all influential patients (Set 1), the estimates of the dropout model parameters ψ_1 and ψ_2 are approximately the same, whereas the estimate of ψ_0 decreases from -1.86 to -1.90 under MAR, and from -1.81 to -1.85 under MNAR. The same can be seen after removing the patients #10, #27, #28, # 114, #139, and #154, who have large overall C_i and $C(\boldsymbol{\psi})$ values (Set 2). Considering the treatment effect at 1 year, its estimate decreases from -5.18 to -5.45 under the MAR assumption, and from -5.71 to -6.12 under the MNAR assumption, resulting in a decrease of the p-value from 0.048 to 0.040 and from 0.030 to 0.021 under MAR and MNAR respectively.

There is no impact on the likelihood ratio test for MAR against MNAR after removing all influential patients, the deviance G^2 only changes slightly from 2.5 to 2.6. If this likelihood ratio test would follow a standard χ^2_1-distribution, we would fail to reject the null hypothesis, which leads us to the MAR assumption. However, the test of MAR against MNAR is non-standard and it cannot be used as such (Rotnitzky *et al.*, 2000; Jansen *et al.*, 2006b). Moreover, recall we have shown in Chapter 6 that one can never test for the assumption of MNAR versus MAR missingness.

Further, after removing the second set of influential patients, that is, patients #10, #27, #28, # 114, #139, the estimate of the treatment effect at 1 year increases from -5.18 to -5.06 under the MAR analysis, yielding a slightly increased borderline p-value, whereas it decreases with 0.01 under the MNAR analysis and together with a decreased standard error the latter yields a small decrease in the p-value. The deviance G^2 for the likelihood ratio test for MAR against MNAR remains 2.5

Finally, we perform the same analyses on the third set, with patients #68 and # 185 removed. Both for the MAR and MNAR analysis, again the estimate of the treatment effect at 1 year decreases quite a lot, from -5.18 to -5.56 and from -5.71 to -6.09 respectively. Consequently, the p-value also drops down from 0.048 to 0.029 under MAR and from 0.030 to 0.018 under the MNAR analysis. The deviance for the likelihood ratio test for MAR changes again from 2.5 to 2.6.

9.3.2 Pattern-Mixture Models

Pattern-mixture models can be of use in the context of sensitivity analysis. Given there are several, quite distinct, strategies to formulate such models, one can consider one strategy as a sensitivity analysis for another one. For example, the sensitivity of simple, identified models can be checked using identifying restrictions. Also, a set of identifying restrictions can be considered, rather than a single one, by way of sensitivity analysis.

Table 9.7: *Age-related macular degeneration trial. Parameter estimates (standard errors) and p-values resulting from the pattern-mixture model using identifying restrictions ACMV, CCMV, and NCMV.*

		ACMV		CCMV		NCMV	
Effect	Parameter						
Parameter estimate (standard error)							
Intercept 4	β_{11}	54.00	(1.47)	54.00	(1.47)	54.00	(1.47)
Intercept 12	β_{21}	53.22	(1.98)	52.89	(1.61)	52.97	(2.20)
Intercept 24	β_{31}	49.43	(2.14)	49.45	(1.79)	49.05	(2.49)
Intercept 52	β_{41}	44.73	(2.69)	44.67	(2.35)	44.40	(2.73)
Treatment 4	β_{12}	-3.11	(2.10)	-3.11	(2.10)	-3.11	(2.10)
Treatment 12	β_{22}	-4.94	(2.81)	-4.26	(2.36)	-4.56	(2.71)
Treatment 24	β_{32}	-4.21	(2.82)	-3.77	(2.55)	-3.79	(2.92)
Treatment 52	β_{42}	-5.19	(2.81)	-4.72	(2.60)	-4.76	(2.90)
p-value							
Intercept 4	β_{11}	$< .0001$		$< .0001$		$< .0001$	
Intercept 12	β_{21}	$< .0001$		$< .0001$		$< .0001$	
Intercept 24	β_{31}	$< .0001$		$< .0001$		$< .0001$	
Intercept 52	β_{41}	$< .0001$		$< .0001$		$< .0001$	
Treatment 4	β_{12}	0.140		0.140		0.140	
Treatment 12	β_{22}	0.083		0.071		0.093	
Treatment 24	β_{32}	0.139		0.140		0.200	
Treatment 52	β_{42}	0.065		0.069		0.101	

Obviously, one can formulate selection models for one's primary analysis, and then fit pattern-mixture models to assess sensitivity. Michiels *et al.* (2002) followed this route. Molenberghs, Michiels and Kenward (1998a) formulated models that combine aspects of both selection models and pattern-mixture models, and used pseudo-likelihood ideas to fit such models.

Table 9.8: *Age-related macular degeneration trial. p-values resulting from the pattern-mixture model using identifying restrictions ACMV, after removing the three subsets as in Table 9.6.*

		Set 1		Set 2		Set 3	
Effect	Parameter						
Parameter estimate (standard error)							
Treatment 4	β_{12}	-3.11	(2.10)	-3.11	(2.10)	-3.11	(2.10)
Treatment 12	β_{22}	-4.94	(2.81)	-4.26	(2.36)	-4.56	(2.71)
Treatment 24	β_{32}	-4.21	(2.82)	-3.77	(2.55)	-3.79	(2.92)
Treatment 52	β_{42}	-5.19	(2.81)	-4.72	(2.60)	-4.76	(2.90)
p-value							
Treatment 4	β_{12}	0.140		0.140		0.140	
Treatment 12	β_{22}	0.083		0.071		0.093	
Treatment 24	β_{32}	0.139		0.140		0.200	
Treatment 52	β_{42}	0.065		0.069		0.101	

In this section, we consider the use of pattern-mixture models for the visual acuity outcome. Based on the discussion in Section 6.2, we will apply the Strategy 1 making use of CCMV, NCMV, and ACMV identification restrictions. After applying each of the three restrictions, the same selection model as before is fitted. The results for the three types of restrictions are shown in Table 9.7. From the estimates and associated standard errors, it is clear that there is little difference in conclusions between the strategies. The estimates for treatment effect and corresponding standard errors obtained under CCMV and NCMV restrictions are underestimated when comparing to ACMV.

In Table 9.7 we observe that for all three strategies the p-value for the treatment effect at 1 year is above the significance level of 0.05 significant, yet it is borderline significant for the ACMV restrictions - which is equivalent to MAR - in line with the conclusions drawn from the selection models in previous section. The p-value is closest to significance and thus to the one from the selection models in case the NCMV restrictions are considered. The Diggle-Kenward MAR and MNAR selection

models also showed a borderline significant treatment effect, its p-value being below 0.05.

The local influence approach applied to the selection models in previous section determined different subsets of influential patients. Fitting pattern-mixture models based on ACMV restrictions after excluding each of these subsets in turn, changes the p-value for treatment effect after 1 year from 0.065 to 0.053, 0.069, and 0.049 respectively, all maintaining the borderline significance.

One will feel comfortable about a significant treatment effect if it held across MAR and a number of MNAR scenarios. Thus, in this case, it is fair to say there is a weak evidence only for a treatment effect.

9.3.3 Latent-Class Mixture Models

In this section we use the latent-class mixture model framework to analyse the visual acuity outcome. First a model building exercise is performed to obtain a latent-class mixture model that is fitted to the data, after which it is used as a classification as well as a sensitivity analysis tool (Section 9.3.4).

Model Building

The heterogeneity linear mixed model considered for the visual acuity is the following:

$$Y_{ij}|q_{i,k}, b_i = \beta_0 + \beta_{1,1} I(t_j = 1) + \beta_{1,2} I(t_j = 2) + \beta_{1,3} I(t_j = 3) \; + \; \beta_{2,j} T_i + b_i,$$

which represents the same full treatment by time interaction with eight mean model parameters as in Section 9.1, yet using a different parametrization for the intercepts, added with a shared intercept b_i. Depending on whether a shared-parameter model or a latent-class mixture model is fitted, the distribution of the shared intercept is the univariate normal distribution, that is, $b_i \sim N(0, d^2)$, or a mixture of g normal distributions with different means but equal variances, that is, $b_i \sim \sum_{k=1}^{g} \pi_k N(\mu_k, d^2)$. The measurement error terms are assumed to be independent and to follow a normal distribution with mean 0 and variance σ^2. The assumed latent-class mixture model first takes the same structure as in Section 8.4, that is,

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_{0,k}, \tag{9.3}$$

after which the shared intercept is added, resulting in

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_{0,k} + \lambda \, b_i. \tag{9.4}$$

Table 9.9: *Age-related macular degeneration trial. Information criteria AIC and BIC, for models with dropout model (9.3) or (9.4), and* $g = 1, 2, 3$.

Model	Dropout Model	g	# Par	-2ℓ	AIC	BIC
1	$\gamma_{0,k}$	1	11	7103.66	7125.66	7163.29
2	$\gamma_{0,k}$	2	14	7049.27	7077.27	7125.15
3	$\gamma_{0,k}$	3	17	6999.89	**7033.89**	**7092.04**
4	$\gamma_{0,k} + \lambda\, b_i$	1	12	7102.32	7124.32	7161.95
5	$\gamma_{0,k} + \lambda\, b_i$	2	15	7046.11	7076.11	7127.42
6	$\gamma_{0,k} + \lambda\, b_i$	3	18	6999.92	7035.92	7097.49

An overview of the models considered is given in Table 9.9. The model building exercise is performed in a forward stepwise selection way as was done in Section 8.5.1 for the first depression trial. Both AIC and BIC criteria opt for the latent-class mixture model with three latent subgroups and no shared intercept in the dropout model. Models for which $g > 3$ could not be fitted, since the variance of the mixture components, d^2, tends to zero as the number of latent subgroups increases.

Parameter estimates with corresponding standard errors and p-values of this latent-class mixture model are depicted in Table 9.10. Clearly, the treatment effect at 1 year is significant.

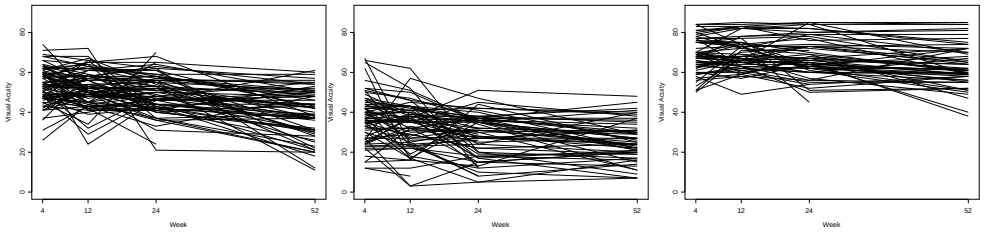
Classification of Subjects

Let us now use posterior probabilities obtained from the fit of the latent-class mixture model to classify the subjects into three subgroups as shown in Section 8.3. The 226 subjects divide into 88, 72, and 66 subjects classified into the first, second, and third group respectively.

In Figure 9.3 the individual profiles of subjects classified into the three latent groups are represented. Clearly, the last group contains subjects with highest visual acuity measurements which remain more or less level throughout the study, whereas subjects classified to the second group have the lowest observations which even decrease during the first 24 weeks. The profiles of subjects of the second group lie in between the ones of the second and third group, which clearly diminish over time. This association between the visual acuity values and the classification in subgroups can be formally confirmed.

Table 9.10: *Age-related macular degeneration trial. Parameter estimates, standard errors, and p-values for the latent-class mixture model with three latent subgroups and dropout model (9.3).*

Effect	Parameter	Estimate	s.e.	p-value
Measurement Model				
Intercept 52	β_0	44.81	1.67	< 0.0001
Intercept 4 – 52	$\beta_{1,1}$	9.74	1.39	< 0.0001
Intercept 12 – 52	$\beta_{1,2}$	8.73	1.39	< 0.0001
Intercept 24 – 52	$\beta_{1,3}$	5.05	1.41	0.0003
Treatment 4	$\beta_{2,1}$	-4.21	2.37	0.0764
Treatment 12	$\beta_{2,2}$	-5.66	2.29	0.0134
Treatment 24	$\beta_{2,3}$	-4.81	2.40	0.0452
Treatment 52	$\beta_{2,4}$	-6.38	2.40	0.0077
Measurement error	σ	10.08	0.28	< 0.0001
Dropout Model				
Intercept Group 1	$\gamma_{0,1}$	-2.62	0.30	< 0.0001
Intercept Group 2	$\gamma_{0,2}$	-2.73	0.33	< 0.0001
Intercept Group 3	$\gamma_{0,3}$	-3.17	0.49	< 0.0001
Shared Effects				
Mean Group 1	μ_1	0.29	1.28	0.8236
Mean Group 2	μ_2	-17.28	0.95	< 0.0001
Variance	d	0.77	1.05	0.4621
Prior probability Group 1	π_1	0.39	0.05	< 0.0001
Prior probability Group 2	π_2	0.31	0.03	< 0.0001
Loglikelihood	ℓ	-3499.9446		



(a) Group 1 (b) Group 2 (c) Group 3

Figure 9.3: *Age-related macular degeneration trial. Classification of the subjects into the three latent subgroups based on a latent-class mixture model.*

Further, this difference between the latent groups is not due to treatment, since the classification of subjects into latent groups is independent of their treatment allocation.

Turning attention to missingness in the three latent groups, it is observed that the dropout percentage decreases from the first to the third subgroup. In the first group 18 out of 88 subjects drop out (20.45%), 12 out of 72 (16.67%) in the second group, and finally 8 out of 66 in the last group (12.12%).

Table 9.11: *Age-related macular degeneration trial. Classification of subjects based on the magnitude of posterior probabilities* π_{ik}.

Decision Rule	Classification	# Patients
$0.80 < \pi_{i1} \leq 1.00$	Clearly Group 1	65
$0.80 < \pi_{i2} \leq 1.00$	Clearly Group 2	58
$0.80 < \pi_{i3} \leq 1.00$	Clearly Group 3	54
$0.60 < \pi_{i1} \leq 0.80$	Group 1	15
$0.60 < \pi_{i2} \leq 0.80$	Group 2	11
$0.60 < \pi_{i3} \leq 0.80$	Group 3	7
$0.55 < \pi_{i1} \leq 0.60$	Doubtful, more likely Group 1	4
$0.55 < \pi_{i2} \leq 0.60$	Doubtful, more likely Group 2	2
$0.55 < \pi_{i3} \leq 0.60$	Doubtful, more likely Group 3	0
$0.45 < \pi_{ik} \leq 0.55$	Uncertain	10

Note again that one should be cautious in over-interpreting this classification in latent groups. As mentioned in Section 8.3 one can consider the magnitude of the maximal posterior probability, which shows insight into how strongly the subjects are allocated to these subgroups. Based on the three posterior probabilities, π_{i1}, π_{i2}, and π_{i3}, the subjects can be classified following the guidelines of Table 9.11. For only 10 out of 226 subjects, that is 4.4%, it is uncertain whether the classification to the particular latent subgroup is appropriate. On the other hand, for most subjects, 210 or 92.9%, the maximal posterior probability is above 0.60 meaning it is clear to which latent group is should be classified.

9.3.4 Sensitivity Analysis

Next to the CC, LOCF and ignorable direct-likelihood analyses as discussed in Section 9.1, the age-related macular degeneration trial has now been reanalysed using (1) MCAR, MAR, and MNAR selection models, (2) the local influence sensitivity tool, (3) pattern-mixture models, and (4) latent-class mixture models. To assess the sensitivity of the modeling assumptions on the conclusions, let us consider the treatment effect at 1 year obtained by the different methods. A comparison of the estimates, standard errors and p-values is provided in Table 9.12. Whereas the pattern-mixture model shows a borderline (in)significant treatment effect at 1 year, it moves to borderline significant for the ignorable direct-likelihood analysis, which assumes MAR missingness, and the MAR Diggle-Kenward selection model, and its significance becomes more prominent in turning to the MNAR Diggle-Kenward selection model and the latent-class mixture model. In conclusion, the primary analysis, which ideally would be based on a model assuming MAR non-response, for instance an ignorable direct-likelihood analysis as in this case, yields a borderline significant treatment effect at 1 year. To assess the sensitivity thereof, this primary analysis is extended with other MAR models under different modeling frameworks which clearly confirms this borderline significance. Expanding this sensitivity analysis by turning attention to MNAR and latent-class mixture models shows a decreased p-value thereby concluding a significant treatment effect at 1 year. Thus, a cautious conclusion is that there is some evidence for a treatment effect at the end of the study.

Table 9.12: *Age-related macular degeneration trial. Estimates, standard errors, and p-values for the treatment effect at 1 year for the latent-class mixture model, the ignorable direct-likelihood analysis, both Diggle-Kenward selection models assuming either MAR or MNAR, as well as the pattern-mixture model using ACMV restrictions.*

	Treatment at 1 year		
Model	Estimate	s.e.	p-value
Latent-Class Mixture Model	-6.38	2.40	0.008
Ignorable Direct-Likelihood	-5.18	2.59	0.047
Pattern-Mixture Model (ACMV)	-5.19	2.81	0.065
MAR Selection Model	-5.18	2.62	0.048
MNAR Selection Model	-5.71	2.63	0.030

10

Conclusion

After the key paper of Rubin (1976) who established incomplete data as a field of study within the domain of statistics, a large amount of research output has been devoted to the problem of missing data. Within the methodological development we can distinguish between the parametric school, based on the likelihood and Bayesian frameworks, and a semi-parametric school, including estimating equations ideas. Even though there is a noticeable divergence between these various lines of thinking, researchers agree that no single modeling approach can overcome the limitation of not having access to the missing data. All parties, that is, academia, industry, and regulatory authorities, emphasize the need for sensitivity analysis, whereas there is less agreement on the kind of sensitivity analysis. An important condition to put forward a particular method as a feasible method within a sensitivity analysis, is the availability of trustworthy and easy-to-use software.

In this thesis, we have shown it is unfortunate that there has been so much emphasis on simple methods, such as complete case analysis or last observation carried forward, which at least require the missingness mechanism to be MCAR. These simple methods have been compared to direct-likelihood analysis, which uses all available information without the need of additional data manipulation and are valid under the less restrictive and more realistic assumption of MAR missingness. Moreover, in case inferences are obtained within the likelihood or Bayesian framework, there is no need to model the missingness process. Consequently, linear mixed models (Verbeke and

Molenberghs, 2000) or generalized linear mixed models (Molenberghs and Verbeke, 2005), within the random-effects model family, can be used for respectively Gaussian and non-Gaussian incomplete longitudinal outcomes. These methods are as simple to conduct as it would be in contexts where data are complete.

In the non-Gaussian case one often opts for a semi-parametric approach such as generalized estimating equations within the marginal model family. Since this method also requires the missingness to be MCAR, alternatives have been proposed such as weighted generalized estimating equations (WGEE, Robins, Rotnitzky and Zhao, 1994) and multiple imputation based generalized estimating equations (MI-GEE, Schafer, 2003), to obtain valid inferences under the MAR assumption. Both methods require only a little amount of programming which can be done using standard statistical software. In this thesis, we have compared both methods using asymptotic and small-sample simulation studies. Theoretically, WGEE is unbiased, which was confirmed by the asymptotic simulation study, yet this cannot be drawn along to small samples, even when every aspect of the analysis is correctly specified. Furthermore, the asymptotic unbiasedness vanishes in case of misspecification in either the missingness or measurement model. On the other hand, MI-GEE proves to be robust under misspecification of either the imputation or measurement model. Moreover, MI-GEE provides less biased and more precise estimates in small to moderate samples compared to WGEE. Consequently, we advice to use MI-GEE in practice above WGEE, despite the asymptotic unbiasedness property of WGEE. Note that, although the focus of this thesis is on missingness in the response, missingness in covariates is often encountered, in which cases MI-GEE can be used whereas WGEE cannot.

Up to here, it has been made clear that the simple ad hoc methods, which have been in common use for a long time, actually belong in the museum of statistics, and the primary analysis should consist of methods which assume the missing data to be MAR. On the other hand though, one can hardly ever rule out the possibility of missing data to be MNAR, which implies that the need may exist to consider MNAR models. Therefore, we have provided an overview of existing MNAR models, with the main focus on the models proposed by Diggle and Kenward (1994) for Gaussian outcomes and by Baker, Rosenberger and DerSimonian (BRD, 1992) for binary outcomes.

An important feature of statistical modelling in the incomplete data setting is that the quality of the fit to the observed data does not render the appropriateness of the implied structure governing the unobserved data. This point is independent of the MNAR route taken, whether a parametric model or a semi-parametric approach is chosen. MNAR models are based on assumptions regarding the unobserved outcomes

which are not verifiable from the available, observed data. Moreover, in this thesis we have proven that the empirical distinction between MNAR and MAR is not possible, in the sense that the fit of each MNAR model to a set of observed data can be reproduced exactly by an MAR counterpart. This so-called MAR bodyguard produces the same fit to the observed data, yet the predictions of the unobserved outcomes given the observed ones will be different. Consequently, unless one is prepared to accept the posited MNAR model in an unquestionable way, one can never test the assumption of an MNAR model for or against MAR. This underlines that the conclusions drawn based on MNAR models are sensitive to the posited and unverifiable model assumptions. Based on these considerations and facts, it is clear that in any incomplete-data setting there cannot be anything that could be called a definitive analysis. A sensible compromise between blindly shifting to MNAR models or ignoring them altogether, is to make them a component of a sensitivity analysis. For instance, after performing a primary analysis based on the MAR assumption, it is advisable to conduct a sensitivity analysis to explore the impact of deviations from this MAR assumption. In this thesis, a review of sensitivity analysis tools both at the level of the model by considering a variety of models, and at the level of the individuals based on global and local influence is given, which are applied to the Diggle-Kenward model and to the BRD model family. In the latter, we extended the local influence approach of Jansen *et al.* (2003) by basing its terminology on cell counts rather than parameters, as well as by perturbing the cell probabilities rather than the model parameters. Although the basis of local influence was to detect influential subjects which drop out non-randomly and thereby seemingly drive the posited MAR selection model in the direction of MNAR, several authors (Verbeke *et al.*, 2001b; Jansen *et al.*, 2006b) have shown that the influential subjects often are influential for other than missingness-related features.

A further route for sensitivity analysis is to consider pattern-mixture models as a complement to selection models (Molenberghs *et al.*, 1998b; Thijs *et al.*, 2002), or as shown in this thesis, use so-called latent-class mixture models. These models are an extension of the shared-parameter model, since both the measurement and dropout processes are allowed to share a set of random effects, conditional upon which both processes are assumed to be independent. Moreover, this model shared features of three different modeling frameworks, by using information from the location and evolution of the response profiles, a selection model concept, and from the dropout patterns, a pattern-mixture idea, to define latent groups and variables, a shared-parameter feature. Through the assumed latent-class structure, the model captures possible heterogeneity between possible latent subgroups of the population. We have

shown that this latent-class mixture model not only can be used as a flexible modeling technique, but that it also serves as a sensitivity analysis tool and it can be applied for further exploration of the latent class membership.

As a final remark, we note that all proposed models can be implemented in standard statistical software. As a final chapter, we show how the MAR-based analyses as well as the analysis based on the Diggle-Kenward model and the local influence approach applied to it, can be conducted using the SAS software. Further, the implementation of a simplified version of the latent-class mixture model as used throughout this thesis is also exemplified in the final chapter. Remark that further investigation of this model is still necessary, for instance, regarding model selection and modeling assumptions, as well as its implementation for more general settings.

11

Software

In this chapter, software implementations are presented for the direct-likelihood, GEE, WGEE and MI-GEE methods – as well as their CC and LOCF counterparts – and for the MNAR Diggle-Kenward selection model, local influence applied to the Diggle-Kenward model, and finally for the latent-class mixture model. In Section 11.1, complete case analysis and last observation carried forward analysis are shown. Next, MAR-based methods are discussed, including direct-likelihood (Section 11.2), WGEE (Section 11.3) and MI-GEE (Section 11.4). Section 11.5 is devoted to the implementation of the full selection MNAR Diggle-Kenward model. Next, the local influence method applied to the Diggle-Kenward model as a sensitivity analysis tool is dealt with in Section 11.6. Finally, Section 11.7 shows the implementation of the flexible latent-class mixture model.

11.1 Simple Analyses

To perform a complete case analysis, subjects for which not all designed measurement have been obtained need to be deleted. When the data are organised 'horizontally' – one record per subject – this is particularly easy. With 'vertically' organized data, some data manipulation is needed and this can be done using the SAS macro `%cc` given below.

```
%macro cc(data=,id=,time=,response=);
%if %bquote(&data)= %then %let data=&syslast;
proc freq data=&data noprint;
tables &id /out=freqsub;
tables &time / out=freqtime;
run;
proc iml;
use freqsub;
read all var {&id,count};
nsub = nrow(&id);
use freqtime;
read all var {&time,count};
ntime = nrow(&time);
use &data;
read all var {&id,&time,&response};
n =  nrow(&response);
complete = j(n,1,1);
ind = 1;
do while (ind <= nsub);
  if (&response[(ind-1)*ntime+ntime]=.) then
    complete[(ind-1)*ntime+1:(ind-1)*ntime+ntime]=0;
  ind = ind+1;
end;
create comp var {&id &time &response complete};
append;
quit;
data cc;
merge &data comp;
if complete=0 then delete;
drop complete;
run;
%mend;
```

Clearly, this macro requires four arguments. The `data=` argument is the data set to be analysed. If this is not specified, the most recent data set is used. The name of the variable in the data set which contains the identification variable is specified by `id=` whereas `time=` specifies the variable indicating the time ordering within a subject. The outcome variable is passed on through the `response=` argument and the name of the output data set, created with the macro, is defined through `out=`. For example, for the first depression trial, the following statement produces the complete case CC data set for the continuous $HAMD_{17}$ depression score:

```
%cc(data=depression,id=patient,time=visit,response=hamd17);
```

After performing this pre-processing, a complete case analysis follows of any type requested by the user. Note that the macro requires the records, corresponding to missing values, to be present in the data set.

When LOCF is of interest, similar steps to those for a complete case analysis need to be performed. For a vertically organized data set, the following SAS macro, `%locf` can be used:

```
%macro locf(data=,id=,time=,response=,out=);
%if %bquote(&data)= %then %let data=&syslast;
proc freq data=&data noprint;
tables &id /out=freqsub;
tables &time / out=freqtime;
run;
proc iml;
use freqsub;
read all var {&id,count};
nsub = nrow(&id);
use freqtime;
read all var {&time,count};
ntime = nrow(&time);
use &data;
read all var {&id,&time,&response};
n =  nrow(&response);
locf = &response;
ind = 1;
do while (ind <= nsub);
  if (&response[(ind-1)*ntime+ntime]=.) then
    do;
      i = 1;
      do while (&response[(ind-1)*ntime+i]^=.);
        i = i+1;
      end;
      lastobserved = i-1;
      locf[(ind-1)*ntime+lastobserved+1:(ind-1)*ntime+ntime]
        =locf[(ind-1)*ntime+lastobserved];
    end;
  ind = ind+1;
end;
create help var {&id &time &response locf};
append;
```

```
quit;
data &out;
merge &data help;
run;
%mend;
```

For the first depression trial, running the following statement produces the LOCF data set for the $HAMD_{17}$ outcome:

```
%locf(data=depression,id=patient,time=visit,response=hamd17,out=locf);
```

The arguments are exactly the same and have the same meaning as in the CC macro. Note that now there is a new response variable created, named `locf`, which should be used in the corresponding analysis programs.

11.2 Direct-Likelihood

As stated in Section 3.1.3, likelihood based inference is valid, whenever the mechanism is MAR and provided the technical condition holds that the parameters describing the nonresponse mechanism are distinct from the measurement model parameters. The log-likelihood then partitions into two functionally independent components, one describing the measurement model, the other one the missingness model. This implies that a likelihood-based software module yields valid inferences, provided the software tool used is able to handle measurement sequences of unequal length. Turning to SAS software, this is the case for the procedures MIXED, NLMIXED, and GLIMMIX. Note that no extra data manipulation is required, in contrast to CC and LOCF.

One note of caution is relevant, however. When residual correlation structures are used for which the order of the measurements within a sequence is important, such as unstructured and AR(1), but not simple or compound symmetry, and intermittent missingness occurs, care must be taken to ensure the *design* order within the sequence, and not the *apparent* order, is passed on. In the SAS procedure MIXED, a statement such as

```
repeated / subject=subject type=un;
```

is correct when every subject has for instance four designed measurements. However, when for a particular subject the second measurement is missing, there is a risk that the remaining measurement are considered the first, second, and third, rather than the first, third, and fourth. This means that the MIXED procedure by default assumes dropout. In that case, the time ordering needs to be specifies explicitly by replacing the above statement by:

```
repeated time / subject=subject type=un;
```

For the GENMOD procedure, the option `withinsubject=time` of the REPEATED statement can be used. Note that this produces GEE and not direct-likelihood.

When the NLMIXED procedure is used, only random effects can be included, and in such a case all relevant information is contained in the actual effects that define the random-effects structure. For example, for a random slope in time all information needed about time is passed on by the RANDOM statement:

```
random intercept time / subject=subject type=un;
```

Thus, in conclusion, a direct-likelihood analysis is no more complex than the corresponding analysis on a data set without missingness.

11.3 Weighted Generalized Estimating Equations

We illustrate WGEE by means of the analysis of the first depression trial discussed in Section 4.4.2. A GENMOD program for the standard GEE analysis would be:

```
proc genmod data=depression descending;
class patient trt visit;
model hamd17 = trt visit trt*visit basval basval*visit/ dist=binomial;
repeated subject=patient / withinsubject=visit type=cs modelse;
run;
```

Let us now discuss the steps to be taken to conduct a WGEE analysis. To compute the weights, one first has to fit the dropout model using, for example, logistic regression. The outcome 'dropout' is binary and indicates whether or not dropout occurs at a given time from the start of the measurement sequence until the time of dropout or the end of the sequence. Covariates in the model are the outcomes at previous occasions ('prev'), supplemented with genuine covariate information. The DROPOUT macro is used to construct the variables 'dropout' and 'prev'.

```
%macro dropout(data=,id=,time=,response=,out=);
%if %bquote(&data)= %then %let data=&syslast;
proc freq data=&data noprint;
tables &id /out=freqid;
tables &time / out=freqtime;
run;
proc iml;
reset noprint;
```

```
use freqid;
read all var {&id};
nsub = nrow(&id);
use freqtime;
read all var {&time};
ntime = nrow(&time);
time = &time;
use &data;
read all var {&id &time &response};
n = nrow(&response);
dropout = j(n,1,0);
ind = 1;
do while (ind <= nsub);
  j=1;
  if (&response[(ind-1)*ntime+j]=.) then print "First Measurement is Missing";
  if (&response[(ind-1)*ntime+j]^=.) then
    do;
      j = ntime;
      do until (j=1);
        if (&response[(ind-1)*ntime+j]=.) then
          do;
            dropout[(ind-1)*ntime+j]=1;
            j = j-1;
          end;
          else j = 1;
      end;
    end;
  ind = ind+1;
end;
prev = j(n,1,1);
prev[2:n] = &response[1:n-1];
i=1;
do while (i<=n);
  if &time[i]=time[1] then prev[i]=.;
  i = i+1;
end;
create help var {&id &time &response dropout prev};
append;
quit;
data &out;
merge &data help;
```

```
run;
%mend;
```

Once the logistic regression has been fitted, the predicted probabilities of dropout are translated into weights. These weights are defined at the individual measurement level and are equal to the product of the probabilities of not dropping out up to the measurement occasion. The last factor is either the probability of dropping out at that time or continuing the study. Let us describe the procedure to construct the inverse weights. At the first occasion, define the weight equal to 1. At other than the last occasion, the quantity of interest equals the cumulative weight over the previous occasions, multiplied by (1−the predicted probability of dropout). At the last occasion *within a sequence where dropout occurs*, it is multiplied by the predicted probability of dropout. At the end of the process, this quantity is inverted to yield the actual weight. This task can be performed using the DROPWGT macro. The arguments are the same as in the DROPOUT macro, except that now also the predicted values from the logistic regression have to be passed on through the `prev=` argument, and dropout indicator is passed on through the `dropout=` argument.

```
%macro dropwgt(data=,id=,time=,pred=,dropout=,out=);
%if %bquote(&data)= %then %let data=&syslast;
proc freq data=&data noprint;
tables &id /out=freqid;
tables &time / out=freqtime;
run;
proc iml;
reset noprint;
use freqid;
read all var {&id};
nsub = nrow(&id);
use freqtime;
read all var {&time};
ntime = nrow(&time);
time = &time;
use &data;
read all var {&id &time &pred &dropout};
n = nrow(&pred);
wi = j(n,1,1);
ind = 1;
do while (ind <= nsub);
    wihlp=1;
    stay=1;
```

```
    /* first measurement */
    if (&dropout[(ind-1)*ntime+2]=1)
      then do;
        wihlp = pred[(ind-1)*ntime+2];
        stay=0;
      end;
    else if (&dropout[(ind-1)*ntime+2]=0)
      then wihlp = 1-pred[(ind-1)*ntime+2];
    /* second to penultimate measurement */
    j=2;
    do while ((j <= ntime-1) & stay);
      if (&dropout[(ind-1)*ntime+j+1]=1)
        then do;
          wihlp = wihlp*pred[(ind-1)*ntime+j+1];
          stay=0;
        end;
      else if (&dropout[(ind-1)*ntime+j+1]=0)
        then wihlp = wihlp*(1-pred[(ind-1)*ntime+j+1]);
      j = j+1;
    end;
    j=1;
    do while (j <= ntime);
      wi[(ind-1)*ntime+j] = wihlp;
      j = j+1;
    end;
    ind = ind+1;
end;
create help var {&id &time &pred &dropout wi};
append;
quit;
data &out;
merge &data help;
data &out;
set &out;
wi = 1/wi;
run;
%mend;
```

Using the DROPOUT and DROPWGT macros, the following code can be used to prepare for a WGEE analysis :

```
%dropout(data=depression, id=patient, time=visit,
```

```
        response=hamd17, out=datahulp);

proc genmod data=datahulp descending;
class trt;
model dropout = prev trt / pred dist=binomial;
ods output obstats=pred;
run;

data pred;
set pred;
keep observation pred;
run;

data datahulp;
merge pred datahulp;
run;

%dropwgt(data=datahulp, id=patient, time=visit, pred=pred,
        dropout=dropout, out=datawgee);
```

To sum up, the dropout indicator and previous outcome variable are defined using the DROPOUT macro, whereafter an ordinary logistic regression is performed. Predicted values are first saved and then merged with the original data. Finally, the predicted values are translated into proper weights using the DROPWGT macro.

After this preparatory endeavor, we merely need to include the weights by means of the WEIGHT (or, equivalently SCWGT) statement within the GENMOD procedure. This statement identifies a variable in the input data set to be used a s the exponential family dispersion parameter weight for each observation. The exponential family dispersion parameter is divided by the WEIGHT variable value for each observation. Whereas the inclusion of the REPEATED statement turns a univariate exponential family model into GEE, the addition of WEIGHT further switches to WGEE. In other words, we just need to add:

```
weight wi;
```

11.4 Multiple-Imputation and GEE

In Chapter 5, the use of multiple imputation in a GEE setting is discussed and exemplified using the first depression trial in Section 5.3. The three tasks of multiple

imputation, that is, imputation, analysis, and inference, can be conducted within SAS. Two key procedures for multiple imputation are MI and MIANALYZE.

The MI procedure is used to generate the imputations. It creates M imputed data sets, physically stored in a single data set with indicator `_imputation_` to separate the various imputed copies from each other.

There are a variety of imputation mechanisms available, distinguishing between non-monotone and monotone sequences, and between continuous and categorical variables (Molenberghs and Kenward, 2007).

For imputations from a multivariate Gaussian imputation model the following MI program can be used for the continuous $HAMD_{17}$ score:

```
proc mi data=depressionhoriz out=miout simple nimpute=5 seed=5;
var trt y4 y5 y6 y7 y8;
run;
```

Note that the data need to be organized 'horizontal', that is, one record per subject, rather than 'vertically'. We will now describe some options available in the PROC MI statement, which are used above. The option `simple` displays simple descriptive statistics and pairwise correlations based on available cases in the input data set. The number of imputations is specified by `nimpute=` and is by default equal to 5. If more than one number is specified, one should use a VAR statement, and the specified numbers must correspond to variables in the VAR statement. The `seed=` option specifies a positive integer, which is used by PROC MI to start the pseudo-random number generator. The default is a value generated from the time of day from the computer's clock. The imputation task is carried out separately for each level of the BY variables. Although not essential, it is useful when an analysis needs to be checked afterwards or when a seed is specified by an external source such as a regulatory authority.

Incomplete categorical outcomes can be imputed by including them into the CLASS statement, in addition to their inclusion in the VAR statement. The following MI program imputes the dichotomized version of the $HAMD_{17}$ score using the conditional logistic regression model:

```
proc mi data=depressionhoriz out=miout nimpute=5 seed=5;
class ybin4 ybin5 ybin6 ybin7 ybin8;
var trt ybin4 ybin5 ybin6 ybin7 ybin8;
monotone logistic(ybin8=ybin4 ybin5 ybin6 ybin7 trt/descending);
run;
proc mi data=miout out=miout2 nimpute=1 seed=100;
by _imputation_;
```

```
class ybin4 ybin5 ybin6 ybin7;
var trt ybin4 ybin5 ybin6 ybin7;
monotone logistic(ybin7=ybin4 ybin5 ybin6 trt/descending);
run;
proc mi data=miout2 out=miout3 nimpute=1 seed=200;
by _imputation_;
class ybin4 ybin5 ybin6;
var trt ybin4 ybin5 ybin6;
monotone logistic(ybin6=ybin4 ybin5 trt/descending);
run;
proc mi data=miout3 out=miout4 nimpute=1 seed=300;
by _imputation_;
class ybin4 ybin5;
var trt ybin4 ybin5;
monotone logistic(ybin5=ybin4 trt/descending);
run;
```

After the imputation task, the imputed data sets are analysed using a standard complete data procedure, such as GEE resulting in MI-GEE. It is important to ensure that the BY statement is used such that a separate analysis is carried out for each data set.

Parameter estimates and their estimated covariance matrices need to be stored in appropriate output data sets, so they can be passed on to the MIANALYZE procedure which combines the M inferences using Rubin's formulae as described in Section 5.1. The MIANALYZE procedure has a generic form, but some care is needed when using it because estimates and accompanying covariance matrices have different names in different SAS procedures, and the output data sets corresponding to these may also be organized somewhat differently. Even though categorical effects and interactions can be used after including them in the CLASS statement, it is safer to to create appropriate indicator variables instead, as sometimes the mapping between parameter estimates and the corresponding precision parameters is not straightforward.

Thus, to prepare for the analysis for the first depression trial, indicator variables are created and the data are sorted by imputation number.

```
data mioutvert;
set mioutvert;
visit4=0;
visit5=0;
visit6=0;
visit7=0;
```

```
visit8=0;
trt0=0;
trt0visit4=0;
trt0visit5=0;
trt0visit6=0;
trt0visit7=0;
basvalvisit4=0;
basvalvisit5=0;
basvalvisit6=0;
basvalvisit7=0;
if visit=4 then visit4=1;
if visit=5 then visit5=1;
if visit=6 then visit6=1;
if visit=7 then visit7=1;
if visit=8 then visit8=1;
if trt=0 then trt0=1;
if visit=4 then basvalvisit4=basval;
if visit=5 then basvalvisit5=basval;
if visit=6 then basvalvisit6=basval;
if visit=7 then basvalvisit7=basval;
run;

proc sort data=mioutvert;
by _imputation_ patient visit;
run;
```

Next, the GENMOD procedure can be called for a GEE analysis, similar to the one presented in Section 11.3:

```
proc genmod data=mioutvert descending;
by _imputation_;
class patient trt visit;
model ybin = trt0 visit4 visit5 visit6 visit7
             trt0visit4 trt0visit5 trt0visit6 trt0visit7
             basval basval*visit4 basval*visit5 basval*visit6 basval*visit7
             / dist=binomial covb;
repeated subject=patient / withinsubject=visit type=cs modelse;
ods output GEEEmpPEst=miparms parminfo=miparminf CovB=micovb;
run;
```

Apart from the change to user-defined coding indicator variables for the categorical covariates in the model, the BY statement has been added, as well as the ODS

statement, to store the parameter estimates and the covariance parameters. For the latter, the `parminfo=` option is used next to the `CovB=` option, to ensure the proper names of the covariate effects are mapped to abbreviations of type `Prm1`, etc. Note that the `CovB=` option only works because the `covb` option was included in the MODEL statement. The parameter estimates are generated by default. The direct output of the GENMOD procedure will be a GEE analysis for each of the five imputed data sets.

Finally, the MIANALYZE procedure combines the M inferences into a single one. To combine the five inferences obtained from the GEE analyses applied to the first depression trial, the following MIANALYZE program can be used:

```
proc mianalyze parms=miparms parminfo=miparminf covb=micovb tcov;
modeleffects intercept trt0 visit4 visit5 visit6 visit7
             trt0visit4 trt0visit5 trt0visit6 trt0visit7
             basval basvalvisit4 basvalvisit5 basvalvisit6 basvalvisit7;
meantrt: test trt0+0.2*trt0visit4+0.2*trt0visit5
              +0.2*trt0visit6+0.2*trt0visit7;
trtatvisits: test trt0+trt0visit4,trt0+trt0visit5,
                  trt0+trt0visit6,trt0+trt0visit7,trt0;
overalltrt: test trt0+trt0visit4,trt0+trt0visit5,
                 trt0+trt0visit6,trt0+trt0visit7,trt0/mult;
run;
```

Parameter estimates and variance-covariance matrices are passed on through a combination of the `parms=` and `covb=` (or `xpxi=`) options. When the `covb=` matrices contain generic names (`Prm1`,...), the mapping between generic and actual parameter names is passed on using `parminfo=`. When one wishes to pass on either data sets of types COV, CORR, or EST, or a data set containing parameter estimates and standard errors, `data=` can be used instead. Including the `wcov`, `vcov`, and `tcov` options will print the within-imputation, between-imputation, and total covariance matrices, respectively. The parameters or effects for which multiple imputation inference is needed are passed on by means of the MODELEFFECTS statement. Categorical effects can be handled as well, after including them in the CLASS statement. However, as stated before, it is safer to create appropriate indicator variables to avoid the use of the CLASS statement. The TEST statement allows testing for hypotheses about linear combinations of the parameters. The statement is based on Rubin (1987), and uses a t distribution, which is the univariate version of the work by Li, Raghunathan and Rubin (1991).

11.5 Diggle-Kenward Model

In this section we exemplify the implementation of the Diggle-Kenward model using the SAS procedure IML by means of the second depression trial data as discussed in Section 6.1.2. For the measurement process, a specific case of the linear mixed model was considered with various fixed effects and a covariance matrix, $\boldsymbol{V}_i$, which was assumed to be of the heterogeneous first-order autoregressive type:

$$\boldsymbol{V}_i = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 & \cdots & \rho\,\sigma_1\sigma_{n-1} \\ \rho\,\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho\,\sigma_1\sigma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho\,\sigma_{n-1}\sigma_1 & \rho\,\sigma_{n-2}\sigma_1 & \cdots & \sigma_n^2 \end{pmatrix}. \tag{11.1}$$

The program can easily be adapted for another form of the linear mixed model (3.8), by just changing this $\boldsymbol{V}_i$ matrix. Before fitting the Diggle-Kenward model, some preparatory work is needed which is shown for the second depression trial data in case of MCAR missingness:

```
proc iml;
use depression;
read all var {id basval group time time2 y} into data;
    id = data[,1];
    basval = data[,2];
    group1 = data[,3];
    group0 = j(nrow(data),1,1)-group1;
    time = data[,4];
    timegroup0 = time#group0;
    timegroup1 = time#group1;
    time2 = data[,5];
    time2group0 = time2#group0;
    time2group1 = time2#group1;
    intercept=j(nrow(data),1,1);
create x var {intercept basval group0 time time2 timegroup0 time2group0};
append;
    y = data[,6];
create y var {y};
append;
    beta=6.84//-.36//-.21//-2.53//.16//.51//-.03;
    sigma1=sqrt(16.32);
    sigma2=sqrt(36.08);
```

```
    sigma3=sqrt(39.88);
    sigma4=sqrt(38.99);
    sigma5=sqrt(40);
    sigma6=sqrt(50);
    rho = .665;
    psi= -1.9;
    initial=beta//sigma1//sigma2//sigma3//sigma4//sigma5//sigma6//rho//psi;
create initial var {initial};
append;
    nsub=259;
    ntime=6;
create nsub var {nsub};
append;
create ntime var {ntime};
append;
quit;
```

Using IML the matrices `x` and `z` are created, which contain all $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ design matrices ($i = 1, \ldots, N$), as well as the vector `y` of all $\boldsymbol{Y}_i$ response vectors ($i = 1, \ldots, N$). Next, a vector of initial values for the parameters in the model is specified through `initial`. Finally, the number of subjects N and the number of time points n are required and passed on by `nsub` and `ntime` respectively.

As initial values for the parameters of the measurement process the parameter estimates of a ignorable direct-likelihood analysis based on the corresponding linear mixed model are used. For the parameters of the dropout model, initial values are shown in Table 11.1 for each of the three missingness mechanisms, MCAR ($\psi_1 = \psi_2 = 0$), MAR ($\psi_2 = 0$), and MNAR, respectively.

Table 11.1: *Initial values for the parameters of the dropout model.*

	Dropout Mechanism			
Parameter	MCAR	MAR	MNAR	MNAR + Covariate
ψ_0	1	$\widehat{\psi}_{0,MCAR}$	$\widehat{\psi}_{0,MAR}$	$\widehat{\psi}_{0,MNAR}$
ψ_1		1	$\widehat{\psi}_{1,MAR}$	$\widehat{\psi}_{1,MNAR}$
ψ_2			1	$\widehat{\psi}_{2,MNAR}$
ψ_3				1

Next, the Diggle-Kenward model can be fitted using IML code, which is available

from the CenStat website[1].

In the IML program, the module `loglik` evaluates the log-likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ for given set of parameters. Since this module requires integration over the missing data, the integrand is calculated in the module `integr`. The log-likelihood depends on the assumed mean and covariance structure and therefore one needs to adapt the module to the case considered. The following part of the module defines the mean and covariance structure for the second depression trial:

```
beta=parameters[1:7];
sigma1=parameters[8];
sigma2=parameters[9];
sigma3=parameters[10];
sigma4=parameters[11];
sigma5=parameters[12];
sigma6=parameters[13];
rho=parameters[14];
sigma11=sigma1**2;
sigma22=sigma2**2;
sigma33=sigma3**2;
sigma44=sigma4**2;
sigma55=sigma5**2;
sigma66=sigma6**2;
sigma12=rho*sigma1*sigma2;
sigma13=(rho**2)*sigma1*sigma3;
sigma14=(rho**3)*sigma1*sigma4;
sigma15=(rho**4)*sigma1*sigma5;
sigma16=(rho**5)*sigma1*sigma6;
sigma23=rho*sigma2*sigma3;
sigma24=(rho**2)*sigma2*sigma4;
sigma25=(rho**3)*sigma2*sigma5;
sigma26=(rho**4)*sigma2*sigma6;
sigma34=rho*sigma3*sigma4;
sigma35=(rho**2)*sigma3*sigma5;
sigma36=(rho**3)*sigma3*sigma6;
sigma45=rho*sigma4*sigma5;
sigma46=(rho**2)*sigma4*sigma6;
sigma56=rho*sigma5*sigma6;
```

and for a particular subject i the mean and covariance matrix is selected using:

```
mui = xi*beta;
```

[1] http://www.censtat.be/software/

```
vi = j(ntime,ntime,1);
vi[1,1] = sigma11;
vi[1,2] = sigma12;
vi[1,3] = sigma13;
...
vi[5,6] = sigma56;
vi[6,6] = sigma66;
```

Next, the log-likelihood is maximized using the Newton-Raphson ridge optimization method (`call nlpnrr`) thereby combining stability and speed. However, in other analyses, it may be necessary to try (several) other optimization methods, and a good number are available in SAS. In the program, we call `nlpnrr` as follows:

```
call nlpnrr(rc,est,"loglik",initial,opt,con);
```

Here, the argument `"loglik"` is the module of the function which has to be maximized, the log-likelihood in this case. The initial values to start the optimization method are listed in `initial`. The `opt` argument indicates an options vector that specifies details of the optimization process. Maximization instead of minimization is indicated by `opt[1]=1`. The output printed is controlled by `opt[2]`, which will yield summaries for the optimization start and termination, the iteration history, and initial as well as final parameter estimates. A constraint matrix is specified in `con`, defining lower and upper bounds for the parameters in the first two rows. However, in this case the heterogeneous first-order autoregressive covariance structure is assumed implying no constraints are needed. Finally, all optimization methods return the following results: the scalar return code `rc` and a row vector `est`. The return code indicates the reason for the termination of the optimization process. A positive return code indicates successful termination, whereas a negative one indicates unsuccessful termination, that is, that the result `est` is unreliable. The row vector `est` contains the optimal point - the maximum likelihood estimate - when the return code is positive.

Further, `nlpfdd` is called, which is a subroutine that approximates derivatives by finite differences method,

```
call nlpfdd(maxlik,grad,hessian,"loglik",est);
```

Again `"loglik"` is the module of the log-likelihood function. The vector that defines the point at which the functions and derivatives should be computed is `est`. This subroutine computes the function values `maxlik` - which is in this case the maximum likelihood value, since `est` is the maximum likelihood estimate -, the gradient vector `grad`, and the Hessian matrix `hessian`, which is needed to calculate the information matrix, and thus the covariance matrix `covar` and standard errors `stde`:

```
inf = - hessian;
covar = inv(inf);
var = vecdiag(covar);
stde = sqrt(var);
```

For the second depression trial data, the estimate and standard error for the treatment effect at week 9 are denoted as `diffest` and `diffse` respectively and are calculated as follows:

```
cov36=covar[3,6];
cov37=covar[3,7];
cov67=covar[6,7];
diffvar=var[3]+81*var[6]+(81*81)*var[7]+2*9*cov36+2*81*cov37+2*9*81*cov67;
diffse=sqrt(diffvar);
diffest=est[3]+9*est[6]+81*est[7];
```

Finally, to fit the model under the different missingness mechanisms a few lines in the program need to be adapted. For the model under MAR, we replace the following line, which corresponds to MCAR,

```
psi[1]=parameters[15];
```

by

```
psi[1:2]=parameters[15:16];
```

while in case of the MNAR assumption it is replaced by

```
psi[1:3]=parameters[15:17];
```

Further, under the MAR and MNAR assumption, we have to add one, or two columns of dots respectively, to the constraints matrix `con`.

Note that the model can be expanded by allowing the dropout process to depend on a covariate, such as for instance treatment `trt`. Thus, instead of (6.4), we use the following

$$\text{logit}\left[P(D_i = j \mid D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})\right] = \psi_0 \; + \; \psi_1 y_{i,j-1} \; + \; \psi_2 y_{ij} \; + \; \psi_3 \, \text{trt}_i. \quad (11.2)$$

In this case, the program needs to be adapted as follows. First, in the modules `integr` and `loglik`, we add `trt` and `trti` as global variables. Further, in the `loglik` module, when specifying the parameters, the parameter ψ_3 needs to be specified as well:

```
trt=parameters[18];
```

and where the information on one particular patient is selected, we add:

```
trti = xi[1,2];
```

Next, in the `integr` module, we replace

```
g=exp(psi[1]+psi[2]*lastobs+psi[3]*yd);
```

by

```
g=exp(psi[1]+psi[2]*lastobs+psi[3]*yd+trt*trti);
```

and in the `loglik` module,

```
 g = exp(psi[1]+yobs[j-1]*psi[2]+yobs[j]*psi[3]);
```

is replaced by

```
 g = exp(psi[1]+yobs[j-1]*psi[2]+yobs[j]*psi[3]+trt*trti);
```

Finally, compared to the MNAR program, we again add a column of dots to the constraints matrix `con`.

11.6 Local Influence Applied to Diggle-Kenward Model

In this section the local influence tool applied to the Diggle-Kenward model is exemplified using the SAS software. Using PROC IML, the normal curvature $C_{\boldsymbol{h}}$ of $\zeta(\boldsymbol{\omega})$ in $\boldsymbol{\omega}_0$, in the direction the unit vector $\boldsymbol{h}$ is calculated. The IML code for the second depression trial data can be found on the website.

Before running this program again the matrices `x`, `z`, the vectors `y` and `initial`, and the numbers `nsub` and `ntime` need to be created using PROC IML, analogous as in Section 11.5. The initial parameters used are the estimates of the parameters of the Diggle-Kenward model assuming MAR missingness. In the local influence implementation, the modules `integr` and `loglik` introduced in Section 11.5, are used to calculate the log-likelihood of the Diggle and Kenward (1994) model under the MNAR assumption. This is necessary for the evaluation of $\boldsymbol{\Delta}$ and $\ddot{\boldsymbol{L}}$. Next, the module `delta` calculates the $\boldsymbol{\Delta}$ vector, whereas $\ddot{\boldsymbol{L}}$ is calculated using `call nlpfdd`. Note that again the program needs to be adapted according to the assumed mean and covariance structure in both the modules `loglik` and `delta`. Additionally the calculation of the derivatives

$$\frac{\partial \lambda(y_{id}|\boldsymbol{h}_{id})}{\partial \boldsymbol{\theta}}$$

to obtain the columns $\boldsymbol{\Delta}_i$ of $\boldsymbol{\Delta}$ as shown in Section 7.3.2 needs to be adapted as well since the latter depends on the covariance structure considered. The implementation

for the second depression trial data is based on the derivation for the special case of three measurements, also shown in Section 7.3.2.

Finally, after $\boldsymbol{\Delta}$ is evaluated at the maximum likelihood estimate through

```
delta=delta(est);
```

the dataset `c_matrix` is created containing the following normal curvatures in the direction of the unit vector $\boldsymbol{h}_i$ containing one in the ith position and zero elsewhere,

$$\texttt{c} = C_i, \;\; \texttt{c1} = C_i(\boldsymbol{\beta}), \;\; \texttt{c2} = C_i(\boldsymbol{\alpha}), \;\; \texttt{c12} = C_i(\boldsymbol{\theta}), \text{ and } \texttt{c3} = C_i(\boldsymbol{\psi}),$$

and the normal curvature in the direction of `hmax` $= \boldsymbol{h}_{\max}$ of maximal normal curvature `cmax` $= C_{max}$. Note that the program requires the number of fixed effects, covariance parameters, and the dropout model parameters, which for the second depression trial is given by

```
nbeta=7;
nsigma=7;
npsi=2;
```

The `c_matrix` dataset can now be used to picture the local influence measures.

11.7 Latent-Class Mixture Model

The latent-class mixture model has been implemented using the GAUSS Software as we will show in this section. First, we explain how the software code is built up in general in Section 11.7.1. Afterwards we demonstrate the `pgm` files, which contain necessary functions for the analyses, as well as the main `bat` files, which include the actual code for a particular dataset, in Section 11.7.2 and 11.7.3 respectively.

11.7.1 General Code

The software code to fit a latent-class mixture model consists of four steps: (1) calculation of the maximum likelihood estimates, (2) calculation of the log-likelihood, standard errors and p-values, corresponding to the maximum likelihood estimates, (3) classification of subjects into the latent groups, and (4) drawing inferences. Let us describe the general code for these three steps in turn. First, according to Section 8.2.2, we can implement the estimation of the parameters of the latent-class mixture model as shown in Table 11.2, to obtain the maximum likelihood estimates $\widehat{\boldsymbol{\Omega}}$.

After obtaining the maximum likelihood estimate $\widehat{\boldsymbol{\Omega}}$ using the presented algorithm, the corresponding maximum likelihood value can be calculated as the second step. To

Table 11.2: *Gauss code. General form of estimation procedure to obtain the maximum likelihood estimates of the latent-class mixture model.*

1. Read dataset and create design matrices $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$
2. Give initial values $\boldsymbol{\Omega}_{initial}$ and initialize necessary global parameters (Table 11.3)
3. EM Algorithm: Iteration 1
   - $iteration = 1$;
   - E Step: For $i = 1, \ldots, N$ and $k = 1, \ldots, g$
     - ▷ Calculate $f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\Omega}_{initial})$
     - ▷ Calculate posterior probabilities $\pi_{ik}(\boldsymbol{\Omega}_{initial})$
   - M Step
     - ▷ For $k = 1, \ldots, g$, calculate the mean of posterior probabilities to obtain estimate $\boldsymbol{\pi}^{(1)} = (\pi_1^{(1)}, \ldots, \pi_g^{(1)})$ (maximum of $\mathcal{O}_1$ function)
     - ▷ Maximize $\mathcal{O}_2$ function π_{ik} using Newton-Raphson numerical technique to obtain estimates $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\alpha}^{(1)})$
   - $iteration = iteration + 1$;
4. EM Algorithm: Iterate until convergence

   $\varepsilon = \text{tol} + 1$;

   Repeat until $(|\varepsilon| < \text{tol})$ or $(iteration > \text{maxiter})$
   - E Step at iteration t: For $i = 1, \ldots, N$ and $k = 1, \ldots, g$
     - ▷ Calculate $f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\Omega}^{(t)})$
     - ▷ Calculate posterior probabilities $\pi_{ik}(\boldsymbol{\Omega}^{(t)})$
   - M Step at iteration t
     - ▷ For $k = 1, \ldots, g$, calculate the mean of posterior probabilities to obtain estimate $\boldsymbol{\pi}^{(t+1)}$ (maximum of $\mathcal{O}_1$ function)
     - ▷ Maximize $\mathcal{O}_2$ function π_{ik} using Newton-Raphson numerical technique to obtain estimates $(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\psi}^{(t+1)}, \boldsymbol{\alpha}^{(t+1)})$
   - $\varepsilon = \mathcal{O}(\boldsymbol{\Omega}^{(t+1)}) - \mathcal{O}(\boldsymbol{\Omega}^{(t)})$;
   - $iteration = iteration + 1$;

   → The EM algorithm converges to the maximum likelihood estimate $\widehat{\boldsymbol{\Omega}}$

Table 11.3: *Global variables.*

Global	Explanation
`_N`	number of subjects
`_g`	number of latent groups
`_n`	number of time points
`_qpoints`	number of quadrature points for numerical Gaussian-Quadrature
`[_x1,_x2]`	interval for numerical Gaussian-Quadrature integration
`_opgtol`	tolerance for Newton-Raphson numerical maximization of $\mathcal{O}_2$
`tol`	tolerance for EM algorithm to maximize log-likelihood
`maxiter`	maximum number of iteration for the EM algorithm

this end, a function is written in GAUSS. Using this function, the covariance matrix of the maximum likelihood estimate can be obtained, and consequently we can calculate standard errors and corresponding p-values for the model parameters. Note that the p-values are based on Wald tests, which are valid for the measurement model parameters, but perhaps not for the dropout model parameters or shared effects. Next, as shown in Section 8.3, subjects can be classified into the latent groups, using the posterior probabilities. Finally, inferences can be drawn, since we now have the maximum likelihood estimates of the parameters and the corresponding covariance matrix.

11.7.2 GAUSS PGM Files `shared.pgm` and `lcmm.pgm`

The GAUSS code we developed assumes only one subject-specific effect in the latent-class mixture model, that is, b_i, a shared intercept. Further, this shared intercept is assumed to have equal covariance across the different mixture components, that is, $d_1^2 = \ldots = d_g^2 = d^2$. Finally, the residual covariance matrices are assumed to be equal and of a simple structure, that is, $\boldsymbol{\Sigma}_i^{(1)} = \ldots = \boldsymbol{\Sigma}_i^{(g)} = \boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}_n$. These simplifications of the latent-class mixture model lead to $\boldsymbol{Y}_i | q_{ik} = 1, b_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + b_i, \sigma^2 \boldsymbol{I}_n)$, with $b_i \sim \sum_{k=1}^{g} \pi_k N(\mu_k, d^2)$.

Note when $g = 1$, the latent-class mixture model results in a classical shared-parameter model, which results for the measurement model in $\boldsymbol{Y}_i | b_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + b_i, \sigma^2 \boldsymbol{I}_n)$, with $b_i \sim N(0, d^2)$. Therefore, we created two files in GAUSS, `shared.pgm` and `lcmm.pgm`, containing functions for the shared-parameter model and the latent-class mixture model with $g \geq 2$ respectively. Both files can be found on the website.

Next, regarding the dropout model, we have split the file into two parts. In the first part, the dropout model does not depend on the shared intercept, that is,

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, b_i, q_{ik})] = \boldsymbol{\gamma}_k \boldsymbol{w}_{ij}, \tag{11.3}$$

whereas in the second part, the shared intercept is added to the dropout model, resulting in

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \boldsymbol{\gamma}_k \boldsymbol{w}_{ij} + \lambda\, b_i. \tag{11.4}$$

Let us now give an overview of the functions which are given in the `pgm` files. First, `shared.pgm` includes two functions, `integrand` and `shared`. The function `integrand` calculates for a given value of the shared intercept b_i the value of the integrand

$$f(\boldsymbol{y}_i^o|b_i, \boldsymbol{\theta}) f(d_i|b_i, \boldsymbol{\psi}) f(b_i|\boldsymbol{\alpha}),$$

with $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\alpha}$, the unknown parameters of the measurement model, the dropout model, and of the shared effect distribution, respectively. This function is used in the function `shared`, which represents the log-likelihood function

$$\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}) = \sum_{i=1}^{N} \ln \left\{ \int f(\boldsymbol{y}_i^o|b_i, \boldsymbol{\theta}) f(d_i|b_i, \boldsymbol{\psi}) f(b_i|\boldsymbol{\alpha})\ db_i \right\}.$$

The integration is computed over a finite interval $[x_1, x_2]$ using Gauss-Legendre quadrature, a numerical integration technique. Both functions appear twice in the `shared.pgm` file, and are distinguished by suffix 1 and 2. Suffix 1 corresponds to assuming dropout model (11.3), which yields that the integrand

$$f(\boldsymbol{y}_i^o|b_i, \boldsymbol{\theta}) f(b_i|\boldsymbol{\alpha})$$

is calculated in `integrand1` and the log-likelihood function in `shared1` is of the following form

$$\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}) = \sum_{i=1}^{N} \ln \left\{ f(d_i|\boldsymbol{\psi}) \int f(\boldsymbol{y}_i^o|b_i, \boldsymbol{\theta}) f(b_i|\boldsymbol{\alpha})\ db_i \right\},$$

whereas suffix 2 corresponds to assuming (11.4), for which both functions, `integrand2` and `shared2`, are as described above. Note that assuming (11.3) for the classical shared-parameter model essentially comes down to the MCAR missingness assumption.

Next, in the `lcmm.pgm` file, there are five function; `integrand`, `fydik`, `q2negagive`, `q1function`, and `lcmm`. Again, a suffix distinguishes between the assumption (11.3) or (11.4) for the dropout model. In the `integrand2` function, the value of the integrand

$$f(\boldsymbol{y}_i^o|b_i, q_{ik=1}, \boldsymbol{\theta}) f(d_i|b_i, q_{ik=1}, \boldsymbol{\psi}) f(b_i|q_{ik=1}, \boldsymbol{\alpha})$$

is given for a certain value of the shared intercept b_i. Under (11.3), the `integrand1` function results in computing

$$f(\boldsymbol{y}_i^o|b_i, q_{ik=1}, \boldsymbol{\theta})f(b_i|q_{ik=1}, \boldsymbol{\alpha}).$$

Next, this integrand is used in the calculation of $f_{ik}(y_i^o, d_i|q_{ik} = 1, \boldsymbol{\Omega})$ in the `fydik` function. Under the assumption of (11.3), we use `fydik1` or

$$f_{ik}(y_i^o, d_i|q_{ik} = 1, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) = f(d_i|q_{ik=1}, \boldsymbol{\psi}) \int f(\boldsymbol{y}_i^o|b_i, q_{ik=1}, \boldsymbol{\theta})f(b_i|q_{ik=1}, \boldsymbol{\alpha})\ db_i,$$

whereas under (11.4) the function `fydik2` is computed as follows:

$$f_{ik}(y_i^o, d_i|q_{ik} = 1, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) = \int f(\boldsymbol{y}_i^o|b_i, q_{ik=1}, \boldsymbol{\theta})f(d_i|b_i, q_{ik=1}, \boldsymbol{\psi})f(b_i|q_{ik=1}, \boldsymbol{\alpha})\ db_i.$$

Again, the integration is conducted over a finite interval $[x_1, x_2]$ using Gauss-Legendre quadrature numerical integration. Note that with these functions, the E step of the EM algorithm can be performed, as well as the calculation of the updated estimate for $\boldsymbol{\pi}$. However, to obtain the updated estimates for $(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})$, we need to maximize the $\mathcal{O}_2$ function, which is calculated in the `q2negative`. Since the optimization procedure `optmum` in GAUSS is only able to minimize functions, `q2negative1` and `q2negative2` actually are the negative of the $\mathcal{O}_2$ function, $-\mathcal{O}_2$. After the E and M step are conducted, we need to calculate ε of the particular iteration. To this end, in addition to the $\mathcal{O}_2$ function, we also need the $\mathcal{O}_1$ function, which is given in `q1function`. Up to here, we can perform the EM algorithm, and consequently we can obtain the maximum likelihood estimates. To get the maximum likelihood value however, we need an extra function, `lcmm`, which computes the log-likelihood value (8.4), or

$$\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{g} \pi_k f_{ik}(y_i^o, d_i|q_{ik} = 1, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) \right\}.$$

11.7.3 GAUSS BAT File for First Depression Trial

Next to the `pgm` files, presented in the previous section, there is a main Gauss file, that is, a `bat` file, which contains the code as shown in general in Section 11.7.1. Let us demonstrate this through the first depression trial data as discussed in Section 8.5.

The files `fdtspm.bat` and `fdtlcmm.bat` are provided on the website to fit respectively a shared-parameter model and a latent-class mixture model to the first depression trial data. As in the `pgm` files, both files split into two parts, distinguishing between either assuming dropout model (11.3) or (11.4).

First of all, the library `optmum` needs to be loaded as well as the necessary `pgm` file, for instance for the latent-class mixture model that is:

```
library optmum;
#include \gauss\myproc\lcmm.pgm;
```

In line with Table 11.2, the code contains 4 steps. The first step consist of reading the data set into GAUSS and create the design matrices. Note that this part of the programs needs to be adapted to one's data set and to the analysis at hand. For the first depression trial this becomes:

```
/* read data set */
dataset="c:\\depression";
open handle=^dataset for read;
data=readr(handle,10000);
call close(handle);
naam=getname(dataset);

/* specify necessary variables */
trt=data[.,9];
visit=data[.,2];
trtvisit=trt.*visit;
basval=data[.,10];
visitsq=visit.*visit;
trttime=data[.,6];
_subject=data[.,1]; /* necessary variable: subject indicator _subject */
_y=data[.,3];       /* necessary variable: response _y */
_time=visit;        /* necessary variable: time variable _time */
intercept=ones(rows(_y));

/* create design matrices */
_x=intercept~trt~visit~trtvisit~basval~visitsq;
_z=intercept;
_w=intercept~visit;
```

Next, GAUSS will either compute or request the necessary global parameters shown in Table 11.3. Also, the initial values for the EM algorithm need to be given. For the required global parameters and initial values a request appears, such as for instance for the initial values for the fixed effects parameters:

```
Give initial values for the fixed effects of the measurement model (beta):
```

In this case, one needs to enter the initial values for the fixed effects one by one.

In the analysis of the first depression trial, we used 40 quadrature points and the integration was done over the interval $[-20, 20]$. Further, the tolerance for the

Newton-Raphson numerical maximization of $\mathcal{O}_2$ was set to 0.00001, and for the EM algorithm, to maximize the log-likelihood, it was chosen to be 0.0001. Finally, the maximal number of iteration of the EM algorithm was 1000. To get sensible initial values, we fitted the linear mixed model with the same fixed and random effects as used in the heterogeneity model, as well as a logistic regression for the dropout model to the data. Since this is not taking into account any latent group structure, we can use the parameter estimates as initial values for the shared-parameter model. Afterwards, the parameters estimates of the shared-parameter model are used as initial values for the two-group latent-class mixture model, and so on. Note that this technique does not provide initial values for the group-specific parameters, and thus we get these by trial and error. Further, when the dropout model includes the shared intercept, the initial value for the extra parameter, λ, was also chosen by trial and error to be 0.10. Obviously, the same goes for the prior probabilities of the latent group components.

Next, the maximum likelihood estimate, denoted by `mle`, is obtained by means of the EM algorithm as described in Section 8.2.2. Afterwards this maximum likelihood estimate is printed as well as the maximized log-likelihood value, that is, the log-likelihood evaluated at the maximum likelihood estimate. Further, the Hessian of the log-likelihood function is calculated using the `hessp` function in GAUSS. Using this Hessian, evaluated in the maximum likelihood estimates, the covariance matrix, `covmatrix`, is obtained and consequently also the variances (`variances`) and standard errors (`stderrors`) corresponding to the maximum likelihood estimates. Finally, p-values corresponding to Wald tests are calculated.

As mentioned in Section 8.3, subjects can be classified into the different latent groups based on the posterior probabilities, which is done in the next step of the program. The matrix `postprob` contains the posterior probabilities for all subjects and latent groups, whereas the vector `classification` contains for each subject the number of the latent group to which it is classified.

Finally, in the last part, certain hypotheses can be tested. In the first depression trial, we test the hypothesis of treatment effect at the last visit as follows:

```
estdiff=mle[2]+5*mle[4];
vardiff=variances[2]+5*5*variances[4]+2*5*covmatrix[2,4];
sediff=sqrt(vardiff);
zvaldiff=estdiff/sqrt(vardiff);
pvaldiff=2*(1-cdfn(abs(zvaldiff)));
"estimate of treatment effect at time=5: " estdiff;
"standard error of treatment effect at time=5: " sediff;
"p-value of treatment effect at time=5: " pvaldiff;
```

Bibliography

Afifi, A. and Elashoff, R. (1966) Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association*, **61**, 595–604.

Agresti, A. (1990) *Categorical Data Analysis.* New York: Wiley.

Agresti, A. (2002) *Categorical Data Analysis.* (2nd ed.). New York: Wiley.

Aitkin, M. and Rubin, D. (1985) Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B*, **47**, 67–75.

Albert, A. and Lesaffre, E. (1986) Multiple group logistic discrimination. *Computers and Mathematics, with Applications*, **2**, 209–224.

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

Baker, S. G. (1995) Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics*, **51**, 1042–1052.

Baker, S. G. and Laird, N. M. (1988) Regression analysis for categorical variables with outcome subject to non-ignorable non-response. *Journal of the American Statistical Association*, **83**, 62–69.

Baker, S. G., Rosenberger, W. F. and DerSimonian, R. (1992) Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**, 643–657.

Beunckens, C., Molenberghs, G. and Kenward, M. G. (2005) Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*, **2**, 379–386.

Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2007a) A latent-class mixture model for incomplete longitudinal gaussian data. *Biometrics*, **Accepted**.

Beunckens, C., Sotto, C. and Molenberghs, G. (2007b) A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis*, **Accepted**.

Beunckens, C., Sotto, C., Molenberghs, G. and Verbeke, G. (2007c) An integrated sensitivity analysis of the slovenian public opinion survey data. *Journal of the Royal Statistical Society, Series C*, **Submitted**.

Böhning, D. (1999) *Computer-Assisted Analysis of Mixtures and Applications. Meta-analysis, disease mapping and others.* Chapman & Hall/CRC.

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

Carpenter, J. R., Kenward, M. G. and Vansteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, **3**, 571–584.

Chen, T. and Fienberg, S. E. (1974) Two-dimenstional contingency tables with both completely and partially cross-classified data. *Biometrics*, **30**, 629–642.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998) Analysis of longitudinal binary data from multi-phase sampling (with discussion). *Journal of the Royal Statistical Society, Series B*, **60**, 71–87.

Conaway, M. R. (1992) The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association*, **87**, 817–824.

Conaway, M. R. (1993) Non-ignorable non-response models for time-ordered categorical variables. *Applied Statistics*, **42**, 105–115.

Cook, R. D. (1979) Influential observations in linear regression. *Journal of the American Statistical Association*, **74**, 169–174.

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

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* London: Chapman and Hall.

Copas, J. B. and Li, H. G. (1997) Inference from non-random samples (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 55–96.

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dempster, A. P. and Rubin, D. B. (1983) Overview. In: *Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography* (Eds. W. G. Madow, I. Olkin and D. B. Rubin), pp. 3–10. New York: Academic Press.

Detke, M. J., Lu, Y., Goldstein, D. J., McNamara, R. K. and Demitrack, M. A. (2002) Duloxetine 60mg once daily dosing versus placebo in the acute treatment of major depression. *Journal of Psychiatric Research*, **36**, 383–390.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data (2nd ed.).* Oxford Science Publications. Oxford: Clarendon Press.

Diggle, P. J. and Kenward, M. G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 45–97.

Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models.* Heidelberg: Springer-Verlag.

Fay, R. E. (1986) Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, **81**, 354–365.

Follmann, D. and Wu, M. (1995) An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151–168.

Freeman, G. H. and Halton, J. H. (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, **38**, 141–149.

Gibbons, R. D., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H., Greenhouse, J. B., Shea, M., Imber, S., Sotsky, S. and Watkins, J. (1993) Soms conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry*, **50**, 739–750.

Glynn, R. J., Laird, N. M. and Rubin, D. B. (1986) Selection modeling versus mixture modeling with nonignorable nonresponse. In: *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag.

Hamilton, M. (1960) A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, **23**, 56–62.

Hartley, H. O. and Hocking, R. (1971) The analysis of incomplete data. *Biometrics*, **27**, 783–808.

Heckman, J. J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**, 475–492.

Hens, N., Aerts, M., Molenberghs, G., Thijs, H. and Verbeke, G. (2005) Kernel weighted influence measures. *Computational Statistics and Data Analysis*, **48**, 467–487.

Heyting, A., Tolboom, J. and Essers, J. (1992) Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine*, **11**, 2043–2061.

Hjort, N. L. and Claeskens, G. (2003) Frequentist model average estimators. *Journal of the American Statistical Association*, **98**, 879–899.

Hogan, J. W. and Laird, N. M. (1997) Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239–257.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006a) Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science*, **21**, 52–69.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G. and Kenward, M. G. (2006b) The nature of sensitivity in monotone missing not at random models. *Computational Statistics and Data Analysis*, **50**, 830–858.

Jansen, I. and Molenberghs, G. (2006) Pattern-mixture models for categorical outcomes with non-monotone missingness. In progress.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H. and Van Steen, K. (2003) A local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 409–418.

Jennrich, R. I. and Schluchter, M. D. (1986) Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, **42**, 805–820.

Kenward, M. G. (1998) Selection models for repeated measurements with nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine*, **17**, 2723–2732.

Kenward, M. G., Goetghebeur, E. J. T. and Molenberghs, G. (2001) Sensitivity analysis of incomplete categorical data. *Statistical Modelling*, **1**, 31–48.

Kenward, M. G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random. *Statistical Science*, **12**, 236–247.

Kenward, M. G., Molenberghs, G. and Thijs, H. (2003) Pattern-mixture models with proper time dependence. *Biometrika*, **90**, 53–71.

Laird, N. M. (1994) Discussion to Diggle, P. J. and Kenward, M. G.: Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 84.

Laird, N. M. and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lang, J. B. and Agresti, A. (1994) Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**, 625–632.

Lavori, P. W., Dawson, R. and Shera, D. (1995) A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, **14**, 1913–1925.

Lesaffre, E. and Verbeke, G. (1998) Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Li, K., Raghunathan, T. and Rubin, D. (1991) Large-sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *Journal of the American Statistical Association*, **86**, 1065–1073.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Little, R. J. A. (1992) Regression with missing X's: A review. *Journal of the American Statistical Association*, **87**, 1227–1237.

Little, R. J. A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

Little, R. J. A. (1994a) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.

Little, R. J. A. (1994b) Discussion to diggle, p. j. and kenward, m. g.: Informative dropout in longitudinal data analysis. *Applied Statistics*, **90**, 1112–1120.

Little, R. J. A. (1995) Modeling the drop-out 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: Wiley.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data.* New York: Wiley.

Little, R. J. A. and Yau, L. (1996) Intention-to-treat analysis in longitudinal studies with dropout. *Biometrics*, **52**, 1324–1333.

Liu, G. and Gould, A. L. (2002) Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of Biopharmaceutical Statistics*, **12**, 207–226.

Mallinckrodt, C. H., Clark, W. S., Carroll, R. J. and Molenberghs, G. (2003a) Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics*, **13**, 179–190.

Mallinckrodt, C. H., Clark, W. S. and Stacy, R. D. (2001a) Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics*, **(1 & 2)**, 9–21.

Mallinckrodt, C. H., Clark, W. S. and Stacy, R. D. (2001b) Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. *Drug Information Journal*, **35**, 4, 1215–1225.

Mallinckrodt, C. H., Sanger, T. M., Dub, S., DeBrota, D. J., Molenberghs, G., Carroll, R. J., Potter, W. Z. and Tollefson, G. D. (2003b) Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry*, **53**, 754–760.

Mallinckrodt, C. H., Watkin, J. G., Molenberghs, G. and Carroll, R. J. (2004) Choice of the primary analysis in longitudinal clinical trials. *Pharmaceutical Statistics*, **3**, 161–169.

McLachlan, G. and Peel, D. (2000) *Finite mixture models.* New York: Wiley.

Meng, X. L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Scinica*, **10**, 538–573.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T. and Thijs, H. (2002) Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine*, **21**, 1023–1042.

Michiels, B., Molenberghs, G. and Lipsitz, S. R. (1999) Selection models and pattern-mixture models for incomplete categorical data with covariates. *Biometrics*, **55**, 978–983.

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M. G. (2007) Every missing not at random model has got a missing at random bodyguard. *Journal of the Royal Statistical Society, Series B*, **Submitted**.

Molenberghs, G. and Kenward, M. (2007) *Missing Data in Clinical Studies.* New York: John Wiley.

Molenberghs, G., Kenward, M. G. and Goetghebeur, E. (2001a) Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *Applied Statistics*, **50**, 15–29.

Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997) The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, **84**, 33–44.

Molenberghs, G. and Lesaffre, E. (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

Molenberghs, G. and Lesaffre, E. (1999) Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18**, 2237–2255.

Molenberghs, G., Michiels, B. and Kenward, M. (1998a) Pseudo-likelihood for combined selection and pattern-mixture models for missing data problems. *Biometrical Journal*, **40**, 557–572.

Molenberghs, G., Michiels, B., Kenward, M. G. and Diggle, P. J. (1998b) Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.

Molenberghs, G., Michiels, B. and Lipsitz, S. (1999) A pattern-mixture odds ratio model for incomplete categorical data. *Communications in Statistics: Theory and Methods*, **28**, 2843–2869.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C. and Carroll, R. J. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**, 445–464.

Molenberghs, G., Thijs, H., Kenward, M. G. and Verbeke, G. (2003) Sensitivity analysis for continuous incomplete longitudinal outcomes. *Statistica Neerlandica*, **57**, 112–135.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer-Verlag.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E. and Kenward, M. G. (2001b) Mastitis in dairy cattle: Local influence to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93–113.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimisation. *The Computer Journal*, **7**, 303–313.

Neuhaus, J. M. (1992) Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research*, **1**, 249–273.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, **59**, 25–35.

Pharmacological Therapy for Macular Degeneration Study Group (1997) Interferon α-IIA is ineffective for patients with choroiadal neovascularization secondary to age-related macular degeneration. *Archives of Ophthalmology*, **115**, 865–872.

Plackett, R. L. (1965) A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Pocock, S. J. (1983) *Clinical Trials: A Practical Approach.* Chichester: John Wiley.

Potthoff, R. F. and Roy, S. N. (1964) A generalized multivariate analysis of variance model useful especially fo growth curve problems. *Biometrika*, **51**, 313–326.

Redner, R. and Walker, H. (1984) Mixture densities, maximum likelihood, and the em algorithm. *SIAM Review*, **26**, 195–239.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rotnitzky, A., Cox, D. R., Bottai, M. and Robins, J. (2000) Likelihood-based inference with singular information matrix. *Bernouilli*, **6**, 243–284.

Rotnitzky, A. and Wypij, D. (1994) A note on the bias of estimators with missing data. *Biometrics*, **50**, 1163–1170.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592. With comments by R. J. A. Little and a reply by the author.

Rubin, D. B. (1977) Formalizing subjective notions about the effect of nonresponse in sample surveys. *Journal of the American Statistical Association*, **72**, 538–543.

Rubin, D. B. (1978) Multiple imputation in sample surveys - a phenomenological bayesian approach to nonresponse. In: *Imputation and Editing of Faulty or Missing Survey Data*. Washington, DC: U.S. Department of Commerce.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Rubin, D. B. (1994) Discussion to Diggle, P. J. and Kenward, M. G.: Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 80–82.

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.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Schafer, J. L. (1999) Multiple imputation: a primer. *Statistical Methods and Medical Research*, **8**, 3–15.

Schafer, J. L. (2003) Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, **57**, 19–35.

Scharfstein, D. O., Rotnizky, A. and Robins, J. M. (1999) Adjusting for non-ignorable drop-out using semiparametric non-response models (with discussion). *Journal of the American Statistical Association*, **94**, 1096–1146.

Schwartz, D. and Lellouch, J. (1967) Explanatory and pragmatic attitudes in therpeutic trials. *Journal of Chronic Diseases.*, **20**, 637–648.

Sheiner, L. B., Beal, S. L. and Dunne, A. (1997) Analysis of non-randomly censored ordered categorical longitudinal data from analgesic trials. *Journal of the American Statistical Association*, **92**, 1235–1244.

Shen, S., Beunckens, C., Mallinckrodt, C. and Molenberghs, G. (2006) A local influence sensitivity analysis for incomplete longitudinal depression data. *Journal of Biopharmaceutical Statistics*, **16**, 365–384.

Shih, W. J. and Quan, H. (1997) Testing for treatment differences with dropouts present in clinical trials – a composite approach. *Statistics in Medicine*, **16**, 1225–1239.

Siddiqui, O. and Ali, M. W. (1998) A comparison of the random-effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*, **8**, 545–563.

Stasny, E. (1986) Estimating gross flows using panel data with nonresponse: an example of the canadian labour force survey. *Journal of the American Statistical Association*, **81**, 42–47.

Stiratelli, R., Laird, N. and Ware, J. (1984) Random-effects models for serial observations with dichotomous response. *Biometrics*, **40**, 961–972.

TenHave, T. R., Kunselman, A. R., Pulkstenis, E. P. and Landis, J. R. (1998) Mixed effects logistic regression models for longitudinal binary response data with informative dropout. *Biometrics*, **54**, 367–383.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G. and Curran, D. (2002) Strategies to fit pattern-mixture models. *Biostatistics*, **3**, 245–265.

Thijs, H., Molenberghs, G. and Verbeke, G. (2000) The milk protein trial: influence analysis of the dropout process. *Biometrical Journal*, **42**, 617–646.

Vach, W. and Blettner, M. (1995) Logistic regression with incompletely observed categorical covariates-investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*, **12**, 1315–1330.

Van Steen, K., Molenberghs, G., Verbeke, G. and Thijs, H. (2001) A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal*, **1**, 125–142.

Vansteelandt, S., Goetghebeur, E., Kenward, M. and Molenberghs, G. (2006) Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, **16**, 953–979.

Verbeke, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–222.

Verbeke, G., Lesaffre, E. and Spiessens, B. (2001a) The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, **35**, 419–439.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M. G. (2001b) Sensitivity analysis for non-random dropout: A local influence approach. *Biometrics*, **57**, 7–14.

Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Wu, M. C. and Bailey, K. R. (1988) Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, **44**, 175–188.

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 modeling the censoring process. *Biometrics*, **44**, 175–188.

Zhao, L. P., Lipsitz, S. and Lew, D. (1996) Regression analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165–1182.

Samenvatting

In een longitudinale studie wordt een bepaald kenmerk herhaaldelijk gemeten over de tijd. Op deze manier levert elk indivudu meer dan één waarneming aan, resulterend in een zogenaamd responsprofiel. Het geldig analyseren van dergelijke longitudinale studies is niet evident, en hangt onder andere af van het type van de responsvariabele. Het *linear-mixed model* wordt algemeen aanvaard als de basis voor de analyse van normaal verdeelde longitudinale gegevens. Voor niet-normaal verdeelde gegevens is er geen algemeen aanvaarde tegenhanger voor dit linear-mixed model. De longitudinale modellen in deze context omvatten (1) marginale modellen , (2) random-effecten (of individu-specifieke) modellen, en (3) voorwaardelijke modellen. Twee belangrijke vertegenwoordigers zijn de *generalized estimating equations* (GEE, Liang and Zeger, 1986) binnen de marginale familie en het *generalized linear-mixed model* (GLMM, Molenberghs and Verbeke, 2005) binnen de random-effecten familie.

Het geldig analyseren wordt nog een stap complexer als de gegevens onvolledig zijn. Een speciaal en veel voorkomend geval van onvolledigheid is wanneer een aantal individuen de geplande studie niet voleindigen. Men spreekt dan van *dropout*, of uitval. In het algemeen komt onvolledigheid voor door redenen die buiten de controle staan van de onderzoekers. Daarenboven kan deze onvolledigheid in verband staan met de metingen waarin men geïnteresseerd is. Dit duidt erop dat over het algemeen het noodzakelijk is ook de aandacht te richten op het onderliggende proces dat deze onvolledigheid bepaalt. Vermits men nooit zeker kan zijn over de exacte vorm van dit proces moeten er veronderstellingen gemaakt worden.

Rubin (1976) introduceert een formeel kader binnen het gebied van onvolledigheid, waarbij hij onderscheid maakt tussen drie mechanismen die onvolledigheid besturen, zijnde (1) *missing completely at random* (MCAR), waarbij het ontbreken van gegevens niet gerelateerd is aan de respons; (2) *missing at random* (MAR), waarbij onvolledigheid kan afhangen van *geobserveerde* respons; en (3) *missing not at random* (MNAR), waarbij het is gerelateerd aan respons, zowel de geobserveerde als de ontbrekende

gegevens. MCAR is de eenvoudigste onderstelling wegens onafhankelijkheid tussen onvolledigheid en respons, doch tegelijk zeer restrictief en zelden waar. MNAR is de meest flexibele veronderstelling, doch de relatie tussen onvolledigheid en *niet geobserveerde respons* maakt het formuleren van modellen en het schatten van parameters complex. Tevens leveren dergelijke modellen slechts identificeerbare parameters op mits zware en niet te toetsen veronderstellingen. Jammer genoeg zijn dergelijke veronderstellingen meestal niet zichtbaar en de preciese implicaties ervan ongekend.

Tegelijkertijd zijn er verscheidene modelfamilies vastgesteld om het meetproces en onvolledigheidsproces gezamenlijk te analyseren. In een *selectie model* (Little and Rubin, 1987) wordt de gezamenlijke verdeling van de metingen en het onvolledigheidsproces ontbonden wordt in de marginale verdeling van de metingen en de voorwaardelijke verdeling van het onvolledigheidsproces, gegeven de metingen. De omgekeerde ontbinding wordt het *pattern-mixture model* genoemd (Little, 1993, 1994a). Wanneer verondersteld wordt dat een aantal random effecten zowel de metingen als het onvolledigheidsproces beïnvloeden, en dat gegeven deze random effecten deze processen onafhankelijk zijn, dan spreekt men van een *shared-parameter model*. Een belangrijk concept in het gebied van onvolledigheid is *ignorability* (Rubin, 1976), wat stelt dat onder bepaalde voorwaarden het onvolledigheidsproces genegeerd kan worden, indien de interesse uitgaat naar conclusies aangaande het meetproces. Wanneer likelihood of Bayesiaanse methoden gebruikt worden, kan ignorability toegepast worden bij MCAR en MAR onvolledigheid, maar binnen het frequentistische kader is de strikte MCAR veronderstelling nodig om geldige analyse te bekomen.

Het is duidelijk dat sinds de publicatie van Rubin (1976), die onvolledigheid vastgesteld heeft als een onderzoeksgebied binnen statistiek, er een groot deel van het onderzoek is gewijd aan het probleem van onvolledigheid. Ondanks dat er een merkbaar onderscheid is in de verschillende denkwijzen zichtbaar in de methodologische ontwikkelingen, toch zijn onderzoekers het eens dat geen enkel model de beperking, van de ontbrekende gegevens nooit te zullen kennen, kan tenietdoen. Aan de ene hand benadrukken alle partijen, zijnde, de academische wereld, de industrie, en de regulerende autoriteiten, de nood voor sensitiviteitsanalyse, terwijl er aan de andere kant minder overeenkomst is over het soort van sensitiviteitsanalyse. Een belangrijke voorwaarde om een bepaalde methode voor te stellen als een haalbare methode binnen een sensitiviteitsanalyse, is de beschikbaarheid van vertrouwde en makkelijk te gebruiken software.

In deze thesis is het aangetoond dat het betreurenswaardig is dat er zoveel nadruk gelegd is op simpele methoden zoals *complete case* analyse of *last observation carried forward*, die tenminste eisen dat het onvolledigheids proces MCAR is. Deze

simpele methoden hebben we vergeleken met *direct-likelihood* analyse, waarbij alle beschikbare informatie gebruikt wordt zonder dat er bijkomende datamanipulatie nodig is. Bovendien is een direct-likelihood analyse geldig onder de minder strikte en meer realistische MAR onderstelling. In geval de conclusies verkregen zijn binnen het likelihood of Bayesiaans kader is het niet nodig om het onvolledigheidsproces te modelleren, en bijgevolg kan het linear-mixed model of het generalized linear-mixed model binnen de random-effecten familie gebruikt worden voor respectivelijk normaal verdeelde en niet-normaal verdeelde onvolledige longitudinale gegevens. Deze methoden zijn even gemakkelijk te implementeren dan wanneer er geen ontbrekende gegevens zouden geweest zijn.

Zoals eerder aangehaald wordt in het geval van niet-normaal verdeelde gegevens geopteerd voor de semi-parametrische GEE methode binnen de marginale familie. Aangezien deze methode vereist dat er voldaan is aan de MCAR onderstelling, zijn er alternatieven voorgesteld zoals *weighted generlized estimating equations* (WGEE, Robins *et al.*, 1994) en generlized estimating equations gebaseerd op *multiple imputation* (MI-GEE, Schafer, 2003), zodat de conclusies geldig zijn onder de MAR aanname. Voor beide methoden dient slechts een kleine dosis programmering uitgevoerd wat mogelijk is met standaard statistische software. In deze thesis hebben we deze twee methoden vergeleken gebruik makende van asymptotische simulaties en simulaties met kleine steekproefgrootte. In theorie is WGEE onvertekend, wat bevestigd werd door de asymptotische simulaties, maar deze eigenschap kan niet doorgetrokken worden voor kleinere steekproeven, zelfs niet wanneer elk aspect van de analyse correct gespecifieerd is. Daarenboven verdwijnt deze asymptotische onvertekening in geval van misspecificatie in het model dat de metingen of het onvolledigheidsproces beschrijft. Aan de andere kant heeft MI-GEE zijn robustheid bewezen onder misspecificatie van het het model voor de metingen of de imputatie. Bovendien resulteert MI-GEE in schattingen die minder vertekening en meer precisie vertonen in kleine tot middelmatige steekproeven vergeleken met WGEE. Omwille van deze opmerkingen adviseren we in de praktijk het gebruik van MI-GEE boven WGEE, ondanks de asymptotische onvertekening van WGEE. Ondanks de focus van deze thesis gericht is op onvolledigheid in de responsvariabele, merken we op dat onvolledigheid in de covariaten ook vaak voorkomend is, en in dit geval kan MI-GEE toegepast worden terwijl WGEE niet mogelijk is.

Tot hier toe hebben we duidelijk gesteld dat de simpele ad hoc methoden, die lange tijd in gebruik waren, eigenlijk een plaats verdienen in het museum voor statistiek, en dat de primaire analyse zou moeten bestaan uit methoden die het MAR mechanisme onderstellen voor de onvolledigheid. Aan de andere kant kan men echter

de mogelijkheid van MNAR onvolledigheid niet als zodanig uitsluiten, wat de nood impliceert om MNAR modellen te beschouwen. Daarom hebben we in deze thesis een overzicht gegeven van bestaande MNAR modellen, waarbij de aandacht voornamelijk gevestigd is op de modellen voorgesteld door Diggle and Kenward (1994) voor normaal verdeelde gegevens, en door Baker *et al.* (1992) voor binaire gegevens.

Een belangrijk kenmerk van het statistisch modelleren in geval er onvolledigheid optreedt, is dat kwaliteit van de fit voor de geobserveerde waarnemingen niets zegt over de geschiktheid van de structuur voor de ontbrekende gegevens die uit deze fit volgt. Dit hangt niet af van welke MNAR methode er gebruikt wordt, of nu een parametrische of niet-parametrische aanpak gekozen wordt. MNAR modellen zijn gebaseerd op veronderstellingen aangaande de ontbrekende gegevens en zijn derhalve niet verifieerbaar door middel van de beschikbare, geobserveerde waarnemingen. Bovendien hebben we in deze thesis aangetoond dat het empirische onderscheid tussen MNAR en MAR niet mogelijk is, in de zin dat de fit van elk MNAR model voor een verzameling van geobserveerde gegevens exact kan gereproduceerd worden door een MAR equivalent. Deze zogenaamde MAR *bodyguard* levert identiek dezelfde fit voor de geobserveeerde gegevens, maar de voorspellingen van de ontbrekende waarnemingen gegeven de geobserveerden is mogelijks verschillend. Een gevolg hiervan is dat men nooit de aanname van een MNAR model ten opzichte van een MAR model kan testen, tenzij men bereid is om het gestelde MNAR model te aanvaarden zonder meer. Dit benadrukt de sensitiviteit van de conclusies gebaseerd op MNAR modellen aangaande de vooropgestelde en niet verifieerbare modelonderstellingen. Dankzij deze bemerkingen en feiten is het duidelijk dat in geen enkel setting met ontbrekende gegevens er gesproken kan worden over één definitieve analyse. Een logisch compromis tussen blindelings te gaan voor MNAR modellen of ze volledig te negeren, is deze MNAR modellen te gebruiken als onderdelen voor een sensitiviteitsanalyse. Zo is het aan te raden om na een primaire analyse gebaseerd op de MAR assumptie een sensitiviteitsanalyse uit te voeren om de impact van de afwijkingen van deze MAR assumptie te onderzoeken.

In deze thesis geven we een overzicht van methoden die gebruikt kunnen worden in een sensitiviteitsanalyse, zowel op het niveau van het model door een waaier van modellen te beschouwen, als op het niveau van de individuen gebaseerd op *global* en *local influence*, beiden toegepast op het Diggle-Kenward model en de BRD modelfamilie. De local influence aanpak voor BRD modellen van Jansen *et al.* (2003) hebben we uitgebreid door enerzijds de terminologie te baseren op de aantallen per cel in plaats van de parameters, en anderzijds de celkansen lichtjes te verstoren in plaats van de de model parameters. Alhoewel de basis van local influence was om invloedrijke

individuen op te sporen, zijnde individuen die *non-randomly* uitvallen en op die manier het gestelde MAR model schijnbaar leiden in de MNAR richting, verscheidene auteurs (Verbeke *et al.*, 2001b; Jansen *et al.*, 2006b) hebben aangetoond dat deze invloedrijke individuen vaak invloedrijk zijn om andere redenen dan aangaande onvolledigheid.

Een verdere stap in sensitiviteitsanalyse is om naast selectiemodellen ook pattern-mixture modellen te beschouwen (Molenberghs *et al.*, 1998b; Thijs *et al.*, 2002), of zogenaamde *latent-class mixture models*, zoals aangetoond in deze thesis. Deze laatste is een uitbreiding van het shared-parameter model, vermits zowel het model voor de metingen als het model dat onvolledigheid beschrijft mogelijks één of meerdere random effecten gemeenschappelijk hebben, en gegeven deze random effecten worden beide processen onafhankelijk verondersteld. Bovendien deelt dit model eigenschappen met de drie verschillende modelfamilies, door informatie te gebruiken van de locatie en de evolutie van de respons profielen, een duidelijk selectie model concept, en van de dropout patronen, een pattern-mixture idee, om op die manier latente groepen en variabelen te definiëren, wat dan weer een karakteristiek is van het shared-parameter model. Door de veronderstelling van een latente structuur vangt het model mogelijke diversiteit op tussen latente groepen in de populatie. We hebben aangetoond dat dit latent-class mixture model niet alleen gebruikt kan worden als een flexibele modeltechniek, maar ook binnen sensitiviteitsanalyse, en om latente groepen verder te bestuderen.

Tenslotte merken we nog op dat alle voorgestelde modellen geimplementeerd kunnen worden in standard statistische software. Daarom hebben we in een laatste hoofdstuk getoond hoe de MAR analyses alsook de analyse gebaseerd op het Diggle-Kenward model en de local influence hierop toegepast, kan worden uitgevoerd gebruik makend van de SAS software. Verder bevat dit laatset hoofdstuk ook de implementatie in GAUSS van een vereenvoudigde versie van het latent-class mixture model zoals gebruikt doorheen deze thesis. Merk op dat er nog verder onderzoek aangaande dit model nodig is, zoals bijvoorbeeld in verband met modelselectie en modelonderstellingen, alsook de implementatie van een meer algemene versie.