Limburgs Universitair Centrum

Faculteit Wetenschappen

Statistical Models for Incomplete Longitudinal Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen: Wiskunde
aan het Limburgs Universitair Centrum te verdedigen door

BART MICHIELS

Promotor:
Prof. dr. G. Molenberghs

1999

Voor Ragna

Dankwoord

Graag zou ik iedereen willen danken die bijgedragen heeft aan het tot stand komen van deze thesis.

Op de eerste plaats komt dan natuurlijk mijn promotor Prof. Dr. Geert Molenberghs. Zonder zijn hulp en steun zou dit werk nooit tot stand zijn gekomen. Bedankt Geert, voor je uitstekende begeleiding en je aanstekelijk enthousiasme voor wetenschappelijk onderzoek. Je hebt mij binnengeleid in de boeiende wereld van de biostatistiek, en mij zo perfect voorbereid op mijn verdere carrière.

Ook mijn andere collega's op het LUC ben ik dank verschuldigd. Ze zorgden voor een aangename werksfeer, en zowel voor onderzoek, onderwijs, als consulting was het prettig samenwerken.

Tenslotte wil ik mijn familie en vrienden bedanken. De permanente steun en interesse van mijn ouders, Ragna, Veerle, en zovele anderen, was onmisbaar.

Bart Michiels

Diepenbeek, 3 juni 1999

Contents

1 Introduction **1**

2 Key Examples **9**

2.1 Fluvoxamine Data 9

2.2 The Vorozole Study 14

3 Missing Data Terminology **17**

3.1 Missing Data 18

3.1.1 Categorical Data 19

3.1.2 Missing Data Mechanisms 20

3.1.3 Ignorability 22

3.2 Approaches to Incomplete Data 24

3.2.1 EM-algorithm 24

3.2.2 Multiple Imputation 25

4 Protective Estimation **29**

4.1 The Protective Estimator for Gaussian Data 30

4.2 The Protective Estimator for Categorical Data 33

4.3 Likelihood Estimation 37

4.4 Pseudo-Likelihood Estimation 42

4.5 Variance Estimation 43

4.5.1 Delta Method 43

4.5.2 EM Aided Differentiation 46

4.5.3 Multiple Imputation 48

4.5.4 Illustration 49
4.6 Examples 50
4.6.1 Fluvoxamine Data 50
4.6.2 Koch Dataset 55
4.7 Conclusion 55

5 Missing at Random for Pattern-Mixture Models 59
5.1 Available Case Missing Value Restriction 60
5.2 Non-Monotone Patterns: A Counterexample 65
5.3 Conclusion 66

6 Selection Models and Pattern-Mixture Models for Incomplete Data With Covariates 69
6.1 Marginal Modelling of Incomplete Categorical Data 70
6.2 Selection Models Versus Pattern-Mixture Models 71
6.2.1 Notation 71
6.2.2 Selection Models 72
6.2.3 Pattern-Mixture Models 74
6.2.4 Identifying Restrictions 75
6.2.5 Relative Merits of Both Families 77
6.2.6 Precision Estimation with Pattern-Mixture Models 79
6.3 Analysis of Fluvoxamine Data 80
6.3.1 Selection Modelling for Side Effects 80
6.3.2 Pattern-Mixture Modelling for Side Effects 83
6.3.3 Comparison for Side Effects 87
6.3.4 Selection Modelling for Therapeutic Effect 89
6.3.5 Pattern-Mixture Modelling for Therapeutic Effect 90
6.3.6 Comparison for Therapeutic Effect 94
6.3.7 Different Missing Data Mechanisms 95
6.4 Conclusion 95

7 Pseudo-Likelihood Estimation for a Combined Selection and Pattern-Mixture Model 99

7.1 Pseudo-Likelihood . . . 100
7.1.1 Definition and Properties . . . 100
7.1.2 Missing Data Mechanisms . . . 102
7.2 A Trivariate Loglinear Model . . . 103
7.3 No Underlying Joint Density . . . 108
7.4 Fluvoxamine Data . . . 111
7.5 Conclusion . . . 112

8 Sensitivity Analysis for a Longitudinal Quality of Life Measure in a Cancer Trial 115

8.1 A Repeated-Measures Model . . . 116
8.2 Exploratory Analysis . . . 117
8.2.1 The Average Evolution . . . 117
8.2.2 The Variance Structure . . . 118
8.2.3 The Correlation Structure . . . 119
8.2.4 The Variogram . . . 121
8.3 A Selection Model Formulation . . . 122
8.4 A Pattern-Mixture Model Formulation . . . 129
8.5 Conclusion . . . 137

9 Concluding Remarks and Further Research 139

Samenvatting 141

References 145

List of Tables

2.1 Fluvoxamine Data, Completers. Each cell gives the outcome at the four visits. The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern. 11

2.2 Fluvoxamine Data, Incomplete Patterns. Each cell gives the outcome at the four visits (a "." indicates the corresponding measurement is missing). The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern. 12

2.3 Fluvoxamine Data. Each cell gives the dichotomized (0 vs. 1/2/3) outcome at the first 3 visits (a "." indicates the corresponding measurement is missing). The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern. 13

2.4 Fluvoxamine Data. Each cell gives the dichotomized (0 vs. 1/2/3) outcome at the first and last visit (a "." indicates the corresponding measurement is missing). The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern (those with a missing covariate are indicated separately). 14

4.1 Four Sets of Artificial Data. Each time a contingency table for the completers ($Y_1 = 1/2, Y_2 = 1/2$), and an additional contingency table for the dropouts ($Y_1 = 1/2$) is given. 39

4.2 Parameter Estimates (Standard Errors) for the Artificial Data from Table 4.1. Methods of Estimation Are: Likelihood (Untransformed and Transformed), Protective Estimation With Delta Method, EM Algorithm, and Multiple Imputation. 40

4.3 Estimated Cell Probabilities (Standard Errors) for the Fluvoxamine Data (all quantities were multiplied by 1000). The cell gives the outcomes at the 3 times considered. 6 Methods were used: likelihood estimation once for the completers only, and once assuming MAR, protective estimation using the Delta method, the EM algorithm and multiple imputation to calculate standard errors, and finally pseudo-likelihood using the protective assumption. 51

4.4 Parametric Models for the Fluvoxamine Data, Side Effects, Likelihood Based Estimation. The cell gives the outcomes at the 3 times considered. Measurement probabilities are multiplied by 1000. 53

4.5 Estimated Cell Probabilities for the Koch Dataset (all quantities were multiplied by 100). Stratification for sex is indicated by male (M) and female (F); stratification for area is indicated by area 1 (1) and area 2 (2). A + indicates that the corresponding stratificator is not used. . 56

6.1 Fluvoxamine Data, Side Effects: Selection Model (full) 81

6.2 Fluvoxamine Data, Side Effects: Selection Model (reduced) 82

6.3 Fluvoxamine Data, Side Effects: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (full) 85

6.4 Fluvoxamine Data, Side Effects: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (reduced) 86

6.5 Fluvoxamine Data, Therapeutic Effect: Selection Model (full) 89

6.6 Fluvoxamine Data, Therapeutic Effect: Selection Model (reduced) . 90

6.7 Fluvoxamine Data, Therapeutic Effect: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (full) 91

6.8 Fluvoxamine Data, Therapeutic Effect: Pattern-Mixture Model, Multiple Imputation (MI) (full) 93

6.9 Fluvoxamine Data, Therapeutic Effect: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (reduced) 94
6.10 Fluvoxamine Data, Side Effects: Different Pattern-Mixture Models (full) 96

7.1 Four Sets of Artificial Data. Each time a contingency table for the completers ($Y_1 = 0/1, Y_2 = 0/1$), and an additional contingency table for the dropouts ($Y_1 = 0/1$) is given. 104
7.2 Parameter Estimates (Standard Errors) for the Artificial Data from Table 7.1, based on a Loglinear Model 105
7.3 Parameter Estimates (Standard Errors) for the Artificial Data from Table 7.1, based on a Pseudo-Likelihood containing Three Logistic Models 110
7.4 Interpretation of the Parameters (logits of the means of the random variables for α_1, α_2, and α_3, and log odds ratios for α_4 and α_5) 110
7.5 Fluvoxamine Data, Side Effects (full) 113
7.6 Fluvoxamine Data, Side Effects (reduced) 113

8.1 Vorozole Study: Selection Model 125
8.2 Vorozole Study: First Pattern-Mixture Model 134
8.3 Vorozole Study: Second Pattern-Mixture Model 135

List of Figures

6.1 Fluvoxamine Data, Selection Model: Probabilities of Side Effects w.r.t. Age (a) at the First Occasion, (b) at the Last Occasion, (c) at Any Occasion 83

6.2 Fluvoxamine Data, Pattern-Mixture Model: Probabilities of Side Effects w.r.t. Age (a) at the First Occasion, (b) at the Last Occasion, (c) at Any Occasion 87

8.1 Vorozole Study: Individual Profiles for Change (Raw, Detrended, and Standardized) 118

8.2 Vorozole Study: Mean Profiles and 95% Confidence Intervals 119

8.3 Vorozole Study: Variance Function 120

8.4 Vorozole Study: Scatterplot Matrix 121

8.5 Vorozole Study: Variogram 123

8.6 Vorozole Study: Fitted Profiles (averaging the predicted means for the incomplete and complete measurement sequences, without random effects) 126

8.7 Vorozole Study: Fitted Profiles (averaging the predicted means for the incomplete and complete measurement sequences, including the random effects) 127

8.8 Vorozole Study: Observed Dropout per Treatment Arm 128

8.9 Vorozole Study: Individual Profiles per Dropout Pattern 131

8.10 Vorozole Study: Mean Profiles per Dropout Pattern 132

8.11 Vorozole Study: Fitted Selection and First Pattern-Mixture Model . . 133

8.12 Vorozole Study: Fitted Selection and Second Pattern-Mixture Model 136

Chapter 1

Introduction

Many clinical studies record relevant outcome measures repeatedly over time. Historically, such designs were often motivated by data monitoring reasons, and analysis still relied on univariate techniques (the last observed measurement, endpoint analysis, summary measures). Currently, there is a tendency towards the use of genuine longitudinal data techniques. Longitudinal data analysis is a very active area of research and only recently have a number of book references become available (Lindsey 1993, Longford 1993, Diggle, Liang and Zeger 1994, Hand and Crowder 1995, Verbeke and Molenberghs 1997, Vonesh and Chinchilli 1997). The linear mixed model for normally distributed endpoints (Laird and Ware 1982) is perhaps the most widespread, supported by the availability of flexible software: the SAS procedure MIXED (Littell *et al.* 1996), the SPlus function LME, SPSS, etc. While the linear mixed model, or its extension that incorporates serial association (Diggle 1988), is well established, its application is still not straightforward.

In contrast to normal data, there is less agreement on models for non-normal data (binary data, counts). Among the most popular ones we find generalized estimating equations (Liang and Zeger 1986) and generalized linear mixed models (Wolfinger and O'Connell 1993, Breslow and Clayton 1993).

In a repeated measures design it is common that some variables fail to be recorded for everybody. Sometimes, there is just one missing value for a subject, but it is not unusual in practice for some sequences of measurements to terminate early for reasons outside the control of the investigator, and any unit so affected is often called

a dropout. For categorical outcomes, incomplete data entail that a subject is not always classified into a single outcome category but rather into a set of categories, whereas the actual single category represents the complete data. The frequent occurence of dropout means that this is a common problem when analysing incomplete longitudinal data, and it might therefore be necessary to accommodate dropout in the modelling process. As a result of including the missingness process, one can obtain correct inference, and answer a question of scientific interest about the missingness.

Following Laird (1988), major problems with incomplete longitudinal data are the difficulties with implementing existing methods, efficiency loss, and the introduction of bias. Possible approaches are, apart from complete case analyses, univariate analyses with adjustments for variance estimates, two-step analyses, and likelihood based methods. Laird further distinguishes between likelihood based methods that include an explicit model for dropout and methods that only model the measurement process. Heyting, Tolboom and Essers (1992) discuss shortcomings of available methods, and advocate the use of more modern methodology that was primarily developed for sample surveys with non-response and for observational studies. These authors consider the following common causes of dropout: recovery, lack of improvement, unwanted signs or symptoms that may be related to the investigational treatment, unpleasant study procedures, intercurrent health problems, external reasons that seem to be unrelated to the trial procedures or to the progress of the patient.

When refering to the missing data mechanism (non-response process) we will use terminology first introduced by Rubin (1976), and further developed in Little and Rubin (1987, Ch. 6). A non-response process is said to be missing completely at random (MCAR) if the missingness is independent of both unobserved and observed data and missing at random (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither completely random nor random is termed non-random (MNAR). In the context of likelihood inference, and when the parameters describing the measurement process and the parameters describing the missingness process are distinct, MCAR

and MAR are ignorable, while a MNAR process is non-ignorable. If the process is ignorable, then a valid analysis can be obtained through a likelihood-based analysis that ignores the missing data mechanism. This leads to considerable simplification in the analysis. Furthermore, there are situations where the MAR assumption gives better results than MNAR (Rubin, Stern and Vehovar 1995). In many examples, however, the reasons for missingness are many and varied and it is therefore difficult to justify on a priori grounds the assumption of random missingness. Arguably, in the presence of non-random missingness, a wholly satisfactory analysis of the data is not feasible. While the treatment of missing data that are missing at random requires some caution, one needs to be even more careful with non-randomly missing data. This contradicts a common belief that, with the availability of methods for incomplete data, fitting models is of the same level of complexity as any other statistical model building exercise and that in fact routine testing for the non-randomness of the non-response process is possible. However, many instances of the contrary have been reported. A classical example is found in Little and Rubin (1987, Section 11.6). Several issues are discussed in Molenberghs *et al.* (1999). It is illustrated how models are identifiable by virtue of model assumptions, which are usually impossible to verify merely on statistical grounds. In addition to the potential occurrence of non-unique, boundary solutions, and solutions that violate constraints, we show that models often yield the same or similar fits to the observed data, but produce qualitatively different predictions for the unobserved data. Other issues are presented in Molenberghs, Goetghebeur, Lipsitz and Kenward (1999).

Work on incomplete categorical data has largely been devoted to partially classified contingency tables (see e.g., Baker and Laird 1988). Molenberghs, Kenward and Lesaffre (1997) introduce a method for the analysis of longitudinal ordinal data with non-random dropout. Their approach is based on Diggle and Kenward (1994), who treat non-random dropout in continuous longitudinal data. The EM algorithm (Dempster, Laird and Rubin 1977) is extensively used to maximize the likelihood in case of incomplete categorical data, but other proposals have been made as well: Molenberghs and Goetghebeur (1997) have introduced a simple method to construct and maximize the observed data likelihood, whilst still formulating their models at

the complete data level. An overview of methods for missing data in longitudinal data is given in Laird (1988). The author distinguishes between ignorable and non-ignorable missingness, in the context of both normally distributed and categorical data. Lehnen and Koch (1974) present a saturated likelihood approach. They have to assume that the missingness is completely random. Log-linear models for ignorable incomplete data can be fitted using the EM algorithm (Fuchs 1982). Baker and Laird (1988) and Stasny (1986) consider models for non-ignorable nonresponse with categorical outcomes. A general framework is provided by Fay (1986). Conaway *et al.* (1992) use loglinear models and perform fitting within GLIM, with the aid of the EM algorithm. Park and Brown (1994) follow a similar line of reasoning.

Diggle and Kenward (1994) propose an MNAR analysis by letting the probability of dropout depend on the possibly unobserved outcomes. This probability model is then combined with a linear mixed model for the measurement process. Further approaches are proposed by Schluchter (1988), Laird, Lange and Stram (1987), Wu and Bailey (1988, 1989), Wu and Carroll (1988). These last authors use random-effects models to describe the censoring or non-response process. Greenlees, Reece and Zieschang (1982), combine the probit of the dropout probability with the general linear measurement model. Little (1995) gives a careful review of the different modelling approaches.

Brown (1990) has shown that some progress can be made when the missingness is assumed to depend on the unobserved values, but not on the observed measurements. For multivariate Gaussian data, Brown constructs an estimator for the mean and covariance parameters of the joint normal distribution. He called this a protective estimator. Brown studies an example where measurements are made relatively far apart in time, such that the influence of a previous measurement on nonresponse is negligible. He allows for the presence of all possible response patterns, but each subject is measured at the first occasion. The categorical counterpart for this protective estimator can be found in Michiels and Molenberghs (1997). In Chapter 4, one can find the protective estimator in both the normal and the categorical case.

Most methods are formulated within the selection modelling frame (Little and

Rubin 1987) as opposed to pattern-mixture modelling (Little 1993). A selection model factors the joint distribution of the measurement and response mechanisms into the marginal measurement distribution and the response distribution, conditional on the measurements. This is intuitively appealing since the marginal measurement distribution would be of interest also with complete data. Further, Little and Rubin's taxonomy (MCAR, MAR, MNAR) is most easily developed in the selection setting. However, it is often argued that, especially in the context of non-random missingness models, selection models, although identifiable, should be approached with caution. This point is well illustrated in Glynn, Laird and Rubin (1986). One is confronted with fundamentally untestable assumptions, a point raised by many discussants to Diggle and Kenward (1994).

Recently, Little (1993, 1994a, 1995) has been promoting the use of pattern-mixture models as a viable alternative. His work is based on earlier material, such as Rubin (1977) where the idea was used in a sensitivity analysis within a fully Bayesian framework. Further references include Glynn, Laird and Rubin (1993), Little and Rubin (1987), and Rubin (1987). In 1989, an entire issue of the Journal of Educational Statistics was devoted to this theme and especially Wainer (1989), as well as the accompanying discussion, deserves particular attention. Within the family of pattern-mixture models, the marginal response distribution is coupled with the measurement process, given the response pattern. Of course, these models are by construction under-identified, since the complete data distribution of an incomplete pattern is a priori a contradiction. Little solves this problem through the use of identifying restrictions. In other words, inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers. In a fully Bayesian context, both selection modelling and pattern-mixture modelling has been used to investigate sensitivity (Rubin 1977).

Although selection models and pattern-mixture models are interchangeable from a probabilistic point of view, in the sense that they represent different factorizations of the same joint distribution, in practice they encourage different kinds of simplifying assumptions. For this reason, it is important to consider their relative merits as scientific models, especially when the probability of missingness depends on the

unobserved outcomes. One attraction of selection models is that they fit naturally into Little and Rubin's taxonomy, whereas pattern-mixture models appear not to do so. In Molenberghs, Michiels, Kenward and Diggle (1998), it is shown, on the contrary, that the classical taxonomy of missing data models can also be applied to pattern-mixture models. Clearly, since MCAR is merely independence between measurement and dropout processes, it is common to both settings. If missingness is restricted to dropout, MAR corresponds to what we call available case missing value (ACMV) restrictions. Under this identifying restriction, a distribution that is impossible to estimate directly in one pattern is set equal to the corresponding distribution over the patterns for which all necessary components are observed. A formal definition of ACMV can be found in Chapter 5.

Now that different missing data mechanisms are established in a selection model and a pattern-mixture model context, a purely philosophical debate about the relative merits of the selection model and pattern-mixture model paradigms is unhelpful. Instead, the choice of one of the models should be based on the statistical and scientific merits of proposed missing value models on their own terms. For example, if the question of scientific interest regards the treatment effect, averaged over all dropout patterns, then choosing a selection model seems to be obvious. On the other hand, if one is interested in the treatment effect, for various dropout patterns separately, then a pattern-mixture model is a natural choice.

Furthermore, we advocate the use of pattern-mixture models as a tool to assess sensitivity of a selection model to the modelling assumptions, or vice versa (Molenberghs, Michiels and Lipsitz 1999, Michiels, Molenberghs and Lipsitz 1998). Explicitly, extra confidence in the conclusions can be gained if two analyses, one within each framework, coincide in key aspects, such as covariate dependencies, strength of association between outcomes, etc. In Chapter 6, a selection model and a pattern-mixture model with covariates will be fitted to the same set of data. We will also show for the pattern-mixture model how precision estimates can conveniently be obtained using profile likelihood and multiple imputation. The main emphasis is on a marginal model (Molenberghs and Lesaffre 1994) since the use of identifying restrictions in this context is less than straightforward. In this context, we choose

to restrict attention mainly to MAR processes since they can be expressed easily in both selection and pattern-mixture frameworks (Molenberghs, Michiels, Kenward and Diggle 1998). Another example can be found in Chapter 8, where a repeated-measures model is used to fit the data from a breast cancer study to a selection model and a pattern-mixture model. In the analysis of this set of continuous outcomes, the MAR assumption is relaxed. It is then shown how pattern-mixture models can be used without explicit use of identifying restrictions (Michiels *et al.* 1998).

An alternative method to combine selection models and pattern-mixture models consists in using a pseudo-likelihood (Arnold and Strauss 1991), containing the interesting parts of both models. Although this looks appealing, only the MAR/ACMV case is straightforward (Molenberghs, Michiels and Kenward 1998). Problems arise if one wants to include non-random missingness. The pseudo-likelihood application can be found in Chapter 7.

Chapter 2

Key Examples

In this chapter, two datasets that are often used throughout this manuscript, are introduced. The first dataset contains data from a psychiatric study, where as main outcomes the therapeutic effect and the side effects are measured. The second dataset was collected at the Janssen Research Foundation, where a quality of life quantity is studied. Other datasets are introduced whenever the need arises.

2.1 Fluvoxamine Data

The data come from a multicentre study involving 315 patients that were treated by fluvoxamine for psychiatric symptoms described as possibly resulting from a dysregulation of serotonine in the brain. Patients with one or more of the following diagnoses were included: depression, obsessive, compulsive disorder and panic disorder. After recruitment of the patient in the study, he or she was investigated at four visits: at weeks 2, 4, 8 and 12. On the basis of about twenty psychiatric symptoms, the *therapeutic effect* and the *side effects* were scored at each visit in an ordinal manner. Side effect is coded as (0) = no; (1) = not interfering with functionality of patient; (2) = interfering significantly with functionality of patient; (3) = the side effects surpass the therapeutic effect. Similarly, the effect of therapy is recorded on a four point ordinal scale: (0) no improvement or worsening; (1) minimal improvement (not changing functionality); (2) moderate improvement (partial disappearance of symptoms) and (3) important improvement (almost disappearance

of symptoms). Thus, side effects occur if new symptoms occur while there is therapeutic effect if old symptoms disappear. Also 3 covariates are measured for these patients: age, a continuous covariate, and sex and the occurence of psychiatric antecedents, two binary covariates. The outcomes can be found in Tables 2.1 and 2.2. These data were analysed by Molenberghs and Lesaffre (1994), Kenward, Lesaffre and Molenberghs (1994), Molenberghs, Kenward and Lesaffre (1997), Michiels and Molenberghs (1995, 1997), Molenberghs, Michiels and Lipsitz (1999), Michiels, Molenberghs and Lipsitz (1998) and Molenberghs *et al.* (1999). A detailed account is given in Lesaffre, Molenberghs and Dewulf (1996).

Since many cells are empty or sparsely filled, we only use a dichotomized version of the data, where category 0 (no effect) is contrasted with the others (category 1) for both outcomes. 299 patients have a measurement at the first time point, including 224 completers. A summary of the data can be found in Table 2.3, where the counts are given for the patients with respect to their measurements at times 2, 3, and 4. We will use the data in this form in Chapter 4.

In Chapter 6, we will restrict attention to the outcomes at times 2 and 5, also in a dichotomized version. Here, we will also use the covariates. Therefore, 6 patients drop out due to missing covariate levels. This leads to 293 patients, whose results are shown in Table 2.4. For these 293 patients, the age varies from 16 up to 75 years, with a mean of 42.26, and a standard deviation of 13.18. There were 104 men (35%), and 189 patients (65%) had psychiatric antecedents.

We wish to thank Solvay Duphar N.V. for the kind permission to use their data.

Table 2.1: Fluvoxamine Data, Completers. Each cell gives the outcome at the four visits. The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern.

Cell	Side	Ther	Cell	Side	Ther	Cell	Side	Ther	Cell	Side	Ther
0000	84	9	1000	24	31	2000	1	12	3000	2	1
0001	4	0	1001	2	1	2011	0	1	3001	1	0
0010	2	1	1010	1	1	2100	0	22	3100	0	4
0011	1	0	1011	4	0	2101	0	2	3101	0	1
0032	1	0	1013	0	1	2102	0	1	3110	0	1
0100	2	0	1100	22	7	2110	1	12	3111	1	4
0101	1	0	1101	2	1	2111	5	11	3113	0	1
0110	1	1	1102	0	1	2113	0	1	3200	0	6
0111	6	0	1103	0	3	2120	0	1	3210	0	6
0122	0	1	1110	6	6	2131	0	1	3211	0	11
			1111	37	14	2200	0	1	3212	0	1
			1112	2	0	2210	0	3	3220	0	1
			1121	0	1	2211	3	5	3221	0	2
			1122	2	0	2220	0	2	3222	0	3
			1200	0	1	2222	2	3	3232	0	1
			1211	3	2	2233	0	1	3310	0	2
			1221	1	0	2310	0	1	3320	0	2
			1222	0	2	2311	0	2	3321	0	2
			1310	0	1				3322	0	4
			1322	0	1				3333	0	3
tot.	102	12	tot.	106	74	tot.	12	82	tot.	4	56

Total (completers)	224	224

Table 2.2: Fluvoxamine Data, Incomplete Patterns. Each cell gives the outcome at the four visits (a "." indicates the corresponding measurement is missing). The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern.

Cell	Side	Ther	Cell	Side	Ther	Cell	Side	Ther	Cell	Side	Ther
000.	6	2	00..	5	1	0...	9	4	.100	1	0
001.	2	0	01..	2	0	1...	6	6	.101	0	1
010.	1	0	10..	2	2	2...	7	9	...0	1	0
012.	1	0	11..	4	2	3...	9	12	...2	0	1
100.	1	1	12..	2	2				....	14	14
110.	2	1	13..	1	1						
111.	3	3	20..	1	0						
112.	0	1	21..	0	6						
122.	0	1	22..	5	3						
123.	1	0	23..	2	1						
133.	0	1	31..	0	2						
211.	1	1	32..	0	1						
320.	0	1	33..	2	5						
322.	0	3									
333.	0	3									

Total (monotone missingness)	75	75
Total (non-monotone missingness)	16	16

Table 2.3: Fluvoxamine Data. Each cell gives the dichotomized (0 vs. 1/2/3) outcome at the first 3 visits (a "." indicates the corresponding measurement is missing). The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern.

Cell	Side	Ther
000	94	11
001	6	1
010	4	0
011	8	2
100	31	46
101	5	3
110	26	52
111	68	127
00.	5	1
01.	2	0
10.	3	2
11.	16	23
0..	9	4
1..	22	27
Total	299	299

Table 2.4: Fluvoxamine Data. Each cell gives the dichotomized (0 vs. 1/2/3) outcome at the first and last visit (a "." indicates the corresponding measurement is missing). The numbers given under Side (side effect) and Ther (therapeutic effect) are the number of patients with this measurement pattern (those with a missing covariate are indicated separately).

Cell	Side	Ther
00	89	11
01	13	1
10	56 (1)	123 (1)
11	61 (4)	84 (4)
0.	25 (1)	7
1.	49	67 (1)
Total	293 (6)	293 (6)

2.2 The Vorozole Study

The data come from a randomized phase III trial comparing the new potent and selective third generation aromatase inhibitor *vorozole* (VOR) with *megestrol acetate* (MEG) in postmenopausal advanced breast cancer patients.

This study was an open-label, multicenter, parallel group design conducted at 67 North American centers. Patients were randomized to either vorozole (225 patients, 2.5 mg taken once daily) or megestrol acetate (227 patients, 40 mg four times daily). The patient population consisted of postmenopausal patients with histologically confirmed estrogen-receptor positive metastatic breast carcinoma. All 452 randomized patients were followed until disease progression or death. The main objective was to compare the treatment group with respect to response rate while secondary objectives included a comparison relative to duration of response, time to progression, survival, safety, pain relief, performance status and quality of life. In Chapter 8, we will focus on overall quality of life, measured by the total Functional Living Index: Cancer (FLIC) (Schipper, Clinch and McMurray 1984). Precisely, a

higher FLIC score is the more desirable outcome.

Patients underwent screening and for those deemed eligible a detailed examination at baseline (occasion 0) took place. Further measurement occasions were month 1, then from month 2 at bi-monthly intervals until month 44.

The median age was 66 years for VOR, and 67 for MEG, and the means were respectively 65.1 (s.e. 9.8) and 65.6 (s.e. 10.0) years. The mean duration of breast cancer was 6.8 (s.e. 5.4) years for VOR, and 6.9 (s.e. 5.5) years for MEG. The average total FLIC score was 116.3 (s.e. 1.5) for VOR, and 117.1 (s.e. 1.3) for MEG. These total FLIC scores were calculated based on 199 resp. 213 patients. Full details of this study are reported in Goss *et al.* (1998).

Goss *et al.* (1998) analysed the data and found no significant differences: the response rate was 9.7% for VOR, versus 6.8% for MEG (p=0.24); clinical benefit from treatment was demonstrated in 23.5% of VOR-treated patients versus 27.2% of MEG-treated patients (p=0.42). They also analysed FLIC using a two-way ANOVA model with effects for treatment, disease status, as well as their interaction. Again, no significant difference was found.

We wish to thank the Janssen Research Foundation for the kind permission to use their data.

Chapter 3

Missing Data Terminology

In virtually all longitudinal studies missing data arise. Some studies are designed such that the number of measurements per subject is variable or even random. The measurement times themselves can vary across subjects and can be random as well. We term these studies *unbalanced*. In such unbalanced studies it is usually not possible to identify non-response, unless measurement times have been recorded, even for occasions at which no measurement was actually taken. In contrast, in a *balanced* study the number of measurements per subject is fixed and the measurements are usually taken at an approximately common set of occasions. In this situation, missing observations can be identified without ambiguity. For this reason, we will focus attention on missing data in the balanced case. Furthermore, the specific case of *dropout* (i.e., a subject is completely observed until a certain point in time, where after no more measurements are taken) can be handled in the unbalanced case similarly to the balanced case, so restricting to balanced studies does not influence the treatment of dropout.

A potential source of confusion is the fact that in a part of the literature also the random effects are viewed as missing variables, which are then estimated using a generalization of the EM algorithm (Laird and Ware 1982). The use of the EM algorithm is discussed in Section 3.2.1. It ought to be clear that this is *not* the type of missing data envisaged here. We are concerned with missing outcome variables, i.e., measurements that potentially could have been obtained, in contrast with the random effects, which are latent variables.

A missing data formalism is given in Section 3.1, where the missing data terminology, largely due to Rubin (1976) and Little and Rubin (1987), is used as a standard framework to deal with missing data mechanisms and their effect on the analysis. Section 3.2 gives some methods to deal with incomplete data, with emphasis on the EM algorithm and the theory of Multiple Imputation.

3.1 Missing Data

Assume that for subject s in the study a sequence of *measurements* Y_{st} is designed to be measured at occasions $t = 1, \ldots, T$. Recall that we restricted attention to balanced designs, so it is planned to obtain an equal number of measurements per subject, and one furthermore intends to obtain these measurements at approximately the same times. The outcomes are grouped into a vector $\boldsymbol{Y}_s = (Y_{s1}, \ldots, Y_{sT})'$. In addition, for each time t define

$$R_{st} = \begin{cases} 1 & \text{if } Y_{st} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

The *missing data indicators* R_{st} are grouped into a vector $\boldsymbol{R}_s$ which is of the same length as $\boldsymbol{Y}_s$. The process generating $\boldsymbol{R}_s$ is referred to as the missing data process. An hierarchy of missing data patterns can be considered. When missingness is due to attrition, all measurements for a subject from baseline onwards up to a certain measurement time are recorded, after which all data are missing. This pattern is often called a *dropout* pattern. It is then possible to replace the information contained in the vector $\boldsymbol{R}_s$ by a single indicator variable. We will use D_s to indicate the last observed measurement occasion. Attrition is a particular *monotone* pattern of missingness. If at one or more intermittent times an observation is missing, the pattern is called *non-monotone*.

Partition $\boldsymbol{Y}_s$ into two subvectors such that $\boldsymbol{Y}_s^o$ is the vector containing those Y_{st} for which $R_{st} = 1$ and $\boldsymbol{Y}_s^m$ contains the remaining components. These subvectors are referred to as the *observed* and *missing* components respectively. The following terminology is adopted:

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

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

Covariates $\boldsymbol{X}_s$: apart from the outcomes, additional information is measured. This information can be collected before or during the study. The covariate vector is allowed to change for different outcome components t and can include continuous as well as discrete variables. We assume no missing values appear in $\boldsymbol{X}_s$. Methods in case of missing covariates have been explored by several authors (Little 1992, Robins, Rotnitzky and Zhao 1994, Zhao, Lipsitz and Lew 1996).

3.1.1 Categorical Data

In case the measurements $\boldsymbol{Y}$ are categorical, we can represent the data by means of a contingency table. First, one has to split the data with respect to their covariate level. Suppose the study contains subjects with $i = 1, \ldots, N$ different covariate levels. For each subject $s = 1, \ldots, n_i$ within level i, one intends to measure a series of covariate vectors $\boldsymbol{X}_i$ (which are the same for all subjects within this covariate level) and outcomes Y_{ist}, one for each of $t = 1, \ldots, T$ measurement occasions. Each outcome component Y_{ist} can take on c_t distinct values. If continuous covariates are included, n_i will be small and often even equal to one.

For each covariate level i, the complete data $(\boldsymbol{Y}_{is}, \boldsymbol{R}_{is})$, $s = 1, \ldots, n_i$, are grouped in a *contingency table* $\boldsymbol{Z}_i^c$ of dimensions $2 \times \cdots \times 2 \times c_1 \times \cdots \times c_T$. These dimensions are given by the number of possible outcomes: for $\boldsymbol{R}_{st}$ this is 2, and for $\boldsymbol{Y}_{st}$ this is c_t. In case of dropout, the contingency table reduces to a table with dimensions $c_0 \times c_1 \times \cdots \times c_T$, where c_0 is in most cases equal to T. The observed data are not $\boldsymbol{Z}_i^c$ but merely $\boldsymbol{Z}_i$, a partially classified table. These cell counts can be thought of as arising by summing over the appropriate rows or columns in the corresponding complete table. We then have a linear relationship between observed and complete

quantities: $\boldsymbol{Z}_i = C_i \boldsymbol{Z}_i^c$. We call the matrix C_i which consists of 0's and 1's the coarsening matrix in agreement with Molenberghs and Goetghebeur (1997) and Heitjan and Rubin (1991). The corresponding cell-probabilities will be denoted by $\boldsymbol{\nu}_i^c$ for the complete table, and $\boldsymbol{\nu}_i$ for the observed table. Again, the coarsening matrix gives the relation between both: $\boldsymbol{\nu}_i = C_i \boldsymbol{\nu}_i^c$. The multinomial cell probability vector $\boldsymbol{\nu}_i^c$ has entries

$$\nu^c_{ir_1 \ldots r_T k_1 \ldots k_T} = P(R_{is1} = r_1, \ldots, R_{isT} = r_T, Y_{is1} = k_1, \ldots, Y_{isT} = k_T | \boldsymbol{X}_i, \boldsymbol{\theta}), \tag{3.1}$$

with $\boldsymbol{\theta}$ a vector of parameters of interest and $\boldsymbol{X}_i$ the design matrix for the covariate level i, constructed from the covariate information.

3.1.2 Missing Data Mechanisms

Consider the density of the full data

$$f(\boldsymbol{y}_s, \boldsymbol{r}_s | \boldsymbol{X}_s, \boldsymbol{\theta}),$$

where $\boldsymbol{X}_s$ is the design matrix and $\boldsymbol{\theta}$ is a vector that parameterizes the joint distribution. We will use $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ to describe the measurement and missingness processes respectively. The design matrix then splits up in two parts: $\boldsymbol{X}_s^\beta$ and $\boldsymbol{X}_s^\alpha$.

A useful taxonomy, constructed by Rubin (1976) and further developed in Little and Rubin (1987), is based on the factorization:

$$f(\boldsymbol{y}_s, \boldsymbol{r}_s | \boldsymbol{X}_s, \boldsymbol{\theta}) = f(\boldsymbol{y}_s | \boldsymbol{X}_s^\beta, \boldsymbol{\beta}) f(\boldsymbol{r}_s | \boldsymbol{y}_s, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha}), \tag{3.2}$$

where 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. This factorization forms the basis of *selection modelling* as the second factor corresponds to the (self-)selection of individuals into 'observed' and 'missing' groups. An alternative taxonomy can be built based on so-called *pattern-mixture models*. These are based on the factorization

$$f(\boldsymbol{y}_s, \boldsymbol{r}_s | \boldsymbol{X}_s, \boldsymbol{\theta}) = f(\boldsymbol{y}_s | \boldsymbol{r}_s, \boldsymbol{X}_s^\beta, \boldsymbol{\beta}) f(\boldsymbol{r}_s | \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha}). \tag{3.3}$$

Indeed, (3.3) can be seen as a mixture of different populations, characterized by the observed pattern of missingness. After initial mention of these models (Little and Rubin 1987, Glynn, Laird and Rubin 1986), they are receiving more attention lately (Little 1993, 1994a, 1995, Hogan and Laird 1997, Ekholm and Skinner 1998).

In the special case of categorical outcomes, factorization is done on $\boldsymbol{\nu}_i^c$. We will use the following notation: a selection model is given by

$$\nu^c_{ir_1\ldots r_T k_1\ldots k_T}(\boldsymbol{\theta}^S) = \mu^{Sc}_{ik_1\ldots k_T}(\boldsymbol{\beta}^S)\ \phi^{Sc}_{ir_1\ldots r_T|k_1\ldots k_T}(\boldsymbol{\alpha}^S), \tag{3.4}$$

with $\boldsymbol{\theta}^S = ((\boldsymbol{\beta}^S)', (\boldsymbol{\alpha}^S)')'$. So the marginal measurement probabilities $\boldsymbol{\mu}_i^{Sc}$ are given by

$$\mu^{Sc}_{ik_1\ldots k_T}(\boldsymbol{\beta}^S) = P(Y_{is1} = k_1, \ldots, Y_{isT} = k_T | \boldsymbol{X}_i^\beta, \boldsymbol{\beta}^S),$$

and the missingness probabilities, conditional on the outcomes, $\boldsymbol{\phi}_i^{Sc}$ are defined as

$$\phi^{Sc}_{ir_1\ldots r_T|k_1\ldots k_T}(\boldsymbol{\alpha}^S) = P(R_{is1} = r_1, \ldots, R_{isT} = r_T | Y_{is1} = k_1, \ldots, Y_{isT} = k_T, \boldsymbol{X}_i^\alpha, \boldsymbol{\alpha}^S).$$

Alternatively, a pattern-mixture model is based on the factorization

$$\nu^c_{ir_1\ldots r_T k_1\ldots k_T}(\boldsymbol{\theta}^P) = \phi^{Pc}_{ir_1\ldots r_T}(\boldsymbol{\alpha}^P)\ \mu^{Pc}_{ik_1\ldots k_T|r_1\ldots r_T}(\boldsymbol{\beta}^P), \tag{3.5}$$

with $\boldsymbol{\theta}^P = ((\boldsymbol{\beta}^P)', (\boldsymbol{\alpha}^P)')'$. Now, $\boldsymbol{\mu}^{Pc}_{i\cdot|r_1\ldots r_T}$ are the complete data measurement probabilities for response pattern $R_1 = r_1, \ldots, R_T = r_T$:

$$\mu^{Pc}_{ik_1\ldots k_T|r_1\ldots r_T}(\boldsymbol{\beta}^P) = P(Y_{is1} = k_1, \ldots, Y_{isT} = k_T | R_{is1} = r_1, \ldots, R_{isT} = r_T, \boldsymbol{X}_i^\beta, \boldsymbol{\beta}^P).$$

Clearly, they cannot be fully identified and additional assumptions will be required. The missingness probabilities are expressed in an unconditional form:

$$\phi^{Pc}_{ir_1\ldots r_T}(\boldsymbol{\alpha}^P) = P(R_{is1} = r_1, \ldots, R_{isT} = r_T | \boldsymbol{X}_i^\alpha, \boldsymbol{\alpha}^P). \tag{3.6}$$

The natural parameters of selection models and pattern-mixture models have a different meaning, and transforming one probability model into one of the other framework is in general not straightforward, even not for normal measurement models. When a selection model is used, it is often mentioned that one has to make untestable assumptions about the missing data mechanism (discussion of Diggle

and Kenward 1994, Molenberghs, Kenward and Lesaffre 1997). Because in pattern-mixture models different densities (possibly with different parameters) are considered for each of the observed values of $\boldsymbol{R}$, it is explicit which parameters cannot be identified. Little (1993) suggests the use of identifying relationships between identifiable and non-identifiable parameters. Thus, even though these identifying relationships are also unverifiable (Little 1995), the advantage of pattern-mixture models is that the verifiable and unverifiable assumptions can easily be separated.

The assumptions about the missing data mechanism were originally defined for selection models. This classical taxonomy is based on the second factor of (3.2):

$$f(\boldsymbol{r}_s|\boldsymbol{y}_s, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha}) = f(\boldsymbol{r}_s|\boldsymbol{y}_s^o, \boldsymbol{y}_s^m, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha}). \tag{3.7}$$

If (3.7) is independent of the measurements, i.e., when it assumes the form $f(\boldsymbol{r}_s|\boldsymbol{X}_s^\alpha, \boldsymbol{\alpha})$, then the process is termed *missing completely at random* (MCAR).

If (3.7) is independent of the unobserved (missing) measurements $\boldsymbol{Y}_s^m$, but depends on the observed measurements $\boldsymbol{Y}_s^o$, thereby assuming the form $f(\boldsymbol{r}_s|\boldsymbol{y}_s^o, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha})$, then the process is referred to as *missing at random* (MAR).

Finally, when (3.7) depends on the missing values $\boldsymbol{Y}_s^m$, the process is referred to as *informative* missingness or *missing not at random* (MNAR). An informative process is allowed to depend on $\boldsymbol{Y}_s^o$.

It is important to note that the above terminology is independent of the statistical framework chosen to analyse the data. This is to be contrasted with the terms *ignorable* and *non-ignorable* missingness. The latter terms depend crucially on the inferential framework (Rubin 1976).

3.1.3 Ignorability

If one uses likelihood based estimation, ignorability can be defined. The full data likelihood contribution for subject s assumes the form

$$L^*(\boldsymbol{\theta}|\boldsymbol{X}_s, \boldsymbol{y}_s, \boldsymbol{r}_s) \propto f(\boldsymbol{y}_s, \boldsymbol{r}_s|\boldsymbol{X}_s, \boldsymbol{\theta}).$$

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

$$L(\boldsymbol{\theta}|\boldsymbol{X}_s, \boldsymbol{y}_s^o, \boldsymbol{r}_s) \propto f(\boldsymbol{y}_s^o, \boldsymbol{r}_s|\boldsymbol{X}_s, \boldsymbol{\theta})$$

with

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

Under an MAR process, we obtain

$$\begin{aligned} f(\boldsymbol{y}_s^o, \boldsymbol{r}_s|\boldsymbol{X}_s, \boldsymbol{\theta}) &= \int f(\boldsymbol{y}_s^o, \boldsymbol{y}_s^m|\boldsymbol{X}_s^\beta, \boldsymbol{\beta}) f(\boldsymbol{r}_s|\boldsymbol{y}_s^o, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha}) d\boldsymbol{y}_s^m \\ &= f(\boldsymbol{y}_s^o|\boldsymbol{X}_s^\beta, \boldsymbol{\beta}) f(\boldsymbol{r}_s|\boldsymbol{y}_s^o, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha}), \end{aligned} \tag{3.8}$$

i.e., the likelihood factorizes into two components of the same functional form as the general factorization (3.2) of the complete data. If further $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are disjoint in the sense that the parameter space of the full vector $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ is the product of the individual parameter spaces then inference can be based on the marginal observed data density only. This technical requirement is referred to as the separability condition.

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 'informativeness' are synonyms in this context. A formal derivation is given in Rubin (1976), where it is also shown that the same requirements hold for Bayesian inference, but that frequentist inference is ignorable only under MCAR. Of course, ignorability is unhelpful when at least part of the scientific interest is directed towards the missingness process.

Classical examples of the more stringent condition with frequentist methods are ordinary least squares and the generalized estimating equations approach of Liang and Zeger (1986). These GEE define an asymptotically unbiased estimator only under MCAR. Robins, Rotnizky and Zhao (1995) have established that some progress can be made under MAR and even under informative processes. Their method is based on including weights that depend on the missingness probability, proving the point that at least some information on the missingness mechanism should be included and thus that ignorability does not hold.

3.2 Approaches to Incomplete Data

Missing data nearly always entail problems for the practicing statistician. First, inference will often be invalidated when the observed measurements do not constitute a simple random subset of the complete set of measurements. Secondly, even when correct inference would follow, it is not always an easy task to trick standard software into operation on a ragged data structure.

Little and Rubin (1987) give an extensive treatment of methods to analyse incomplete data, many of which are intended for continuous, normally distributed data. Some of these methods were proposed more than fifty years ago. Examples are Yates' (1933) iterated ANOVA and Bartlett's (1937) ANCOVA procedures to analyse incomplete ANOVA designs. The former method is an early example of the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin 1977). This EM-algorithm is discussed in Section 3.2.1.

The computationally simplest technique is a complete case analysis, in which the analysis is restricted to the subjects for whom all intended measurements have been observed. A complete case analysis is popular because it maps a ragged data matrix into a rectangular one, by deleting incomplete cases. An alternative approach, with a similar effect on the applicability of complete data software, is based on imputing missing values. One distinguishes between single imputation and multiple imputation (Rubin 1987). In the first case, a single value is substituted for every 'hole' in the data set and the resulting data set is analysed as if it represented the true complete data. Also in the multiple imputation technique, 'holes' in the data set are filled, but to account for the uncertainty in filling in missing values, the imputation is done multiple times, and each time the complete data are analysed. The theory of multiple imputation is explained in Section 3.2.2.

3.2.1 EM-algorithm

The EM algorithm consists of two components, the *Expectation* and *Maximization* steps. Each step is completed once within each algorithm cycle. Cycles are repeated until a suitable convergence criterion is satisfied. In the expectation step the un-

observed (or missing) data are estimated by their expectations given the observed data and current parameter values. In the maximization step the parameters are estimated using maximum likelihood applied to the observed data augmented by the estimates of the unobserved data. Effectively this maximizes, in each cycle, the expectation of the complete data log likelihood $\mathrm{E}[\log L(\boldsymbol{\theta})]$ where the expectation is taken with respect to the observed data and the current fitted values of $\boldsymbol{\theta}$. Dempster, Laird and Rubin (1977) show that on convergence the fitted parameters are equal to a local maximum of the likelihood function, which is the maximum likelihood estimate in the case of a unique maximum.

Two of the main drawbacks of the EM algorithm are its typically very slow rate of convergence and its lack of direct provision of a measure of precision for the maximum likelihood estimates. Both problems are in fact related and several proposals have been made to overcome them. We mention the technique suggested by Louis (1982), the EM-aided differentation by Meilijson (1989), the "rate matrix" method of Meng and Rubin (1991), and the linear transformation method of Baker (1992). Standard errors and Wald statistics are computed directly from the observed information and score tests are also relatively simple to compute.

We will use Meilijson's (1989) proposal, which is based on the property that the derivative of the complete data score vector coincides with the observed information matrix. It leads to an easy numerical algorithm, using the classical finite differences of the score vector to approximate the derivative. Let the constant that defines the differences be ϵ. To compute the jth column of the information matrix, one changes $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_j$, where all components remain the same, except for the jth one which is changed to $\theta_j + \epsilon$. Then one E step is carried out, yielding $\boldsymbol{Y}(\boldsymbol{\theta}_j)$. Next, the score vector $\boldsymbol{S}_j$ is computed. An approximation for the jth column is given by $(\boldsymbol{S}_j - \boldsymbol{S})/\epsilon$, where $\boldsymbol{S}$ is the score vector at maximum. Replacing all quantities by their estimated values yields a convenient algorithm.

3.2.2 Multiple Imputation

The theory of multiple imputation is presented in Rubin (1987). Several other sources, such as Rubin and Schenker (1986), Little and Rubin (1987), Rubin (1996)

and Schafer (1997), give an excellent account of the technique. As discussed by Rubin and Schenker (1986), the theoretical justification for multiple imputation is most easily understood using Bayesian methodology.

Suppose interest lies in estimating the vector $\boldsymbol{\beta}$, containing the parameters of interest. Rubin (1987) proposed using multiple imputation to "fill-in" the unobserved components of the outcome vectors using the observed data and then use the filled-in data to estimate $\boldsymbol{\beta}$. His method also yields a variance estimator. In order to be able to fill in values, we need the distribution of the missing data, given the observed data and a parameter vector $\boldsymbol{\gamma}$. Multiple imputation is most useful when $\boldsymbol{\gamma}$ is an easily estimated set of parameters, while $\boldsymbol{\beta}$ is complicated to estimate in the presence of missing data.

Recall that the observed data are $\boldsymbol{Y}^o$ and the complete data are $\boldsymbol{Y}$. Multiple imputation uses $\boldsymbol{Y}^o$ to fill in $\boldsymbol{Y}^m$, leading to the complete data $\boldsymbol{Y} = (\boldsymbol{Y}^o, \boldsymbol{Y}^m)$. If we knew the distribution of $\boldsymbol{Y}^m$, with parameter vector $\boldsymbol{\gamma}$, then we could impute $\boldsymbol{Y}^m$ by drawing from the conditional distribution $f(\boldsymbol{Y}^m|\boldsymbol{Y}^o, \boldsymbol{\gamma})$. Since $\boldsymbol{\gamma}$ is unknown, we estimate it from the data, yielding $\widehat{\boldsymbol{\gamma}}$, and use the distribution $f(\boldsymbol{Y}^m|\boldsymbol{Y}^o, \widehat{\boldsymbol{\gamma}})$. Because $\widehat{\boldsymbol{\gamma}}$ is a random variable, we must also take its variability into account in drawing imputations. In Bayesian terms, $\boldsymbol{\gamma}$ is a random variable of which the distribution depends on the data. So we first obtain the posterior distribution of $\boldsymbol{\gamma}$ from the data, a distribution which is a function of $\widehat{\boldsymbol{\gamma}}$.

After formulating the posterior distribution of $\boldsymbol{\gamma}$, we use the following imputation algorithm.

1. Draw $\boldsymbol{\gamma}^*$ from the posterior distribution of $\boldsymbol{\gamma}$, $f(\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{Y}^o)$. We approximate this posterior distribution by a normal.

2. Draw $\boldsymbol{Y}^m$ from $f(\boldsymbol{Y}^m|\boldsymbol{X}, \boldsymbol{Y}^o, \boldsymbol{\gamma}^*)$.

3. Use the completed data $\boldsymbol{Y}$ and the model to estimate the parameter of interest $\boldsymbol{\beta}^*$ and its variance $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*)$, called the within-imputation variance.

These three steps are repeated independently M times, resulting in $\boldsymbol{\beta}^*_m$, $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*_m)$, $m = 1, \ldots, M$.

In case the data to be filled in are categorical, we use a uniform random number generator in step 2 (see Rubin 1987, pp. 169-170). Suppose the count Z is to be distributed over the cells Z_k^c , $k = 1, \ldots, c$. Then, the cumulative probabilities

$$\begin{aligned} \lambda_0 &= 0, \\ \lambda_k &= \frac{\sum_{k'=1}^{k} \nu_{k'}^c}{\nu}, \qquad k = 1, \ldots, c \end{aligned}$$

are calculated and Z draws U_t from a uniform $U[0,1]$ distribution are made. Next, Z_k^c is set equal to $\sum_t(\lambda_{k-1} < U_t \leq \lambda_k)$.

Finally, we combine the estimates obtained after M imputations. The overall estimated parameter vector is the mean of all individual estimates:

$$\boldsymbol{\beta}^* = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\beta}_m^*.$$

The variance is obtained as a weighted sum of the within-imputation variance and the between-imputations variance:

$$\boldsymbol{\Sigma}^* = \boldsymbol{W} + \frac{M+1}{M} \boldsymbol{B}$$

where

$$\boldsymbol{W} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\Sigma}(\boldsymbol{\beta}_m^*),$$

the mean of the within-imputation variances, and

$$\boldsymbol{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\boldsymbol{\beta}_m^* - \boldsymbol{\beta}^*)(\boldsymbol{\beta}_m^* - \boldsymbol{\beta}^*)',$$

the between-imputations variance (Rubin 1987). Based on these variances, one can calculate approximate 95% confidence intervals. Finding an appropriate reference distribution is not an easy matter. Rubin (1987) proposes a multivariate T distribution. Shafer (1997, p. 113) suggests that the approximations by Li, Raghunathan and Rubin (1991) work well in practice. Since in our case the number of imputations will be large, we can certainly rely on the corresponding normal approximation.

Chapter 4

Protective Estimation

For multivariate Gaussian data, Brown (1990) constructs an estimator for the mean and covariance parameters of the joint normal distribution. He assumes that the missingness depends on the unobserved values, but not on the observed measurements, a particular type of MNAR. He called this estimator a protective estimator. Brown (1990) studies an example where measurements are made relatively far apart in time, such that the influence of a previous measurement on nonresponse is negligible. He allows for the presence of all possible response patterns, but each subject is measured at the first occasion. In Section 4.1, the protective estimator as introduced by Brown is given.

In Section 4.2, we will consider repeated categorical measurements, where each subject is observed at the first occasion, and missingness is due to attrition, ruling out nonmonotone patterns. We have constructed a protective estimator in this setting, which can be used both in a selection model, as well as in a pattern-mixture modelling framework (Michiels and Molenberghs 1995, 1997). Estimation of measurement parameters is possible, without explicitly modelling the dropout process. It is well known (see e.g., Baker and Laird 1988) that specific problems arise with non-random missingness models for categorical outcomes. One can be confronted with non-unique, invalid, and/or boundary solutions. In contrast to Brown who only gives necessary conditions for consistent solutions, we will derive a theorem specifying necessary and sufficient conditions for a unique solution in the interior of the parameter space. An intuitive and appealing interpretation of these conditions

is given. An algorithm is presented which consists of the repeated calculation for a bivariate outcome, with the first one always observed and the second one possibly missing. This procedure circumvents working with intractible systems of equations. A connection with direct likelihood estimation is given in Section 4.3, and a link with pseudo-likelihood estimation is established in Section 4.4. Section 4.5 is devoted to variance estimation. The precision estimates will be based on the delta method, the EM algorithm, and on multiple imputation. The relative merits of these techniques are discussed and they are contrasted with the results from likelihood and pseudo-likelihood estimation. Finally, the method is illustrated with two examples in Section 4.6.

4.1 The Protective Estimator for Gaussian Data

Brown (1990) introduced the *protective estimator* for normal data. He called it protective because it is an estimator that retains its consistency over a wide range of non-random missing data mechanisms.

Let $\boldsymbol{Y}$ be a T-dimensional random variable, following a multivariate normal distribution. Let $\boldsymbol{R}$ be a vector of the same dimension as $\boldsymbol{Y}$, indicating the missingness. So, $\boldsymbol{Y}_{st}$ is the tth measure on the sth subject ($s = 1, \ldots, N; t = 1, \ldots, T$). Recall that $\boldsymbol{R}_{st} = 1$ if $\boldsymbol{Y}_{st}$ is observed, and $\boldsymbol{R}_{st} = 0$ if $\boldsymbol{Y}_{st}$ is missing. The joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{R}$ is factorized in a selection modelling way as

$$f(\boldsymbol{y}, \boldsymbol{r}) = n(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\ \omega(\boldsymbol{R} = \boldsymbol{r} | \boldsymbol{Y} = \boldsymbol{y})$$

where $n(.;.)$ is the normal density. There are nearly no conditions on ω, because little is generally known about the missing data mechanism. A GCM (Generalized Censoring Mechanism) is defined as

$$\omega(\boldsymbol{R} = (r_1, \ldots, r_T)' | \boldsymbol{Y} = (y_1, \ldots, y_T)') = \prod_{t=1}^{T} h_t(r_t | y_t)$$

where $h_t(t = 1, \ldots, T)$ are functions bounded between 0 and 1, but without further restrictions.

Under GCM, missingness on each variable depends on that variable alone. Because one usually assumes the first variable to be observed for every subject, we set

h_1 constant. For the other h_t, we know nothing about the form or the conditions we have to put on them.

To obtain estimators for the unknown parameters in the model, Brown uses statistics whose distributions do not depend on the mechanism. It can be proven that this method leads to consistent estimators. To explain the method, we will restrict attention to the case with $T = 3$. Generalization to higher dimensions is straightforward.

The first two moments of Y_1 can be estimated independently from the missing data mechanism, because the first measurement is observed for all subjects. This leads to estimators for μ_1 and σ_{11}. Furthermore, the distributions of $Y_1|(Y_2; R_2 = 1)$, $Y_1|(Y_3; R_3 = 1)$, and $Y_1|(Y_2, Y_3; R_2 = 1, R_3 = 1)$ do not depend on ω (e.g., $f(y_1|(y_2; r_2 = 1)) = n(y_1|y_2; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ are measurement parameters, i.e., functions of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). This leads to the following statistics, which are independent of the mechanism: $\mu_1 - \frac{\sigma_{11}}{\sigma_{22}}\mu_2$; $\frac{\sigma_{12}}{\sigma_{22}}$; $\sigma_{11.2}(= \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}})$.

Under the mild assumption that neither σ_{12} nor σ_{13} equal zero, these statistics lead, using some algebra, to $\mu_2, \sigma_{12}, \sigma_{22}, \mu_3, \sigma_{13}$ and σ_{33}. If one furthermore calculates the partial regression coefficients to predict Y_1:

$$\beta_{12.3} = \frac{\sigma_{12}\sigma_{33} - \sigma_{13}\sigma_{23}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}$$

and

$$\beta_{13.2} = \frac{-\sigma_{12}\sigma_{23} + \sigma_{13}\sigma_{22}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2},$$

the only statistic not yet estimated is given as:

$$\sigma_{23} = \begin{cases} -\frac{\beta_{13.2}\sigma_{12}\sigma_{33} - \beta_{12.3}\sigma_{13}\sigma_{22}}{\beta_{12.3}\sigma_{12} - \beta_{13.2}\sigma_{13}} & \text{if } \rho_{12}^2 \neq \rho_{13}^2 \\ \frac{\sigma_{12}\sigma_{33}}{\sigma_{13}} & \text{if } \rho_{12}^2 = \rho_{13}^2 \neq 0 \text{ and } \beta_{12.3} = 0 \\ \pm\sqrt{\sigma_{22}\sigma_{33}}(\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\frac{\rho_{12}}{\beta_{12.3}} - 1) & \text{if } \rho_{12} = \pm\rho_{13} \text{ and } \beta_{12.3} \neq 0 \end{cases}$$

The next step is to estimate all these statistics, by using the observed data. In the following, the superscripts will indicate whether a measurement is observed (1), missing (0), or marginalized (.). Then the estimates for all the statistics we need

are:

$$\begin{aligned}
\widehat{\mu}_1 &= \bar{Y_1}^{(1..)} \\
\widehat{\sigma}_{11} &= s_{11}^{(1..)} \\
\widehat{\mu}_2 &= \frac{\bar{Y_1}^{(1..)} - \bar{Y_1}^{(11.)}}{b_{12}^{(11.)}} + \bar{Y_2}^{(11.)} \\
\widehat{\sigma}_{22} &= \frac{\widehat{\sigma}_{11} - s_{11.2}^{(11.)}}{(b_{12}^{(11.)})^2} \\
\widehat{\sigma}_{12} &= b_{12}^{(11.)}\widehat{\sigma}_{22} \\
\widehat{\mu}_3 &= \frac{\bar{Y_1}^{(1..)} - \bar{Y_1}^{(1.1)}}{b_{13}^{(1.1)}} + \bar{Y_3}^{(1.1)} \\
\widehat{\sigma}_{33} &= \frac{\widehat{\sigma}_{11} - s_{11.3}^{(1.1)}}{(b_{13}^{(1.1)})^2} \\
\widehat{\sigma}_{13} &= b_{13}^{(1.1)}\widehat{\sigma}_{33}
\end{aligned}$$

We obtain $\widehat{\sigma}_{23}$ by minimizing a residual variance expression for $Y_1|Y_2, Y_3$ in terms of already calculated estimates: define

$$M_{1.23}(y_2, y_3; \sigma_{23}^*) = \widehat{\mu}_1 + \frac{\widehat{\sigma}_{12}\widehat{\sigma}_{33} - \widehat{\sigma}_{13}\sigma_{23}^*}{\widehat{\sigma}_{22}\widehat{\sigma}_{33} - (\sigma_{23}^*)^2}(y_2 - \widehat{\mu}_2) + \frac{\widehat{\sigma}_{13}\widehat{\sigma}_{22} - \widehat{\sigma}_{12}\sigma_{23}^*}{\widehat{\sigma}_{22}\widehat{\sigma}_{33} - (\sigma_{23}^*)^2}(y_3 - \widehat{\mu}_3)$$

Then, $\widehat{\sigma}_{23}$ is that value of σ_{23}^* that minimizes

$$\sum_{s:r_{s1}=r_{s2}=r_{s3}=1} (y_{s1} - M_{1.23}(y_{s2}, y_{s3}; \sigma_{23}^*))^2$$

One can use the steepest descent method to minimize this expression, starting from the sample covariance based on cases where Y_2 and Y_3 are both observed.

Brown also gives necessary conditions to have consistent estimators. In general, these conditions are: $\forall t : \rho_{1t} \neq 0$ and $P(Y_t \text{ is observed}) > 0$. Notice that these conditions are not sufficient.

A drawback of this method is that all missingness patterns (with the first variable observed) are needed to obtain the necessar estimates. If one has a study with dropout only, no complete solution can be found using the protective estimator for normal data.

4.2 The Protective Estimator for Categorical Data

A class of protective estimators for repeated categorical data will be presented next. In contrast to the protective estimator for normal data, in the case of categorical data we have to restrict to dropout to find unique estimates. We will use the notation for categorical data, introduced in Section 3.1.1. Since we do not use covariates in this chapter, and since missingness is restricted to dropout (indicating the time of last observation), the (intended) data, grouped in a contingency table, are $\boldsymbol{Z}^c_{d,k_1\ldots k_T}$, with multinomial cell probabilities

$$\nu^c_{d,k_1\ldots k_T} = P(D_s = d, Y_{s1} = k_1, \ldots, Y_{sT} = k_T). \tag{4.1}$$

Since the same variables are recorded at different times, the outcomes Y_{st} can take on the same number of levels for all times, denoted by r.

We will use the following simplified notation for the selection model and the pattern-mixture model:

$$\nu^c_{d,k_1\ldots k_T} = \mu^c_{k_1\ldots k_T}\ \phi^c_{d|k_1\ldots k_T} \tag{4.2}$$

$$\nu^c_{d,k_1\ldots k_T} = \phi^c_d\ \mu^c_{k_1\ldots k_T|d}. \tag{4.3}$$

Since we only observed the first d measures, the observed data are $Z_{d,k_1\ldots k_d}$. So we can estimate the probabilities

$$\mu^c_{k_1\ldots k_d|d}, \tag{4.4}$$

directly from the data, as well as derived marginal and conditional probabilities. These probabilities are of a pattern-mixture nature and correspond to a "partial classification" of Model (4.3). The aim is to construct the cell probabilities $\mu^c_{k_1\ldots k_T}$, when a selection model is thought appropriate, or $\mu^c_{k_1\ldots k_T|d}$ for pattern-mixture models.

A main assumption for protective estimation is that dropout (possibly) depends on the unobserved outcome, but not on the (previously) observed outcomes. The method used to estimate all the probabilities needed, is based on statistics that are

independent of the missing data mechanism. Factorize the probability $\nu^c_{(D\geq d),k_1k_2\ldots k_t}$ first as

$$\nu^c_{(D\geq d),k_1k_2\ldots k_t} = \nu^c_{(D\geq d),k_2\ldots k_t}\nu^c_{k_1|k_2\ldots k_t;(D\geq d)} \tag{4.5}$$

and alternatively as

$$\begin{aligned}\nu^c_{(D\geq d),k_1k_2\ldots k_t} &= \mu^c_{k_1k_2\ldots k_t}\phi^c_{(D\geq d)|k_1k_2\ldots k_t} \\ &= \mu^c_{k_1k_2\ldots k_t}\phi^c_{(D\geq d)|k_2\ldots k_t} \\ &= \nu^c_{(D\geq d),k_2\ldots k_t}\nu^c_{k_1|k_2\ldots k_t},\end{aligned} \tag{4.6}$$

where we assume, as Brown did, that the first variable is always observed, and hence dropout does not depend on its value. Equating (4.5) and (4.6) yields

$$\nu^c_{k_1|k_2\ldots k_t;(D\geq d)} = \nu^c_{k_1|k_2\ldots k_t}. \tag{4.7}$$

Choosing $t = d$, the left hand side of (4.7) contains the probabilities directly observed through the patterns for which at least d measurements are available; they can be written as $\mu_{k_1|k_2\ldots k_d}$. The right hand side contains the complete data measurement probabilities, marginalized over all dropout patterns, i.e., $\mu^c_{k_1|k_2\ldots k_d}$, and hence (4.7) can be rewritten as $\mu^c_{k_1|k_2\ldots k_d} = \mu_{k_1|k_2\ldots k_d}$. This result means that the conditional probability of the first outcome, given the outcomes on the $d-1$ measures that follow can be estimated directly through the observed data.

Let us summarize the relevant quantities that are directly available from the data. First, table d yields $\mu^c_{k_1\ldots k_d|d}$. Secondly, tables d through T provide the conditional probabilities

$$\mu^c_{k_1|k_2\ldots k_d}. \tag{4.8}$$

In particular, all tables contribute to $\mu^c_{k_1}$, whereas only the last table contributes to $\mu^c_{k_1|k_2\ldots k_T}$. The next step is the computation of $\mu^c_{k_1\ldots k_d}$, for all d. Denote by $\mu_{k_1\ldots k_d}$ the directly observed cell probabilities, estimated from tables d through T.

For $d = 1$, $\mu^c_{k_1}$ follows from the data. Let us first consider the construction of $\mu^c_{k_1k_2}$. Recall that neither the observed $\mu_{k_1k_2}$, nor the $\mu_{k_1k_2|d}$ for $d \geq 2$ are of

direct use; they only contribute through (4.8) by estimating $\mu^c_{k_1|k_2}$. Then, $\mu^c_{k_1k_2}$ is determined by solving the system of equations

$$\sum_{k_2=1}^{r} \mu^c_{k_1|k_2}\mu^c_{k_2} = \mu^c_{k_1}, \qquad k_1 = 1, \dots r, \tag{4.9}$$

where $\mu^c_{k_1|k_2}$ act as coefficients and $\mu^c_{k_2}$ as unknowns. Solving this system yields $\mu^c_{k_2}$, whereafter $\mu^c_{k_1k_2}$ is obtained by a simple multiplication. Writing (4.9) as $M_{1|2}M_2 = M_1$, it clearly follows that a unique solution can be found if and only if the determinant of the matrix $M_{1|2}$ is nonzero. This is equivalent with $\det(M_{12}) = \det(\mu_{k_1k_2}) \neq 0$. Further, in order to obtain a valid solution, one has to guarantee that all components of M_2 are nonnegative. Necessary and sufficient conditions are given in Theorem 1.

We now proceed by induction. Suppose that we have constructed all marginal probabilities, up to order $d-1$. We will construct $\mu^c_{k_1\dots k_d}$. For a fixed multi-index $(k_2, \dots, k_{d-1})$, consider the system of equations

$$\sum_{k_d=1}^{r} \mu^c_{k_1|k_2\dots k_d}\mu^c_{k_d|k_2\dots k_{d-1}} = \mu^c_{k_1|k_2\dots k_{d-1}}, \qquad k_1 = 1, \dots r. \tag{4.10}$$

Solving this system yields $\mu^c_{k_d|k_2\dots k_{d-1}}$ and hence $\mu^c_{k_1k_d|k_2\dots k_{d-1}}$, resulting in

$$\mu^c_{k_1k_2\dots k_{d-1}k_d} = \mu^c_{k_1k_2\dots k_{d-1}} \frac{\mu^c_{k_1k_d|k_2\dots k_{d-1}}}{\mu^c_{k_1|k_2\dots k_{d-1}}}; \tag{4.11}$$

all quantities on the right hand side being determined. Writing (4.10) as

$$M^{(k_2\dots k_{d-1})}_{1|d} M^{(k_2\dots k_{d-1})}_{d} = M^{(k_2\dots k_{d-1})}_{1}, \tag{4.12}$$

we obtain a family of systems of equations, one for each combination $(k_2, \dots, k_{d-1})$. Each one of these systems is exactly of the form (4.9). We now state the conditions for a valid solution.

Theorem 1 *The System of Equations (4.10) has a unique, valid (i.e., nonnegative) solution if and only if*

1. $\det(\mu_{k_1\dots k_d}) \neq 0$, *for* $k_2, \dots, k_{d-1}$ *fixed;*
2. *the column vector* $M^{(k_2\dots k_{d-1})}_1 = (\mu_{k_1|k_2\dots k_{d-1}})_{k_1}$ *is an element of the convex hull of the* r *column vectors, indexed by* k_d, $\left((\mu_{k_1|k_d;k_2\dots k_{d-1}})_{k_1}\right)$.

Note that the vectors $\left((\mu_{k_1|k_d;k_2\ldots k_{d-1}})_{k_1}\right)$ are the columns of $M_{1|d}^{(k_2\ldots k_{d-1})}$. By requiring that this theorem holds for all $d = 2,\ldots T$ and for all $1 \leq k_2,\ldots,k_{d-1} \leq r$, one ensures the existence of an overall solution.

Proof of Theorem 1

Clearly, the matrix equation has a unique root if and only if $\det M_{1|d}^{(k_2\ldots k_{d-1})} \neq 0$, where only k_1 and k_d are free indexes. Whether or not this determinant is zero is not altered by multiplying each column of the matrix with $\mu_{k_d|k_1\ldots k_{d-1}}$. Equivalently, one can construct the determinant of the two-way table, obtained from the observed probabilities $\mu_{k_1\ldots k_d}$ by keeping $k_2,\ldots,k_{d-1}$ fixed.

We need to establish that the solution is nonnegative if and only if the column on the right hand side is an element of the convex hull of the columns of the matrix $M_{1|d}^{(k_2\ldots k_{d-1})}$. However, the convex hull is formed by all elements for which there exists a linear combination of which all coefficients are in the unit interval and summing to 1, i.e., a vector of probabilities. □

These conditions can be interpreted as follows. The model implies that the column distributions in both tables are the same, as indicated by (4.7). Therefore, the single column we observe on the right hand side of (4.10) must be a convex linear combination of the set of column distributions we observe on the left hand side. An interesting consequence is that a negative solution points to a violation of the assumptions: the data can contradict the model, even without applying a model checking procedure. A nonzero determinant is equivalent to a set of column distributions which is of full rank. We could call this condition the "full association" condition. It is worthwile to observe that in the case of binary outcomes ($r = 2$) these conditions reduce to: (1) the odds ratio of the table on the left hand side is different from 1, (2) the (marginal) odds of the only column on the right hand side lies in the interval bounded by the two (conditional) odds of the columns on the left hand side. For the binary case, these conditions were spelled out by Baker and Laird (1988, p.67).

The algorithm presented here is not the only method to find the cell probabilities. By requiring that the cell probabilities $\mu^c_{k_1\ldots k_T}$ sum to one and have (4.8)

as conditionals, one is able to construct a single system of r^T equations in r^T unknowns. Although appealing at first sight, the procedure advocated in this section has several important advantages. First, a potentially complex procedure is broken into a sequence of simple, identical procedures, for which the validity requirements are readily verified. No large matrices have to be inverted. If the method yields a non-valid solution, one can at least compute the estimator for a subset of the outcomes, e.g., the first $t-1$ outcomes, when the first problem occurs at variable t. By removing the "problematic" variables, one can compute the estimator for a maximal subset.

4.3 Likelihood Estimation

The protective estimator for two measurements can also be derived through likelihood estimation, by considering a saturated measurement model and a dropout model which only depends on the unobserved outcome, i.e., the likelihood based on the factorization $\nu^c_{d,k_1k_2} = \mu^c_{k_1k_2}\ \phi^c_{d|k_2}$. This model saturates the degrees of freedom, available in the data. Maximizing the likelihood

$$L \propto \prod_{k_1,k_2} (\mu^c_{k_1k_2}\ \phi^c_{2|k_2})^{z_{2,k_1k_2}} \prod_{k_1} (\mu^c_{k_11}\ \phi^c_{1|1} + \mu^c_{k_12}\ \phi^c_{1|2})^{z_{1,k_1}} \tag{4.13}$$

yields the protective estimator which is also equal to the estimator for the model (Y_1Y_2, Y_2R), discussed in Baker and Laird (1988; see also their Equation 2.1). The likelihood estimator is seemingly attractive in allowing for flexible dependence on outcomes and covariates, but explicit and often untestable assumptions about the dropout process have to be made. The protective estimator leaves the dropout model unspecified (non-parametric) and thus uses less degrees of freedom.

An explicit solution to (4.13) can be derived:

$$\begin{aligned}
\hat{\mu}^c_{11} &= \frac{z_{2,11}}{z_{+,++}}\left(\frac{(z_{1,1}+z_{2,11})z_{2,22}-(z_{1,2}+z_{2,21})z_{2,12}}{z_{2,11}z_{2,22}-z_{2,12}z_{2,21}}\right),\\
\hat{\mu}^c_{12} &= \frac{z_{2,12}}{z_{+,++}}\left(\frac{(z_{1,2}+z_{2,22})z_{2,11}-(z_{1,1}+z_{2,12})z_{2,21}}{z_{2,11}z_{2,22}-z_{2,12}z_{2,21}}\right),\\
\hat{\mu}^c_{21} &= \frac{z_{2,21}}{z_{+,++}}\left(\frac{(z_{1,1}+z_{2,11})z_{2,22}-(z_{1,2}+z_{2,21})z_{2,12}}{z_{2,11}z_{2,22}-z_{2,12}z_{2,21}}\right),\\
\hat{\mu}^c_{22} &= \frac{z_{2,22}}{z_{+,++}}\left(\frac{(z_{1,2}+z_{2,22})z_{2,11}-(z_{1,1}+z_{2,12})z_{2,21}}{z_{2,11}z_{2,22}-z_{2,12}z_{2,21}}\right),\\
\hat{\phi}^c_{2|1} &= \frac{z_{2,11}z_{2,22}-z_{2,12}z_{2,21}}{(z_{1,1}+z_{2,11})z_{2,22}-(z_{1,2}+z_{2,21})z_{2,12}},\\
\hat{\phi}^c_{2|2} &= \frac{z_{2,11}z_{2,22}-z_{2,12}z_{2,21}}{(z_{1,2}+z_{2,22})z_{2,11}-(z_{1,1}+z_{2,12})z_{2,21}}.
\end{aligned}$$

The first four estimates are identical to the protective estimator. Therefore, studying likelihood (4.13) can shed some light on situations where boundary restrictions are violated. Table 4.1 presents four artificial sets of data. The first one satisfies the protective assumptions as the odds in the incomplete table are 1, well between 0.5 and 2. Table 4.1 (b) presents data on the boundary, whereas the data in Table 4.1 (c) violate the protective assumption. Finally, Table 4.1 (d) is included to discuss model fit. Parameter estimates and standard errors are presented in the first column of Table 4.2.

In all four cases, parameter estimates coincide with those found by the protective estimator (third column). For the data in Table 4.1 (b) we estimate $\hat{\phi}_{2|2} = 1.0$, implying that no observations from the second column drop out. In the third example, the violation of the restrictions shows in the estimate for $\phi_{2|2}$. Note that the cell probabilities give no direct hint on parameter space violations. However, the probabilities predicted for the incomplete table are $\mu_{1,11} = 0.8333$, $\mu_{1,12} = -0.0833$, $\mu_{1,21} = 0.4167$, $\mu_{1,22} = -0.1667$. The latter probabilities are found with a pattern-mixture formulation of the protective estimator. The log-likelihood is -423.94.

To avoid parameter space violations, one can reparametrize the probabilities as

Table 4.1: Four Sets of Artificial Data. Each time a contingency table for the completers ($Y_1 = 1/2, Y_2 = 1/2$), and an additional contingency table for the dropouts ($Y_1 = 1/2$) is given.

(a)	50	25	25
	25	50	25
(b)	50	25	50
	25	50	25
(c)	50	25	75
	25	50	25
(d)	80	10	60
	40	20	90

follows:

$$\mu^c_{k_1 k_2} = \exp\left[\alpha_1(k_1 - 1) + \alpha_2(k_2 - 1) + \alpha_3(k_1 - 1)(k_2 - 1) - A(\alpha_1, \alpha_2, \alpha_3)\right] \tag{4.14}$$

where $A(\alpha_1, \alpha_2, \alpha_3)$ is a normalizing constant, and

$$\phi_{2|k_2} = \frac{\exp(\gamma_{k_2})}{1 + \exp(\gamma_{k_2})}. \tag{4.15}$$

In this case, the parameters corresponding to the data in Table 4.1 (b) are $\alpha_1 = 0.6932$, $\alpha_2 = 0$, $\alpha_3 = 1.3863$, $\gamma_1 = 0$, and γ_2 approaches infinity. The log-likelihood at maximum is -390.40, independent of whether the untransformed or transformed likelihoods are used.

When we switch to Table 4.1 (c) and again adopt parametrization (4.14) and (4.15), the log-likelihood at maximum becomes -424.66, obtained for parameters

Table 4.2: Parameter Estimates (Standard Errors) for the Artificial Data from Table 4.1. Methods of Estimation Are: Likelihood (Untransformed and Transformed), Protective Estimation With Delta Method, EM Algorithm, and Multiple Imputation.

		Likelihood		Protective		
Data	Par.	Untr.	Transf.	Delta	EM	MI
(a)	μ_{11}	0.33(0.05)	0.33(0.05)	0.33(0.06)	0.33(0.05)	0.33(0.05)
	μ_{12}	0.17(0.04)	0.17(0.04)	0.17(0.04)	0.17(0.05)	0.17(0.04)
	μ_{21}	0.17(0.04)	0.17(0.04)	0.17(0.04)	0.17(0.04)	0.17(0.04)
	μ_{22}	0.33(0.05)	0.33(0.05)	0.33(0.06)	0.33(0.05)	0.33(0.05)
	$\phi_{2\|1}$	0.75(0.10)	0.75		0.75(0.10)	
	$\phi_{2\|2}$	0.75(0.10)	0.75		0.75(0.10)	
(b)	μ_{11}	0.44(0.04)	0.44(0.04)	0.44(0.06)	0.44(0.04)	0.43(0.04)
	μ_{12}	0.11(0.03)	0.11(0.02)	0.11(0.04)	0.11(0.04)	0.12(0.03)
	μ_{21}	0.22(0.06)	0.22(0.03)	0.22(0.03)	0.22(0.06)	0.21(0.04)
	μ_{22}	0.22(0.06)	0.22(0.03)	0.22(0.05)	0.22(0.06)	0.24(0.04)
	$\phi_{2\|1}$	0.50(0.07)	0.50		0.50(0.07)	
	$\phi_{2\|2}$	1.00(0.23)	1.00		1.00(0.23)	
(c)	μ_{11}	0.53(0.04)	0.50(0.03)	0.53(0.05)	0.53(0.04)	0.50(0.03)
	μ_{12}	0.07(0.03)	0.10(0.02)	0.07(0.05)	0.07(0.03)	0.10(0.02)
	μ_{21}	0.27(0.07)	0.20(0.02)	0.27(0.02)	0.27(0.07)	0.20(0.03)
	μ_{22}	0.13(0.07)	0.20(0.02)	0.13(0.03)	0.13(0.07)	0.20(0.03)
	$\phi_{2\|1}$	0.38(0.06)	0.43		0.38(0.06)	
	$\phi_{2\|2}$	1.50(0.47)	1.00		1.50(0.75)	
(d)	μ_{11}	0.33(0.08)	0.33(0.08)	0.33(0.03)	0.33(0.08)	0.33(0.06)
	μ_{12}	0.17(0.08)	0.17(0.08)	0.17(0.03)	0.17(0.08)	0.17(0.06)
	μ_{21}	0.17(0.05)	0.17(0.05)	0.17(0.05)	0.17(0.05)	0.17(0.04)
	μ_{22}	0.33(0.05)	0.33(0.05)	0.33(0.05)	0.33(0.05)	0.33(0.04)
	$\phi_{2\|1}$	0.80(0.19)	0.80		0.80(0.19)	
	$\phi_{2\|2}$	0.20(0.06)	0.20		0.20(0.06)	

$\alpha_1 = 0.6932$, $\alpha_2 = 0$, $\alpha_3 = 1.6095$, $\gamma_1 = -0.2777$, while γ_2 approaches infinity. These values correspond to dropout probabilities $\phi_{2|1} = 0.4286$ and $\phi_{2|2} = 1.000$. This is a slight decrease of the log-likelihood, but all estimated probabilities are valid. Corresponding marginal probabilities are given in the second column of Table 4.2. They are further divided over completers: $\nu^c_{2,11} = 0.2143$, $\nu^c_{2,12} = 0.1000$, $\nu^c_{2,21} = 0.0857$, $\nu^c_{2,22} = 0.2000$, and incomplete observations: $\nu^c_{1,11} = 0.2857$, $\nu^c_{1,12} = 0.0000$, $\nu^c_{1,21} = 0.1143$, and $\nu^c_{1,22} = 0.0000$. A boundary solution is obtained, where all marginal probabilities are as observed, but the association as observed for the completers differs from the estimated association (consider e.g., the odds ratio). The differences in standard errors are discussed in Section 4.5.

Table 4.1 (d) is generated by first assuming that $Z_{11} = Z_{22} = 100$ and $Z_{12} = Z_{21} = 50$ and also that $\phi_{2|1} = 0.8$ and $\phi_{2|2} = 0.2$. This implies that the completers' table is as in Table 4.1 (d), that the dropouts' full data are $Z_{1,11} = 20$, $Z_{1,12} = 40$, $Z_{1,21} = 10$, $Z_{1,22} = 80$ and thus that the supplemental margin consists of the counts 60 and 90, as shown in Table 4.1 (d). Fitting a protective model to those data with all five methods yields exactly the same estimates. When these parameter estimates are used to construct fitted frequencies, then $Z_{1,jk}$ and $Z_{2,jk}$ (j,k = 1,2) are recovered. The log-likelihood at maximum is -479.43. As an alternative to maximizing likelihood (4.13) let us turn to the MAR likelihood

$$\begin{aligned} L &\propto \prod_{k_1,k_2} (\mu^c_{k_1k_2}\ \phi^c_{2|k_1})^{z_{2,k_1k_2}} \prod_{k_1} (\mu^c_{k_11}\ \phi^c_{1|k_1} + \mu^c_{k_12}\ \phi^c_{1|k_1})^{z_{1,k_1}} \\ &\propto \prod_{k_1,k_2} (\mu^c_{k_1k_2})^{z_{2,k_1k_2}} \prod_{k_1} (\mu^c_{k_1+})^{z_{1,k_1}} \prod_{k_1} (\phi^c_{2|k_1})^{z_{2,k_1+}} \prod_{k_1} (\phi^c_{1|k_1})^{z_{1,k_1+}}. \end{aligned}$$

Then, the fitted probabilities (with standard errors) are $\hat{\mu}_{11} = 0.44(0.03)$, $\hat{\mu}_{12} = 0.06(0.02)$, $\hat{\mu}_{21} = 0.33(0.04)$ and $\hat{\mu}_{22} = 0.17(0.03)$. The dropout probabilities are $\hat{\phi}_{2|k_1=1} = 0.60(0.04)$ and $\hat{\phi}_{2|k_1=2} = 0.40(0.04)$. While the completers' table is recovered, the filled-in dropout table is estimated to be $\hat{Z}_{1,11} = 53.33$, $\hat{Z}_{1,12} = 6.67$, $\hat{Z}_{1,21} = 60.00$ and $\hat{Z}_{1,22} = 30.00$, entirely different from the true underlying structure. However, the value of the maximized log-likelihood is still -479.43. This points to a general problem with missing data: it is often impossible or at least very difficult to establish superiority of one dropout model over another, based on statistical considerations alone. Indeed, whereas the MAR and protective models both fit the

observed data perfectly, they yield entirely different predictors for the underlying dropout table. Arguably, background and/or covariate information should be used to support the model builder's task.

4.4 Pseudo-Likelihood Estimation

While it is not straightforward to generalize the likelihood method, discussed in the previous section, to more than 2 measurement occasions, one can proceed alternatively by means of pseudo-likelihood estimation (Arnold and Strauss 1988). Pseudo-likelihood has been used in the context of spatial data by Cressie (1991) and for correlated binary data by Le Cessie and Van Houwelingen (1994).

Let us concentrate on three measurement occasions. The protective estimator first determines $\mu^c_{k_1k_2}$ and then $\mu^c_{k_1k_3|k_2}$. Each of these steps can be handled by means of likelihood (4.13). This leads naturally to a new expression

$$\begin{aligned} L^* \quad \propto \quad & \left[\prod_{k_1,k_2} (\mu^c_{k_1k_2}\ \phi^c_{2|k_2})^{z_{k_1k_2}} \prod_{k_1} (\sum_{k_2} \mu^c_{k_1k_2}\ \phi^c_{1|k_2})^{z_{k_1}} \right] . \\ & \prod_{k_2} \left[\prod_{k_1,k_3} (\mu^c_{k_1k_3|k_2}\ \phi^{(k_2)c}_{2|k_3})^{z_{k_1k_3|k_2}} \prod_{k_1} (\sum_{k_3} \mu^c_{k_1k_3|k_2}\ \phi^{(k_2)c}_{1|k_3})^{z_{k_1|k_2}} \right] \end{aligned} \tag{4.16}$$

which is merely the product of three components. All counts used in (4.16) are based on the maximal amount of information, except for z_{k_1}, which is calculated using the subjects being observed at the first time point only. This is clearly not a likelihood since the factors are incorrectly assumed to be independent.

While the point estimator is consistent, one has to be careful with estimating the precision (Arnold and Strauss 1988, Geys, Molenberghs and Ryan 1997). In particular, let

$$\boldsymbol{\lambda} = \{\mu^c_{k_1k_2}, \mu^c_{k_1k_3|k_2}, \phi^c_{1|k_2}, \phi^c_{2|k_2}, \phi^{(k_2)c}_{1|k_3}, \phi^{(k_2)c}_{2|k_3}; \text{for all } k_1, k_2, k_3\}$$

and

$$l(\boldsymbol{\lambda}) = \sum_{i=1}^{N} l_i(\boldsymbol{\lambda}) = \ln(L^*(\boldsymbol{\lambda}))$$

where $l_i(\boldsymbol{\lambda})$ is the contribution of subject i to the log pseudo-likelihood and N is the total sample size. Then, $\sqrt{N}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ converges in distribution to the normal

distribution $N(\mathbf{0}, \mathbf{J}(\boldsymbol{\lambda})^{-1}\mathbf{K}(\boldsymbol{\lambda})\mathbf{J}(\boldsymbol{\lambda})^{-1})$ where

$$\mathbf{J}(\boldsymbol{\lambda}) = \mathbf{E}\left(\frac{\partial^2 \mathbf{l}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'}\right)$$

and

$$\mathbf{K}(\boldsymbol{\lambda}) = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{N}} \mathbf{E}\left(\frac{\partial \mathbf{l_i}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}(\frac{\partial \mathbf{l_i}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}})'\right).$$

This result is very close in spirit to the sandwich estimator, known from generalized estimating equations (Liang and Zeger 1986). It yields an asymptotic measure of precision for $\mu^c_{k_1k_2}$ and $\mu^c_{k_1k_3|k_2}$ from which the precision for $\mu^c_{k_1k_2k_3}$ easily follows using a standard delta method argument. Precision for the dropout probabilities is also available. The method will be contrasted with different methods for variance calculation for the protective estimator, described in the next section.

4.5 Variance Estimation

In order to complete the protective estimation procedure, we have to compute an estimate of the variance. Whereas estimating the variance under an MAR mechanism is fairly straightforward (Little and Rubin 1987), it is much more involved under non-random nonresponse. In our case, we are restricted by the fact that the dropout parameters are not estimated. Brown (1990) does not outline a way to estimate the variance of the parameters. We will discuss three procedures. The first one is based on the delta method (see e.g., Agresti 1990). The second one uses EM aided differentiation (see Section 3.2.1). The final technique makes use of multiple imputation (see Section 3.2.2). At the end of this section, the artificial examples considered in Section 4.3 will be revisited.

4.5.1 Delta Method

Throughout, the superscripts c will be dropped from the probabilities. Write (4.12) as $M_{1|d}M_d = M_1$. As M_1 and $M_{1|d}$ are directly observed from the data, their covariance matrices are immediate. Let $V(.)$ indicate the covariance function. First, M_1, the vector of probabilities μ_{k_1}, has multinomial covariance matrix

$$V(M_1) = \frac{1}{n_1}(\text{diag}(M_1) - M_1M_1'), \tag{4.17}$$

n_1 indicating the sample size. In order to conveniently work with the matrix $M_{1|d}$, a matrix of which the columns represent independent multinomial distributions, we introduce some extra notation. The components are $\mu_{k_1|k_d}$, while column k_d (corresponding to the k_dth conditional distribution) is denoted by $\mu_{.|k_d}$. The covariance matrix $V(M_{1|d})$ is block diagonal with blocks

$$V(M_{1|d})(k_d) = \frac{1}{n_{1|k_d}}((\text{diag}(\mu_{.|k_d}) - \mu_{.|k_d}\mu'_{.|k_d}). \tag{4.18}$$

It is convenient to write the components of M_1 as $\mu_{k_1|0}$. Similarly, $\mu_{.|0}$ is defined. Then (4.18) encompasses (4.17) by letting $k_d = 0, 1, \ldots, r$. Thus, in the remainder, k_d will be allowed to assume the value 0 also, which should be read as "unconditional". From these covariance matrices and the observation that $M_d = M_{1|d}^{-1}M_1$, the covariance $V(M_d)$ is obtained by applying the delta method. Note that M_d is a vector valued function, of which the arguments are a nonredundant set of components of M_1 and $M_{1|d}$. The redundancies are given by the following identities:

$$g_{k_d} = \sum_{k_1=1}^{r} \mu_{k_1|k_d} - 1 = 0, \quad k_d = 0, 1, \ldots, r. \tag{4.19}$$

A possible nonredundant set is given by the first $r-1$ components of each probability vector. Let us denote these sets by $\tilde{M}_1$ and $\tilde{M}_{1|d}$, with similar notation for the vectors $\tilde{\mu}_{.|k_d}$. Grouping the vectors $\tilde{\mu}_{.|k_d}$ into a vector $\boldsymbol{T}$ and the remaining $\mu_{r|k_d}$ into $\boldsymbol{Y}$, we can express the total derivative of M_d w.r.t. $\boldsymbol{T}$ as

$$\frac{dM_d}{d\boldsymbol{T}} = \frac{\partial M_d}{\partial \boldsymbol{T}} - \frac{\partial M_d}{\partial \boldsymbol{Y}} \left(\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{Y}}\right)^{-1} \frac{\partial \boldsymbol{G}}{\partial \boldsymbol{T}}$$

where $\boldsymbol{G}$ is the set of functions, described by (4.19). It follows immediately that

$$\begin{aligned}
\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{Y}} &= I_{r+1}, \\
\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{T}} &= I_{r+1} \otimes \mathbf{1}_{1,r-1}, \\
\frac{\partial M_d}{\partial M_1} &= M_{1|d}^{-1}, \\
\frac{\partial M_d}{\partial \mu_{k_1|k_d}} &= -M_{1|d}^{-1} E_{k_1 k_d} M_d, \quad k_1, k_d = 1, \ldots, r,
\end{aligned}$$

where $E_{k_1 k_d}$ is a zero matrix, except for a single 1 in entry (k_1, k_d). Using these expressions, we obtain

$$\Phi = \frac{dM_d}{d\boldsymbol{T}} = (\Phi_1, -M_{1|d}^{-1}\Phi_2)$$

where Φ_1 is an $r \times (r-1)$ matrix of which the columns are $\xi_{.|k_d} - \xi_{.|r}$ $(k_d = 1, \ldots, r-1)$, $\xi_{.|k_d}$ are the columns of $M_{1|d}^{-1}$, and $\Phi_2 = (\Phi_{21}, \ldots, \Phi_{2r})$ with

$$\Phi_{2k_d} = \begin{pmatrix} \mu_{k_d} I_{r-1} \\ -\mu_{k_d} \mathbf{1}_{1,r-1} \end{pmatrix}.$$

Finally, $V(M_d) = \Phi W \Phi'$, with

$$W = \text{cov}(\boldsymbol{T}) = \left(\begin{array}{c|c} V(\tilde{M}_1) & V_{1,1|d} \\ \hline V'_{1,1|d} & V(\tilde{M}_{1|d}) \end{array}\right).$$

The matrix $V(\tilde{M}_1)$ is found from $V(M_1)$ by omitting the last row and column. Replacing $\mu_{.|k_d}$ by $\tilde{\mu}_{.|k_d}$ in (4.18) yields $V(\tilde{M}_{1|d})$. Finally, $V_{1,1|d}$ describes the covariance function of M_1 and $M_{1|d}$. A typical element is

$$\text{cov}(\mu_{k_1}, \mu_{k'_1|k_d}) = \frac{1}{n_{12}} \mu_{k'_1|k_d} (\delta_{k_1 k'_1} - \mu_{k_1}),$$

$k_1, k'_1 = 1, \ldots, r-1; k_d = 1, \ldots, r$, with n_{12} the number of subjects, observed at both occasions.

With similar computations, the covariance matrix of M_{1d} (with this matrix appropriately vectorized) is found to be

$$V(\text{vec}(M_{1d})) = \Phi_3 V(\tilde{\mu}_{.|k_d}) \Phi'_3 + \text{bdiag}_{k_d=1}^{r} (\Phi_{2k_d} V(\tilde{\mu}_{.|k_d}) \Phi'_{2k_d})$$

with Φ_3 given by

$$\Phi_3 = \begin{pmatrix} \mu_{.|1} & & & \\ & \mu_{.|2} & & \\ & & \ddots & \\ & & & \mu_{.|r-1} \\ -\mu_{.|r} & \ldots & & -\mu_{.|r} \end{pmatrix}.$$

The notation "bdiag" stands for a block diagonal matrix with blocks given by the indexed argument.

Probabilities of the form $\mu^c_{k_1 \ldots k_d}$ can be written as a product of probabilities that have been determined:

$$\mu^c_{k_1 \ldots k_d} = \mu^c_{k_1 k_d | k_2 \ldots k_{d-1}} \prod_{t=2}^{d-1} \mu^c_{k_t | k_2 \ldots k_{t-1}}. \tag{4.20}$$

Deriving a variance estimator from this expression is straightforward. One only has to take into account that in each matrix (set of probabilities) a sum constraint applies. This fact needs to be discounted in computing the derivatives.

4.5.2 EM Aided Differentiation

The computations based on the delta method are certainly involved, due to the fact that a linear system of equations needs to be solved. We will show that a computational scheme, based on the EM algorithm (see Section 3.2.1), is useful to circumvent this step. The price to pay is that, although the dropout parameters are not necessary for estimating the model parameters, they are required for variance estimation. Suppose we want to estimate $\mu^c_{k_1 k_d | k_2 \ldots k_{d-1}}$. Without loss of generality, we will describe the algorithm for $\mu_{k_1 k_2}$ (setting $d = 2$ and dropping the superscript). Information is based on two tables: those observed at both occasions, summarized in table $Z_{2,k_1 k_2}$ and those observed at the first occasion only: Z_{1,k_1}.

Choosing starting values $\mu^{(0)}_{k_1 k_2}$, one iterates between the E step and the M step until convergence. The E step first calculates probabilities $\mu^{(t)}_{k_1 | k_2}$, given $\mu^{(t)}_{k_1 k_2}$. Together with the probabilities μ_{1,k_1}, directly computed from the incomplete table Z_{1,k_1}, the probabilities $\mu^{(t)}_{1,k_1 k_2}$ are found by solving

$$\sum_{k_2=1}^{r} \mu^{(t)}_{k_1 | k_2} \mu^{(t)}_{1,k_2} = \mu_{1,k_1}, \tag{4.21}$$

and hence $\mu^{(t)}_{1,k_1 k_2} = \mu^{(t)}_{1,k_2} \mu^{(t)}_{k_1 | k_2}$. From these probabilities and the observed data Z_{1,k_1} the expected counts in the completed table $Z^{(t)}_{1,k_1 k_2}$ are readily found. Note that the dropout probabilities are implicitly determined since

$$\phi_{2|k_1 k_2} = \phi_{2|k_2} = \frac{\mu_{2,k_1 k_2}}{\mu_{k_1 k_2}}.$$

The M step merely sums over both tables $Z^{(t)}_{k_1k_2} = Z^{(t)}_{1,k_1k_2} + Z_{2,k_1k_2}$ and determines an update for the probabilities:

$$\mu^{(t+1)}_{k_1k_2} = \frac{Z^{(t)}_{k_1k_2}}{Z_{++}}. \tag{4.22}$$

Observe the strong connection with the likelihood approach. Indeed, expression (4.22) maximizes the complete data log-likelihood

$$\ell^{(t)} = \sum_{k_1,k_2} Z^{(t)}_{k_1k_2} \ln(\mu^{(t+1)}_{k_1k_2}). \tag{4.23}$$

An important advantage of this technique is that log-likelihood (4.23) can be replaced by another one, such as the independence log-likelihood, thereby opening perspectives of modelling the effect of predictor variables. This is feasible without distorting the protective restrictions as they are used only in the E step.

To calculate the variance, we will use the method proposed by Meilijson (1989) (see Section 3.2.1). Suppose we define $\boldsymbol{\beta}$ to be a nonredundant set of $\mu_{k_1k_2}$. In the E step the only probabilities used are μ_{1,k_1} and $\mu_{1,k_1k_2}(\boldsymbol{\beta}_i)$. The former are fixed, the latter change as they depend on the small perturbations of the parameter vector. This implies that the dropout probabilities are implicitly changed, whereas they are a formal part of the parameter vector, and should remain fixed. A simple solution is to compute the dropout probabilities and consider them as a formal part of the parameter vector:

$$\phi_{2|k_2} = \frac{Z_{2,k_1k_2}}{Z_{k_1k_2}}, \tag{4.24}$$

with the complete data cell counts evaluated at the maximum. The right hand side of (4.24) is independent of k_1 due to the protective assumption.

The corresponding E step will be based on $\mu_{1,k_1}(\boldsymbol{\beta}_i)$ and $\mu_{2,k_1k_2}(\boldsymbol{\beta}_i)$ and hence the correct information matrix is obtained. There is another reason for the use of the $\boldsymbol{\phi}$ parameters. Because the covariance between the $\boldsymbol{\mu}$ and the $\boldsymbol{\phi}$ parameters is in general not zero, unless one assumes MAR, the correct covariance matrix is obtained only if it is based on the full information for both $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$.

In conclusion, the EM algorithm is a simple method to compute parameter estimates and their variances, but two major criticism apply. First, the estimates

can easily be determined using the methods of Section 4.2, making the technique redundant for parameter estimation. Secondly, for variance estimation, estimation of dropout probabilities is required, thereby weakening the advantage of protective estimation. An important advantage of the method is that it can be used to calculate a variance estimator for general T dimensional outcome vectors. In order to do so, the EM computations can be used to replace the variance computations for all $r \times r$ tables that occur, whereafter a delta method argument is applied to combine these variances into a variance for the multivariate cell probabilities.

4.5.3 Multiple Imputation

We will present the method of multiple imputation (see Section 3.2.2) for only two variables. Generalization is discussed at the end of this section.

Our interest lies in estimating the parameter vector $\boldsymbol{\beta}$, containing a nonredundant subset of $\mu^c_{k_1k_2}$. The set of easily estimable parameters $\boldsymbol{\gamma}$ includes μ_{1,k_1} and the conditional probabilities $\mu_{k_1|k_2}$, determined from μ_{2,k_1k_2}.

Using a normal posterior distribution for $\boldsymbol{\theta}$, the algorithm for‘ ‘filling-in’ the data is:

1. Draw $\boldsymbol{\gamma}^*$ from the posterior distribution of $\boldsymbol{\gamma}$. This yields $\mu^*_{k_1|k_2}$ and μ^*_{1,k_1}, which are easily transformed to μ^*_{1,k_1k_2} using the algorithm of Section 4.2.

2. Draw Z^*_{1,k_1k_2} from $f(Z_{1,k_1k_2}|Z_{1,k_1}, \boldsymbol{\gamma}^*)$. As discussed in Section 3.2.2, this is easily realized using a uniform random generator.

3. Calculate the estimate of the parameter of interest, and its estimated variance, using the completed data:

$$\begin{aligned} \hat{\mu}_{k_1k_2} &= \frac{Z^*_{1,k_1k_2} + Z_{2,k_1k_2}}{Z_{+,++}}, \\ \boldsymbol{U} &= \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \frac{1}{Z_{+,++}} \left(\mathrm{diag}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}' \right). \end{aligned}$$

Repeating the previous three steps M times, and combining these results as in Section 3.2.2, the parameters and variances of interest are found.

The method described above can easily be generalized to T dimensional contingency tables. Two obvious methods are: (1) use multiple imputation to estimate the variance of each two-way table, occuring in the computational method outlined in Section 4.2; combine the variance estimators into an estimator for the T dimensional table using e.g., the delta method (similar to the extension suggested in previous section); (2) use multiple imputation to complete all partial tables into T dimensional contingency tables.

Note that, although some of the draws may yield negative μ^*_{1,k_1k_2}, this does not imply that the procedure breaks down. It merely means that the corresponding table Z^*_{1,k_1k_2} will contain structural zeros. Further, a variance estimate is obtained without having to estimate the dropout probabilities, which is closer in spirit to protective estimation than the EM algorithm. This reduction in parameters to be estimated may result in a more efficient variance estimator. Assuming that $\boldsymbol{\gamma}$ is normally distributed is only an approximation, which may result in a bias for small samples (see Section 3.2.2).

4.5.4 Illustration

To illustrate the use of the three protective (variance) estimation procedures, consider the data of Table 4.1. Parameter estimates and standard errors are shown in Table 4.2. In all four cases, the EM based estimator coincides exactly with the likelihood estimator. A disadvantage is that likewise negative probabilities are found. Two solutions to this problem can be proposed. First, one can use a different (parametrization of the) likelihood in the M step. A second solution is to enforce restrictions in the E step. When the conditions of Theorem 1 are not satisfied, the probabilities found from (4.21) are invalid and have to be replaced by the appropriate boundary solution. Applying this technique to Table 4.1 (c) yields exactly the same solution as found with the transformed likelihood. We have omitted it from Table 4.2. A further disadvantage is that the dropout probabilities need to be estimated.

Although the delta method does not require the dropout probabilities, it also suffers from parameter space violations. Moreover, the delta method is known to be

somewhat inefficient, although reasonable agreement is observed for Table 4.1 (a) and the method shows to be superior for Table 4.1 (d). For Table 4.1 (b) caution should be used because parameters lie on the boundary of their space, and for Table 4.1 (c) even parameter estimates are not meaningful.

For Table 4.1 (a), multiple imputation yields the same result as both the direct likelihood and EM methods. It was based on 5000 samples. Although the procedure is more time consuming, we obtain a correct answer without having to estimate the dropout probabilities. With Table 4.1 (b), small differences are seen. The effect of these is that parameters move slightly away from the boundary. Indeed, the direction in which the MI parameters move is the same as the one seen for Table 4.1 (c). In the latter case, the MI estimator yields different results than seen with other methods, but they coincide with the transformed likelihood parameters (and with the EM parameters when a valid solution in the E step is ensured). In other words, using multiple imputation automatically ensures valid parameters, whereas other methods require some additional work such as finding a solution on the boundary, which is quite involved when the number of categories r is large. The standard errors differ between the two methods, but also the standard errors between both likelihood procedures differ, due to the fact that one of the dropout parameters with the transformed likelihood equals infinity, and asymptotic properties should be interpreted with caution. For Table 4.1 (d), the MI standard errors are slightly smaller than the likelihood based standard errors. In conclusion, multiple imputation seems to be a recommendable technique, not only for variance estimation, but also to estimate the parameters.

4.6 Examples

4.6.1 Fluvoxamine Data

The first example is taken from the data introduced in Section 2.1. We used the observations at times 2, 3 and 4, leading to the data of Table 2.3.

Table 4.3 gives the cell probability estimators and estimates of the standard errors under MAR, with all three protective estimators, with pseudo-likelihood and

Table 4.3: Estimated Cell Probabilities (Standard Errors) for the Fluvoxamine Data (all quantities were multiplied by 1000). The cell gives the outcomes at the 3 times considered. 6 Methods were used: likelihood estimation once for the completers only, and once assuming MAR, protective estimation using the Delta method, the EM algorithm and multiple imputation to calculate standard errors, and finally pseudo-likelihood using the protective assumption.

	Likelihood		Protective			PL
Cell	Comp.	MAR	Delta	EM	MI	
			Side Effects			
000	388(31)	355 (28)	342 (33)	342 (34)	343(30)	342 (34)
001	25(10)	23 (91)	31 (15)	31 (21)	30(13)	31 (21)
010	17 (8)	17 (81)	18 (29)	18 (27)	19 (9)	18 (27)
011	33(11)	34(111)	37 (25)	37 (29)	36(12)	37 (28)
100	128(21)	129 (21)	113 (20)	113 (21)	115(20)	113 (21)
101	21 (9)	21 (9)	26 (13)	26 (13)	24(10)	26 (13)
110	107(20)	117 (21)	115(179)	115(179)	129(35)	115(178)
111	281(29)	305 (29)	318(186)	318(180)	305(40)	318(179)
			Therapeutic Effect			
000	45(13)	51(13)	44 (30)	48(-)	50(15)	46 (18)
001	4 (4)	5 (5)	12 (36)	9(-)	7 (6)	15 (15)
010	0 (0)	0 (0)	0 (0)	0(-)	0 (0)	2 (3)
011	8 (6)	9 (6)	7 (5)	7(-)	7 (5)	11 (6)
100	190(25)	177(23)	184(117)	200(-)	205(33)	187 (70)
101	12 (7)	12 (7)	37(105)	22(-)	17(10)	35 (61)
110	215(26)	217(27)	265 (29)	265(-)	254(35)	168(273)
111	525(32)	531(31)	449 (37)	449(-)	461(41)	536(275)

for the subset of completers. Both side effects and therapeutic effect are analysed. Let us discuss the results for side effects first.

Clearly the three protective estimation strategies yield (virtually) the same results, apart from very large standard errors found with both the delta method and with the EM algorithm for cells 110 and 111. This might be explained by the fact that the information for cells 110 and 111 is largely the same because they are imputed from the same cells ($11*$ and $1**$). No violations of the boundary restrictions were encountered. The results obtained using pseudo-likelihood are very close to delta- and EM-results. Point estimates coincide, and standard errors are very similar. Once again, MI seems to yield more precise standard errors, perhaps because there is no need to sacrifice information to the estimation of dropout parameters. To assess the fit of the models, a deviance statistic was computed. To obtain a saturated model, one needs to consider a different probability table for each of the three observed patterns. The deviance is 11.23 for the protective estimator and 9.40 for the MAR model, on 4 degrees of freedom in both cases. Assuming a χ^2 distribution, P values are 0.024 and 0.052 respectively, pointing to a similar (lack of) fit for both models.

An alternative strategy consists of estimating both the measurement parameters and the dropout model, using the model advocated by Molenberghs, Kenward and Lesaffre (1997), using likelihood based estimation. Describing the dropout probability, given the outcomes, by a logistic regression, where the linear predictor describes the effect of both the previous and the current (possibly unobserved) outcome:

$$\text{logit}(\phi_{d|k_{d-1}k_d}) = \beta_0 + \beta_{d-1}k_{d-1} + \beta_d k_d,$$

several models can be considered. Estimates for the marginal cell probabilities and for the dropout parameters are given in Table 4.4. Standard errors are given between parentheses.

Several observations can be made. First, the overall deviances (corresponding to the likelihood of measurement and dropout processes simultaneously) convey a different message than the deviances of the dropout process. The overall deviances do not show a clear distinction between MAR and informative models. Indeed, although both terms in the informative model MNAR(2) are significant, the MAR

Table 4.4: Parametric Models for the Fluvoxamine Data, Side Effects, Likelihood Based Estimation. The cell gives the outcomes at the 3 times considered. Measurement probabilities are multiplied by 1000.

	MCAR	MAR	MNAR(1)	MNAR(2)
Cell				
000	355(28)	355(28)	346(29)	331(44)
001	23 (9)	23 (9)	25(10)	29(17)
010	17 (8)	17 (8)	17 (8)	18 (9)
011	34(11)	34(11)	40(13)	50(25)
100	129(21)	125(21)	115(20)	99(33)
101	21 (9)	21 (9)	22 (9)	22 (9)
110	117(21)	117(21)	108(20)	99(26)
111	305(29)	305(29)	327(29)	353(54)
Dropout Parameter				
β_0	-2.19	-3.56	-4.33	-5.59(0.34)
β_d			1.35	2.71(0.14)
β_{d-1}		0.86		-0.70(0.38)
Deviance(overall)	618.16	613.86	613.69	613.55
Deviance(dropout)	184.98	180.67	174.33	160.08

model and the informative model MNAR(1) with dependence on the current outcome only, describe the data equally well. This seems to be due to the "balance" which is achieved between dropout and measurement processes. Indeed, when the dropout model shows a better fit, achieved by including relevant parameters, the measurement model (with the same number of parameters) can afford to show a greater lack-of-fit. It clearly shows that the likelihood is very flat and similar likelihood values are obtained for conceptually very different models. This observation is in agreement with those made in Section 4.3 regarding the comparison of MAR and protective models for Table 4.1 (d). When either the previous measurement

only or the current observation only are included to describe the dropout process, the latter is the clear winner in terms of fitting the dropout process. It should be remembered, however, that it uses the current (possibly unobserved) value as a covariate and hence it should be considered jointly with the measurement model, at which level the fit is comparable. The conclusion is that at least one outcome should be included in the dropout model. This is in agreement with the results by Molenberghs, Kenward and Lesaffre (1997) who postulated that dropout mainly depends on the size of side effects, whereas a decrease in the therapeutic outcome seems to be responsible for dropout. From the deviances (p. 52) of the models in Table 4.3 we would infer that the previous measurement is the better candidate to describe dropout with respect to the side effects outcome. The fitted cell probabilities reported in Tables 4.3 and 4.4 are of similar magnitude. In contrast to the examples in Section 4.3, the standard errors are smaller for likelihood estimation than for protective estimation.

Therapeutic effect is more complicated because one of the combinations (010) does not occur. EM and delta method show slightly different estimates because EM was enforced to satisfy the boundary conditions. No sensible precision estimates are obtained with the EM method. Multiple imputation yields similar results and automatically satisfies the conditions. Note that the delta estimator still yields a valid set of probabilities. But when the corresponding dropout probabilities are computed, a situation comparable to the one in Table 4.1 (c) occurs e.g., for the cross-classification of the first and the third variable, given the second one equals 0. The counts are $Z_{00|0} = 11$, $Z_{01|0} = 1$, $Z_{10|0} = 46$, $Z_{11|0} = 3$, $Z_{0*|0} = 1$, $Z_{1*|0} = 2$, and hence violation of the restrictions might be due to chance. This means that the hypothetically complete table will have negative (but small) cell counts. Therefore, the delta method estimator can still be considered sensible. The zero cell counts lead to instability of the pseudo-likelihood estimator. A satisfactory solution is provided by applying a continuity correction, i.e., by adding 0.5 to all cell counts. The results then differ slightly from the ones obtained with the delta method.

For therapeutic effect, Molenberghs, Kenward and Lesaffre (1997) found that MAR and MCAR fitted equally well, but the fit was much improved by allowing

for informative dropout. In our analysis, the deviance for the MAR model is 5.57 on 4 degrees of freedom, which has to be contrasted with a deviance of 20.28 on 4 degrees of freedom for the protective estimator. Interpretation of these statistics should be done with caution, as the frequencies in some cells are very small and the estimator lies on the boundary.

4.6.2 Koch Dataset

The second example is taken from Koch *et al.* (1991). Presence or absence of colds during three successive years is recorded on 5554 subjects. Covariates are sex and the area of the residence of the subject. Considering the monotone sequences only, we have a subsample of 3112 subjects. Table 4.5 shows estimated cell probabilities for different strata. Apart from the entire set of data, we also considered stratification by sex (M/F), area (1/2), and by sex and area simultaneously. For all strata, we considered both an MAR and a protective model. All zeros in the table correspond to a boundary estimate and are not due to rounding. It is interesting to observe that a boundary solution for strata combined does not imply a boundary solution for the strata separately (e.g., all data versus stratified by area) and vice versa (e.g., area 1 versus area 1 stratified by sex).

Finally, EM and multiple imputation estimators are slightly different when applied to those tables in which boundary problems occur. As an example, we consider the estimates for the full set of data with EM: 23, 9, 6, 12, 11, 8, 9, and 22 respectively. They are much closer to the MAR solution, with the multiple imputation estimates even closer. This phenomenon is observed for all other tables as well. It means that the correction for parameter space violations is too extreme with the delta method estimator.

4.7 Conclusion

An estimator for a longitudinal categorical data table, subject to dropout, has been proposed. This protective estimator assumes that dropout depends on the unobserved outcomes, but not on the observed ones. Such an estimator had already been

Table 4.5: Estimated Cell Probabilities for the Koch Dataset (all quantities were multiplied by 100). Stratification for sex is indicated by male (M) and female (F); stratification for area is indicated by area 1 (1) and area 2 (2). A + indicates that the corresponding stratificator is not used.

	++	M+	F+	+1	+2	M1	M2	F1	F2
Protective Estimators									
111	23	20	29	21	26	15	21	24	32
112	9	7	10	9	11	9	11	11	11
121	0	3	0	3	6	7	7	8	4
122	18	16	17	15	11	11	10	10	12
211	11	13	10	11	11	11	13	10	9
212	8	7	8	10	6	12	6	10	7
221	0	5	0	4	12	12	14	11	11
222	31	30	27	29	17	24	19	18	16
MAR Estimators									
111	22	17	27	24	16	27	24	35	32
112	10	10	11	12	8	3	11	8	11
121	7	7	7	5	2	0	3	4	4
122	11	11	10	12	17	18	15	12	12
211	10	11	10	10	12	16	10	10	9
212	9	9	8	7	10	2	10	6	7
221	12	13	11	11	3	0	3	10	11
222	20	22	17	19	33	34	25	16	16

proposed for normal data by Brown (1990). The advantage of this technique is that no further assumptions on the missing data process have to be made in order to estimate the measurement parameters. This advantage is also shared by the more familiar MAR assumption. Both can be seen as estimators that are valid under a class of dropout models, rather than under a single mechanism.

The estimator is presented in the selection modelling framework, through derivation of a single set of cell probabilities. Alternatively, the cell probabilities for each pattern separately can be constructed, implying that all tables are completed, regardless of the number of observed components.

A connection with likelihood and pseudo-likelihood based estimation is established. Several estimation techniques have been proposed. A variance estimator can be based on the delta method, the EM algorithm, and on multiple imputation. Whereas the second and especially the first are computationally less demanding, the latter one has the important advantages that no range restriction violations occur. In order to avoid this problem with the other techniques, more complicated parametrizations have to be used.

The method is applied to a set of artificial data, in order to compare the different variance estimators, and also two sets of data have been analysed.

Our procedure can be extended to a modelling approach where covariates are measured, along with the outcomes. Especially the EM algorithm and the multiple imputation method are very promising in this respect. Estimating marginal probabilities and measures of association can be particularly desirable. Such a technique would be very appealing because one can assume a parsimonious model to describe the influence of predictor variables on the measurement probabilities, without having to model the dropout process explicitly in case a protective estimator is chosen. This property is shared with a MAR mechanism. Of course, one has to have evidence that either MAR or protective assumptions are plausible. Preferably, contextual information should be considered.

Chapter 5

Missing at Random for Pattern-Mixture Models

The missing data mechanisms introduced by Rubin (1976) and Little and Rubin (1987) (MCAR, MAR, and MNAR, see Section 3.1.2), are developed for a selection modelling framework. In Section 5.1, we define the available case missing value restrictions (ACMV, Molenberghs, Michiels, Kenward and Diggle 1998), and prove it to be the pattern-mixture counterpart of the MAR assumption. This facilitates a sensitivity analysis based on selection models and pattern-mixture models under the same assumption about the missing data mechanism. Indeed, since MCAR is merely independence and thus equivalent in both frameworks, and since we now have established a pattern-mixture counterpart of MAR, the taxonomy of Little and Rubin (1987) can also be made for pattern-mixture models. Although selection models and pattern-mixture models yield different parameters, a comparison of both concerning e.g., treatment effect, can give extra confidence in the obtained results. It is necessary to note that the equivalence developed here is for the special but important case of (monotone) dropout. A counterexample in the case of non-monotone missingness is given in Section 5.2.

5.1 Available Case Missing Value Restriction

In this section, we will restrict attention to a longitudinal data setting, where missingness is due to dropout. It will be shown in Section 5.2 that the results obtained for this case cannot be generalized to non-monotone patterns.

In a selection model, the joint density $f(\boldsymbol{y}, d)$ is factorized as in Section 3.1.2 as

$$f(\boldsymbol{y}, d) = f(\boldsymbol{y}) f(d|\boldsymbol{y}). \tag{5.1}$$

The MAR assumption, where a subject's missingness mechanism depends on its observed outcomes only, can be written as

$$f(d = t|y_1, \cdots, y_T) = f(d = t|y_1, \cdots, y_t), \text{ for } t = 1, \ldots, T.$$

Remember that in a pattern-mixture model, the joint density of $f(\boldsymbol{y}, r)$ is factorized as

$$f(\boldsymbol{y}, d) = f(d) f(\boldsymbol{y}|d).$$

We will now show how pattern-mixture models can be classified using exactly the same taxonomy as is used for selection models. Furthermore, we establish a link between this classification and the identifying restrictions proposed in Little (1993).

Clearly, selection models and pattern-mixture models coincide under MCAR, since in either case the joint density simplifies to $f(\boldsymbol{y})f(d)$. Next, we show that MAR can be expressed in a pattern-mixture framework through restrictions, related to the *complete case missing value* (CCMV) restrictions (Little 1993), which we call *available case missing value* (ACMV) restrictions. Little's CCMV restrictions set a conditional density of unobserved components given a particular set of observed components equal to the corresponding conditional density in the subgroup of completers. Our ACMV restrictions equate this conditional density to the one calculated from the subgroup of all patterns for which all required components have been observed. It is intuitively more appealing to use a component based on all available information, than based on a smaller, and possibly less similar, group.

In our setting of longitudinal data with dropouts, CCMV can be defined formally

as the condition that

$$\forall t \geq 2, \forall j < t : f(y_t|y_1, \cdots, y_{t-1}, d = j) = f(y_t|y_1, \cdots, y_{t-1}, d = T),$$

whereas ACMV is the condition that

$$\forall t \geq 2, \forall j < t : f(y_t|y_1, \cdots, y_{t-1}, d = j) = f(y_t|y_1, \cdots, y_{t-1}, d \geq t). \tag{5.2}$$

If there are only 2 time points ($T = 2$), then ACMV and CCMV coincide. With these definitions, our main result is:

Theorem 2 *For longitudinal data with dropouts, MAR $\Longleftrightarrow$ ACMV.*

To establish the proof of this theorem, a lemma is needed:

Lemma 1 *In a longitudinal setting with dropout,*
ACMV $\Longleftrightarrow \forall t \geq 2, \forall j < t : f(y_t|y_1, \cdots, y_{t-1}, d = j) = f(y_t|y_1, \cdots, y_{t-1})$.

Proof of Lemma 1

Take $t \geq 2, j < t$, then ACMV leads to:

$$\begin{aligned}
& f(y_t|y_1, \cdots, y_{t-1}) \\
&\quad = \sum_{i=1}^{t-1} f(y_t|y_1, \cdots, y_{t-1}, d = i) f(d = i|y_1, \cdots, y_{t-1}) \\
&\qquad\qquad + f(y_t|y_1, \cdots, y_{t-1}, d \geq t) f(d \geq t|y_1, \cdots, y_{t-1}) \\
&\quad = \sum_{i=1}^{t-1} f(y_t|y_1, \cdots, y_{t-1}, d = j) f(d = i|y_1, \cdots, y_{t-1}) \\
&\qquad\qquad + f(y_t|y_1, \cdots, y_{t-1}, d = j) f(d \geq t|y_1, \cdots, y_{t-1}) \\
&\quad = f(y_t|y_1, \cdots, y_{t-1}, d = j) \left[\sum_{i=1}^{t-1} f(d = i|y_1, \cdots, y_{t-1}) \right. \\
&\qquad\qquad \left. + f(d \geq t|y_1, \cdots, y_{t-1}) \right] \\
&\quad = f(y_t|y_1, \cdots, y_{t-1}, d = j).
\end{aligned}$$

To show the reverse direction, take again $t \geq 2, j < t$.

$$\begin{aligned}
& f(y_t|y_1,\cdots,y_{t-1},d\geq t)f(d\geq t|y_1,\cdots,y_{t-1}) \\
&\quad = f(y_t,d\geq t|y_1,\cdots,y_{t-1}) \\
&\quad = f(y_t|y_1,\cdots,y_{t-1}) - \sum_{i=1}^{t-1} f(y_t|y_1,\cdots,y_{t-1},d=i)f(d=i|y_1,\cdots,y_{t-1}) \\
&\quad = f(y_t|y_1,\cdots,y_{t-1}) - \sum_{i=1}^{t-1} f(y_t|y_1,\cdots,y_{t-1})f(d=i|y_1,\cdots,y_{t-1}) \\
&\quad = f(y_t|y_1,\cdots,y_{t-1})\left[1-\sum_{i=1}^{t-1} f(d=i|y_1,\cdots,y_{t-1})\right] \\
&\quad = f(y_t|y_1,\cdots,y_{t-1},d=j)\left[1-\sum_{i=1}^{t-1} f(d=i|y_1,\cdots,y_{t-1})\right] \\
&\quad = f(y_t|y_1,\cdots,y_{t-1},d=j)f(d\geq t|y_1,\cdots,y_{t-1}).\square
\end{aligned}$$

Proof of Theorem 2

MAR $\Rightarrow$ ACMV

Consider the ratio Q of the complete data likelihood to the observed data likelihood. This gives, under the MAR assumption:

$$\begin{aligned}
Q &= \frac{f(y_1,\cdots,y_T)f(d=i|y_1,\cdots,y_T)}{f(y_1,\cdots,y_i)f(d=i|y_1,\cdots,y_i)} \\
&= \frac{f(y_1,\cdots,y_T)f(d=i|y_1,\cdots,y_i)}{f(y_1,\cdots,y_i)f(d=i|y_1,\cdots,y_i)} \\
&= f(y_{i+1},\ldots,y_T|y_1,\ldots,y_i). \qquad (5.3)
\end{aligned}$$

Further, one can always write:

$$\begin{aligned}
Q &= \frac{f(y_{i+1},\cdots,y_T|y_1,\cdots,y_i,d=i)f(y_1,\ldots,y_i|d=i)f(d=i)}{f(y_1,\ldots,y_i|d=i)f(d=i)} \\
&= f(y_{i+1},\ldots,y_T|y_1,\ldots,y_i,d=i). \qquad (5.4)
\end{aligned}$$

Equating expressions (5.3) and (5.4) for Q we see that

$$f(y_{i+1},\cdots,y_T|y_1,\cdots,y_i,d=i) = f(y_{i+1},\cdots,y_T|y_1,\cdots,y_i). \qquad (5.5)$$

To show that (5.5) implies the ACMV conditions (5.2), we will use the induction principle on t. First, consider the case $t = 2$.

Using (5.5) for $i = 1$, and integrating over $y_3, \cdots, y_T$, we obtain

$$f(y_2|y_1, d = 1) = f(y_2|y_1),$$

leading to, using Lemma 1,

$$f(y_2|y_1, d = 1) = f(y_2|y_1, d \geq 2).$$

Suppose by induction ACMV holds $\forall t \leq i$. We will now prove the hypothesis for $t = i + 1$. Choose $j \leq i$. Then from the induction hypothesis and Lemma 1, it follows that

$$\begin{aligned} \forall j < t \leq i : f(y_t|y_1, \cdots, y_{t-1}, d = j) &= f(y_t|y_1, \cdots, y_{t-1}, d \geq t) \\ &= f(y_t|y_1, \cdots, y_{t-1}). \end{aligned}$$

Taking the product over $t = j + 1, \cdots, i$ then gives

$$f(y_{j+1}, \cdots, y_i|y_1, \cdots, y_j, d = j) = f(y_{j+1}, \cdots, y_i|y_1, \cdots, y_j). \tag{5.6}$$

After integration over $y_{i+2}, \cdots, y_T$, Equation (5.5) leads to

$$f(y_{j+1}, \cdots, y_{i+1}|y_1, \cdots, y_j, d = j) = f(y_{j+1}, \cdots, y_{i+1}|y_1, \cdots, y_j). \tag{5.7}$$

Dividing (5.7) by (5.6) and equating the left and right hand sides, we find that

$$f(y_{i+1}|y_1, \cdots, y_i, d = j) = f(y_{i+1}|y_1, \cdots, y_i).$$

This holds $\forall j \leq i$, and Lemma 1 shows this is equivalent with ACMV.

ACMV $\Rightarrow$ MAR

Starting from the ACMV assumption and Lemma 1, we have

$$\forall t \geq 2, \forall j < t : f(y_t|y_1, \cdots, y_{t-1}, d = j) = f(y_t|y_1, \cdots, y_{t-1}). \tag{5.8}$$

We now factorize the full data density as

$$\begin{aligned} f(y_1, \cdots, y_T, d = i) &= f(y_1, \cdots, y_i, d = i) f(y_{i+1}, \cdots, y_T|y_1, \cdots, y_i, d = i) \\ &= f(y_1, \cdots, y_i, d = i) \prod_{t=i+1}^{T} f(y_t|y_1, \cdots, y_{t-1}, d = i). \end{aligned}$$

Using (5.8), it follows that

$$\begin{aligned} f(y_1,\cdots,y_T,d=i) &= f(y_1,\cdots,y_i|d=i)f(d=i)\prod_{t=i+1}^{T} f(y_t|y_1,\cdots,y_{t-1}) \\ &= f(y_1,\cdots,y_i|d=i)f(d=i)f(y_{i+1},\cdots,y_T|y_1,\cdots,y_i) \\ &= \frac{f(y_1,\cdots,y_i|d=i)f(d=i)}{f(y_1,\ldots,y_i)}f(y_1,\cdots,y_T) \\ &= f(d=i|y_1,\ldots,y_i)f(y_1,\ldots,y_T) \end{aligned} \tag{5.9}$$

An alternative factorization of $f(y,d)$ gives

$$f(y_1,\cdots,y_T,d=i) = f(d=i|y_1,\cdots,y_T)f(y_1,\cdots,y_T). \tag{5.10}$$

It follows from (5.9) and (5.10) that

$$f(d=i|y_1,\cdots,y_T) = f(d=i|y_1,\cdots,y_i).\square$$

An interesting by-product of this theorem is that, since MAR corresponds to a set of (untestable) restrictions (ACMV) in the pattern-mixture framework, MAR itself is also untestable. This fact is often overlooked in the selection framework.

Little (1993) suggested the possibility of using more than the completers to construct identifying restrictions for two practical reasons: (1) the set of completers may be small and (2) there may be a closer similarity between the conditional distributions given $d=t$ and some other incomplete pattern $d=s$, than between those for $d=t$ and the completers, $d=T$.

We suggest the use of the following procedure, which uses the maximum amount of information. First, restrict the dataset to the first two components only. Then, missing data patterns $d=2,\ldots,T$ collapse into a single pattern $d\geq 2$. Applying ACMV restrictions to $d=1$ and $d\geq 2$ leads to the construction of the density $f(y_2|y_1,d=1)=f(y_2|y_1,d\geq 2)$, as in (5.2). Multiplying by $f(y_1|d=1)$ leads to $f(y_1,y_2|d=1)$, thus determining the joint densities of $f(y_1,y_2|d)$ for all $d=1,\cdots,T$. Next, $f(y_3|y_1,y_2,d)$ $(d=1,2)$ can be calculated from $f(y_3|y_1,y_2,d\geq 3)$. We then proceed by induction to construct all joint densities.

5.2 Non-Monotone Patterns: A Counterexample

It has to be noted that the result of Theorem 2 does not hold for general missing data patterns. Consider a bivariate outcome (y_1, y_2) where missingness can occur in both components. Let (r_1, r_2) be the corresponding bivariate missingness indicator, where $r_j = 0$ if y_j is missing and 1 otherwise $(j = 1, 2)$.

Consider the following MAR mechanism:

$$f(r|y) = P(r_1, r_2|y_1, y_2) = \begin{cases} p & \text{if } (r_1, r_2) = (0, 0), \\ q_{y_1} & \text{if } (r_1, r_2) = (1, 0), \\ s_{y_2} & \text{if } (r_1, r_2) = (0, 1), \\ 1 - p - q_{y_1} - s_{y_2} & \text{if } (r_1, r_2) = (1, 1). \end{cases} \tag{5.11}$$

We need to indicate how the concept of ACMV will be translated to this setting. Several proposals can be considered. A trivial extension of the ACMV restrictions in the monotone case, implies for the patterns $r = (1, 0)$ and $r = (0, 1)$:

$$r = (1, 0) \quad : \quad f(y_1, y_2|r = (1, 0)) = f(y_1|r = (1, 0)).f(y_2|y_1, r = (1, 1)), \tag{5.12}$$

$$r = (0, 1) \quad : \quad f(y_1, y_2|r = (0, 1)) = f(y_2|r = (0, 1)).f(y_1|y_2, r = (1, 1)). \tag{5.13}$$

The idea is that the density of missing components, given observed components, is replaced by the corresponding density of patterns for which both are available. Restrictions for the pattern $r = (0, 0)$ will be discussed further.

From condition (5.12) we derive

$$\begin{aligned} \frac{f(r = (1, 0)|y_1, y_2)f(y_1, y_2)}{f(r = (1, 0))} &= \frac{f(r = (1, 0)|y_1)f(y_1)}{f(r = (1, 0))} \frac{f(r = (1, 1)|y_1, y_2)f(y_1, y_2)}{f(r = (1, 1)|y_1)f(y_1)} \\ &\Updownarrow \\ f(r = (1, 0)|y_1, y_2) &= \frac{f(r = (1, 0)|y_1)f(r = (1, 1)|y_1, y_2)}{f(r = (1, 1)|y_1)} \\ &\Updownarrow \\ f(r = (1, 1)|y_1, y_2) &= f(r = (1, 1)|y_1), \end{aligned}$$

since $f(r = (1, 0)|y_1, y_2) = f(r = (1, 0)|y_1) = q_{y_1}$, implying that s_{y_2} is constant.

Similarly, condition (5.13) implies that q_{y_1} is constant:

$$\begin{aligned}
\frac{f(r=(0,1)|y_1,y_2)f(y_1,y_2)}{f(r=(0,1))} &= \frac{f(r=(0,1)|y_2)f(y_2)}{f(r=(0,1))}\frac{f(r=(1,1)|y_1,y_2)f(y_1,y_2)}{f(r=(1,1)|y_2)f(y_2)} \\
&\Updownarrow \\
f(r=(0,1)|y_1,y_2) &= \frac{f(r=(0,1)|y_2)f(r=(1,1)|y_1,y_2)}{f(r=(1,1)|y_2)} \\
&\Updownarrow \\
f(r=(1,1)|y_1,y_2) &= f(r=(1,1)|y_2),
\end{aligned}$$

since $f(r=(0,1)|y_1,y_2) = f(r=(0,1)|y_2) = s_{y_2}$.

Clearly, since both q_{y_1} and s_{y_2} have to be constant, the mechanism needs to be MCAR. In other words, ACMV≡MCAR, independent of the restrictions for $f(y_1,y_2|r=(0,0))$, and hence ACMV and MAR differ.

There are different methods to construct $f(y_1,y_2|r=(0,0))$:

$$\begin{aligned}
f(y_1,y_2|r=(0,0)) &= f(y_1,y_2|r=(1,1)), \\
&= f(y_1|r=(1,1) \text{ or } r=(1,0))f(y_2|y_1,r=(1,1)), \\
&= f(y_2|r=(1,1) \text{ or } r=(0,1))f(y_1|y_2,r=(1,1)).
\end{aligned}$$

The first proposal is CCMV: take the things one does not have from the completers. The second proposal means that we first identify the density of the first component by equating it to the density of the patterns where y_1 is observed ($r=(1,0)$ and $r=(1,1)$), and that we then identify the density of y_2 given y_1 based on the completers. This is in fact also ACMV, which is more relaxed than CCMV, because one takes what one needs from as much cases as possible. The third proposal is analogous to the second, with only y_1 and y_2 interchanged. Although it seems one would have to choose between one of these three proposals, due to the other patterns that lead to MCAR, the different options are in fact exactly the same.

5.3 Conclusion

In a missing data context, the choice of modelling framework needs careful consideration. The simplicity of the classical MCAR, MAR, and MNAR taxonomy is

not a feature particular to the selection modelling approach, since, in the case of monotone missing data, the same taxonomy can be developed for pattern-mixture models. The MAR assumption is translated in the latter case into the ACMV restriction. This intermediate case corresponds to an explicit and reasonably natural set of restrictions on the unidentifiable components of the full data distribution. It is also shown that this equivalence does not hold for non-monotone missing data patterns.

Since we have the same missing data mechanism in both the selection and the pattern-mixture framework, a sensitivity analysis can be carried out. This is done in Chapter 6 for categorical data, and in Chapter 8 for a dataset with continuous outcomes. Furthermore, another advantage of this equivalence is that the interesting parts of both frameworks can be combined into a pseudo-likelihood (see Chapter 7).

Chapter 6

Selection Models and Pattern-Mixture Models for Incomplete Data With Covariates

In the previous chapter, we have established ACMV, the pattern-mixture analogue for the MAR assumption in the selection modelling framework. This leads to a (theoretical) equivalence of selection and pattern-mixture models, since they are both factorizations of the same distribution, and we have an expression for the same missing data mechanism in both frameworks. Here, we study how this result can be used to model longitudinal data. We have analysed two sets of categorical data, and in both cases we came up with similar conclusions (Molenberghs, Michiels and Lipsitz 1999, Michiels, Molenberghs and Lipsitz 1998).

In Section 6.1, we establish the notation used in this chapter. A comparison between selection models and pattern-mixture models is given in Section 6.2. All models are based on the odds ratio model proposed by Dale (1986) and Molenberghs and Lesaffre (1994, 1999). First, selection models are explored, and precision estimates are derived. Then we look at pattern-mixture models. Here, identifying restrictions are needed. Since we assumed MAR as modelling assumption for the selection model, we chose to use ACMV for the pattern-mixture model. Both models yield different parameters, therefore a comparison of their relative merits is included. We end the section with a discussion of the precision estimates for the pattern-mixture

model, based on profile likelihood and on multiple imputation. Two examples with categorical outcomes are discussed in Section 6.3, where all issues of the previous section are encountered.

6.1 Marginal Modelling of Incomplete Categorical Data

We adopt notation introduced in Chapter 3. The observed data are $\boldsymbol{Z}_i$, a partially classified table of the complete data $\boldsymbol{Z}_i^c$. The cell counts and corresponding probabilities of the margin can be thought of as arising by summing over the appropriate rows or columns in the corresponding complete table. We then have a linear relationship between observed and complete quantities: $\boldsymbol{Z}_i = C_i \boldsymbol{Z}_i^c$ and $\boldsymbol{\nu}_i = C_i \boldsymbol{\nu}_i^c$. We call the matrix C_i which consists of 0's and 1's the coarsening matrix, in agreement with Molenberghs and Goetghebeur (1997) and Heitjan and Rubin (1991). Then the kernel of the multinomial (observed) loglikelihood is

$$\ell(\theta; \boldsymbol{Z}) = \sum_{i=1}^{N} \boldsymbol{Z}_i' \ln(\boldsymbol{\nu}_i)$$

subject to the constraints $\sum_k \nu_{ik} = 1$, where the summation index k cycles through all (multi-indexed) cells of $\boldsymbol{\nu}_i$.

We develop estimation of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ (either $\boldsymbol{\theta}^S$ or $\boldsymbol{\theta}^P$). Following McCullagh and Nelder (1989), the score equations are given by

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\nu}_i}{\partial \boldsymbol{\theta}} \right)' \boldsymbol{V}_i^- (\boldsymbol{Z}_i - n_i \boldsymbol{\nu}_i),$$

with $\boldsymbol{V}_i = \mathrm{diag}(\boldsymbol{\nu}_i) - \boldsymbol{\nu}_i \boldsymbol{\nu}_i'$. Further,

$$\left(\frac{\partial \boldsymbol{\nu}_i}{\partial \boldsymbol{\theta}} \right)' = \left(\frac{\partial \boldsymbol{\nu}_i^c}{\partial \boldsymbol{\theta}} \right)' \left(\frac{\partial \boldsymbol{\nu}_i}{\partial \boldsymbol{\nu}_i^c} \right)' = \left(\frac{\partial \boldsymbol{\nu}_i^c}{\partial \boldsymbol{\theta}} \right)' C_i'. \tag{6.1}$$

Implementing a Newton-Raphson algorithm involves computation of second derivatives of the likelihood for which expressions can be found in Molenberghs and Lesaffre (1999). Alternatively, the log-likelihood can be fed to a numeric optimizer such as the GAUSS-procedure OPTMUM or the S-PLUS-function NLMINB.

We assume that the data are collected following a multinomial sampling scheme and let composite generalized linear models hold:

- for the parameters of the selection model:

$$\boldsymbol{\eta}_i^S(\boldsymbol{\mu}_i^{Sc}) = \boldsymbol{X}_i^{S\beta}\boldsymbol{\beta}^S, \tag{6.2}$$

$$\boldsymbol{\xi}_i^S(\boldsymbol{\phi}_i^{Sc}) = \boldsymbol{X}_i^{S\alpha}\boldsymbol{\alpha}^S, \tag{6.3}$$

- and for the parameters of the pattern-mixture model:

$$\boldsymbol{\eta}_i^P(\boldsymbol{\mu}_{i|d}^{Pc}) = \boldsymbol{X}_{i|d}^{P\beta}\boldsymbol{\beta}_d^P, \qquad d = 1, \ldots, T, \tag{6.4}$$

$$\boldsymbol{\xi}_i^P(\boldsymbol{\phi}_i^{Pc}) = \boldsymbol{X}_i^{P\alpha}\boldsymbol{\alpha}^P. \tag{6.5}$$

  Denote $\boldsymbol{\beta}^P = (\boldsymbol{\beta}_d^P)_{d=1,\ldots,T}$, $\boldsymbol{X}_i^{P\beta} = (\boldsymbol{X}_{i|d}^{P\beta})_{d=1,\ldots,T}$, $\boldsymbol{X}_i^S = (\boldsymbol{X}_i^{S\beta}, \boldsymbol{X}_i^{S\alpha})$, and $\boldsymbol{X}_i^P = (\boldsymbol{X}_i^{P\beta}, \boldsymbol{X}_i^{P\alpha})$.

Note that, in a selection model, the dropout probabilities are modelled conditional on the outcomes. This implies that $\boldsymbol{X}_i^{S\alpha}$ contains, apart from the covariates included in the study, also the outcome variables.

Choices for the vector link functions $\boldsymbol{\eta}_i^S$, $\boldsymbol{\eta}_i^P$, $\boldsymbol{\xi}_i^S$ and $\boldsymbol{\xi}_i^P$ will be discussed later.

6.2 Selection Models Versus Pattern-Mixture Models

A specific model choice is based on the form of (6.2) and (6.3), or (6.4) and (6.5), reflected in the matrix $\partial\boldsymbol{\nu}_i^c/\partial\boldsymbol{\theta}$ in (6.1). We will discuss selection models and pattern-mixture models in turn. To simplify notation, a specific bivariate setting of the notation in Chapter 3 will be considered. Extension to the general case is straightforward albeit heavy in notation.

6.2.1 Notation

Assume complete data consist of a design matrix and a categorical outcome (with c levels) measured on two occasions for each subject. Assume further that each

subject is seen at the first occasion, with only part of them measured at the second occasion. The observed multinomial data consist of a set of complete $c \times c$ tables $\boldsymbol{Z}_{i2}$ with counts Z_{i2jk} $(j,k = 1,\ldots,c)$ and a supplemental margin $\boldsymbol{Z}_{i1}$ with counts Z_{i1j}, where $j = 1,\ldots,c$. The (hypothetical) full data amount to two $c \times c$ tables Z^c_{idjk} with $d = 1,2$ and $j,k = 1,\ldots,c$. Obviously, the relation between complete and observed counts is $Z_{i2jk} = Z^c_{i2jk}$ and $Z_{i1j} = \sum_{k=1}^{c} Z^c_{i1jk}$. Adopting the convention that the counts of all tables corresponding to design level i are represented as vectors in lexicographic ordering, and further that $\boldsymbol{Z}_i = (\boldsymbol{Z}'_{i2}, \boldsymbol{Z}'_{i1})'$ with a similar expression for $\boldsymbol{Z}^c_i$, we deduce that the coarsening matrix C_i in this case is given by

$$C = C_i = \left(\begin{array}{c|c} C_{i0} & 0 \\ \hline 0 & C_{i1} \end{array}\right) = \left(\begin{array}{c|c} I_{c^2} & 0_{c^2,c^2} \\ \hline 0_{c,c^2} & I_c \otimes 1_{1,c} \end{array}\right), \tag{6.6}$$

with $I_.$ the identity matrix, $0_.$ a matrix of zeros, $1_.$ a matrix of ones and $\otimes$ the Kronecker product.

6.2.2 Selection Models

As before, we denote the probability for an observation with design $\boldsymbol{X}^S_i$ to fall into category (j,k) of the dth table by

$$\nu^{Sc}_{idjk}(\boldsymbol{\theta}^S) = \mu^{Sc}_{ijk}(\boldsymbol{\beta}^S)\, \phi^{Sc}_{id|jk}(\boldsymbol{\alpha}^S). \tag{6.7}$$

Since in this section, we work only in the selection model setting, the superscript S will be omitted. There are obvious constraints on these probabilities. For each i:

$$\sum_{d=1}^{2}\sum_{j=1}^{c}\sum_{k=1}^{c} \nu^c_{idjk} = \sum_{j=1}^{c}\sum_{k=1}^{c} \mu^c_{ijk} = 1 \;\text{ and }\; \sum_{d=1}^{2} \phi^c_{id|jk} = 1 \text{ for all } j,k.$$

In this section, we define $\phi_{ijk} = \phi^c_{i2|jk} = 1 - \phi^c_{i1|jk}$, the probability that a measurement is made at the second occasion, given that the complete data are $(Y_{i1} = j, Y_{i2} = k)$.

When the complete data $\boldsymbol{Z}^c_i$ would be available, the information required to estimate the measurement parameters $\boldsymbol{\beta}$ could be obtained from the collapsed table with entries $Z^c_{i1jk} + Z^c_{i2jk}$, while the parameters of $\boldsymbol{\alpha}$ would follow from the pairs (Z^c_{i1jk}, Z^c_{i2jk}) for all (j,k). For the partially observed table however, we have to fit

the observed data likelihood with cell probabilities ν_{i2jk} and

$$\nu_{i1j+} = \sum_{k=1}^{c} \nu^c_{i1jk} = \sum_{k=1}^{c} \mu^c_{ijk}(1 - \phi_{ijk}).$$

In general, the latter expression does not split into a $\boldsymbol{\mu}$ and a $\boldsymbol{\phi}$ part.

To fully specify (6.2) and (6.3), we will choose link functions for the left hand sides of the form:

$$\boldsymbol{\eta}_i(\boldsymbol{\mu}^c_i) = D_\mu \ln(A_\mu \boldsymbol{\mu}^c_i) \text{ and } \boldsymbol{\xi}_i(\boldsymbol{\phi}^c_i) = D_\phi \ln(A_\phi \boldsymbol{\phi}^c_i), \tag{6.8}$$

where A_μ and A_ϕ are matrices containing zeros and ones, used to construct sums of probabilities (e.g., probabilities of collapsed tables), and D_μ and D_ϕ are contrast matrices (with entries equal to 0, 1 or -1). In other words, log contrasts of the probabilities are equated to a set of linear predictors. The logistic model forms a special case, but the general form was also used by McCullagh and Nelder (1989) and Lang and Agresti (1994). Log odds ratios to model associations can be incorporated in this formulation.

Observing that $\boldsymbol{\nu}^c_i = \boldsymbol{\nu}^c_i(\boldsymbol{\mu}^c_i, \boldsymbol{\phi}^c_i)$ (see Equation 6.7), we find

$$\frac{\partial \boldsymbol{\nu}^c_i}{\partial(\boldsymbol{\mu}^c_i, \boldsymbol{\phi}^c_i)} = \left(\begin{array}{c|c} F_i & M_i \\ \hline I - F_i & -M_i \end{array}\right)$$

with $F_i = \text{diag}(\boldsymbol{\phi}^c_i)$ and $M_i = \text{diag}(\boldsymbol{\mu}^c_i)$. Introducing some extra notation

$$\begin{aligned} T_{\eta i} &= \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}^c_i}\right) = D_\mu(\text{diag}(A_\mu \boldsymbol{\mu}^c_i))^{-1} A_\mu, \\ T_{\xi i} &= \left(\frac{\partial \boldsymbol{\xi}_i}{\partial \boldsymbol{\phi}^c_i}\right) = D_\phi(\text{diag}(A_\phi \boldsymbol{\phi}^c_i))^{-1} A_\phi \end{aligned}$$

the score equations become

$$\frac{\partial \ell}{\partial(\boldsymbol{\beta}, \boldsymbol{\alpha})} = \sum_{i=1}^{N} \left(\begin{array}{c|c} X^\beta_i & 0 \\ \hline 0 & X^\alpha_i \end{array}\right)' \left(\begin{array}{c|c} T^{-1}_{\eta i} & 0 \\ \hline 0 & T^{-1}_{\xi i} \end{array}\right)' \left(\begin{array}{c|c} F_i & M_i \\ \hline I - F_i & -M_i \end{array}\right)' C_i' \boldsymbol{V}_i^- \boldsymbol{S}_i \tag{6.9}$$

with $\boldsymbol{S}_i = \boldsymbol{Z}_i - n_i \boldsymbol{\nu}_i$. Solving these equations can be done using a Newton-Raphson algorithm, as discussed in Section 6.1. The inverse of the matrix of second derivatives, evaluated at the maximum of the likelihood function, provides an estimator of the precision.

6.2.3 Pattern-Mixture Models

For pattern-mixture models, we factorize the complete data probabilities as products of marginal dropout parameters and measurement probabilities conditional on the dropout pattern:

$$\nu^{Pc}_{idjk}(\boldsymbol{\theta}^P) = \phi^{Pc}_{id}(\boldsymbol{\alpha}^P)\ \mu^{Pc}_{ijk|d}(\boldsymbol{\beta}^P). \tag{6.10}$$

In this section, only the pattern-mixture model setting will be treated, and therefore the superscript P will be omitted. The constraints on these probabilities are, for each i:

$$\sum_{j=1}^{c}\sum_{k=1}^{c}\mu^c_{ijk|d} = 1, \quad d = 1,2 \ \text{ and } \ \sum_{d=1}^{2}\phi^c_{id} = 1.$$

To derive the score equation, we need to adapt the notation slightly. The measurement probabilities for pattern $d = 1, 2$ are collected into a vector $\boldsymbol{\mu}^c_{i|d}$ and the dropout parameters into $\boldsymbol{\phi}^c_i$. The design for the measurement part has 2 components $\boldsymbol{\eta}_{i|d} = \boldsymbol{\eta}_i(\boldsymbol{\mu}^c_{i|d}) = \boldsymbol{X}^{\beta}_{i|d}\boldsymbol{\beta}_d$ $(d = 1, 2)$. As in Equations 6.8, we can write the link functions as

$$\boldsymbol{\eta}_{i|d} = \boldsymbol{\eta}_i(\boldsymbol{\mu}^c_{i|d}) = D_{\mu|d}\ln(A_{\mu|d}\boldsymbol{\mu}^c_{i|d}),\ d = 1,2 \ \text{ and } \ \boldsymbol{\xi}_i(\boldsymbol{\phi}^c_i) = D_{\phi}\ln(A_{\phi}\boldsymbol{\phi}^c_i),$$

where again $A_{\mu|d}$ and A_ϕ are sum matrices, and $D_{\mu|d}$ and D_ϕ are contrast matrices. This leads to

$$\begin{aligned} T_{\eta i|d} &= \left(\frac{\partial\boldsymbol{\eta}_{i|d}}{\partial\boldsymbol{\mu}^c_i}\right) = D_{\mu|d}(\mathrm{diag}(A_{\mu|d}\boldsymbol{\mu}^c_{i|d}))^{-1}A_{\mu|d}, \\ T_{\xi i} &= \left(\frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\phi}^c_i}\right) = D_{\phi}(\mathrm{diag}(A_{\phi}\boldsymbol{\phi}^c_i))^{-1}A_{\phi} \end{aligned}$$

We can now write the score equations as:

$$\frac{\partial\ell}{\partial(\boldsymbol{\beta},\boldsymbol{\alpha})} = \sum_{i=1}^{N}\left(\begin{array}{c|c} X^{\beta}_{i|1} & 0 \\ \hline X^{\beta}_{i|2} & 0 \\ \hline 0 & X^{\alpha}_i \end{array}\right)'\left(\begin{array}{c|c|c} T^{-1}_{\eta i|1} & 0 & 0 \\ \hline 0 & T^{-1}_{\eta i|2} & 0 \\ \hline 0 & 0 & T^{-1}_{\xi i} \end{array}\right)'\left(\begin{array}{c|c|c} F_i & 0 & \boldsymbol{\mu}^c_{i|1} \\ \hline 0 & I-F_i & -\boldsymbol{\mu}^c_{i|2} \end{array}\right)' C_i'\boldsymbol{V}_i^{-}\boldsymbol{S}_i. \tag{6.11}$$

In order to maximize the pattern-mixture likelihood, we need to discuss identifying restrictions first.

6.2.4 Identifying Restrictions

As mentioned in Section 3.1.2, a pattern-mixture model is chronically under-identified. In the case described above, the incomplete pattern ($d = 1$) would provide information about the first measurement, but neither about the second one, nor about the association between both. To be specific, for the incomplete pattern, only the probabilities $\mu^c_{ij|1}$ are identified, leaving the $\mu^c_{ik|1j}$ inestimable.

We will describe a solution to this problem, by first considering a measurement model and secondly combining it with a particular form of identifying restrictions on the model parameters.

A possible modelling approach is to consider a bivariate model for the completers, i.e., a Dale model (Dale 1986), and a univariate model for the incomplete observations, i.e., a logistic regression model. We will term this the *minimal approach*. The Dale model, applied to two binary outcomes Y_{i1} and Y_{i2}, supplements a logistic regression for each of the outcomes separately, with an odds ratio:

$$\psi_i = \frac{P(Y_{i1}=0, Y_{i2}=0|\boldsymbol{X}_i)P(Y_{i1}=1, Y_{i2}=1|\boldsymbol{X}_i)}{P(Y_{i1}=0, Y_{i2}=1|\boldsymbol{X}_i)P(Y_{i1}=1, Y_{i2}=0|\boldsymbol{X}_i)}. \tag{6.12}$$

In terms of the probabilities μ^c_{ijk}, we define

$$\eta_{i1} = \ln\left(\frac{\mu^c_{i+1}}{(1-\mu^c_{i+1})}\right) = \boldsymbol{X}^\beta_{i(1)}\boldsymbol{\beta}, \tag{6.13}$$

$$\eta_{i2} = \ln\left(\frac{\mu^c_{i1+}}{(1-\mu^c_{i1+})}\right) = \boldsymbol{X}^\beta_{i(2)}\boldsymbol{\beta}, \tag{6.14}$$

$$\eta_{i3} = \ln\psi_i = \ln\left(\frac{\mu^c_{i11}(1-\mu^c_{i1+}-\mu^c_{i+1}+\mu^c_{i11})}{(\mu^c_{i1+}-\mu^c_{i11})(\mu^c_{i+1}-\mu^c_{i11})}\right) = \boldsymbol{X}^\beta_{i(3)}\boldsymbol{\beta}. \tag{6.15}$$

Here $\boldsymbol{X}^\beta_{i(t)}$, $t = 1, \ldots, 3$ is a row vector containing design and covariate information. Their union is $\boldsymbol{X}^\beta_i$. The model extends naturally to multiple ordinal outcomes (Molenberghs and Lesaffre 1994). Explicit solutions for the probabilities can be found:

$$\mu^c_{i1+} = \frac{\exp(\eta_{i1})}{1+\exp(\eta_{i1})},$$

$$\mu^c_{i+1} = \frac{\exp(\eta_{i2})}{1+\exp(\eta_{i2})},$$

and

$$\mu^c_{i11} = \begin{cases} \frac{1+(\mu^c_{i1+}+\mu^c_{i+1})(\psi_i-1)-S(\mu^c_{i1+},\mu^c_{i+1},\psi_i)}{2(\psi_i-1)} & \text{if } \psi_i \neq 1, \\ \mu^c_{i1+}\mu^c_{i+1} & \text{if } \psi_i = 1, \end{cases}$$

with

$$S(q_1, q_2, \psi) = \sqrt{[1 + (q_1 + q_2)(\psi - 1)]^2 + 4\psi(1 - \psi)q_1 q_2}.$$

The above expression was studied by Plackett (1965), Mardia (1970) and Dale (1986). Details on the estimation of the covariance matrix can be found in Molenberghs and Lesaffre (1994, 1999). Molenberghs and Lesaffre (1994) extended the Dale model to multivariate ordinal outcomes. They generalized the computations of the bivariate Plackett distribution in order to establish the multivariate cell probabilities. The Plackett distribution is also used in GEEs when the odds ratio is used to measure the association (Lipsitz, Laird and Harrington 1991). Thus, the Dale model combines logistic regression for each of the measurements with marginal global odds ratios to describe the association between outcomes. It belongs to the family of marginal measurement models (Liang, Zeger and Qaqish 1992).

Often, one is interested in model parameters for the full set of repeated outcomes. Little (1993, 1995) proposes the use of identifying restrictions: identify unknown probabilities by equating them to functions of known probabilities. In our bivariate setting, we identify $\mu^c_{ik|1j}$ by equating them to appropriate functions of μ_{i2jk}. The simplest example is the set termed complete case missing value (CCMV) restrictions: $\mu^c_{ik|1j} = \mu^c_{ik|2j} = \mu_{ik|2j}$. Other restrictions are discussed by Little (1993). Another set of restrictions is defined in Molenberghs, Michiels, Kenward and Diggle (1998). They define the available case missing value (ACMV) restrictions, which are equivalent to MAR. These restrictions have already been introduced in Chapter 5. It is also stated that ACMV and CCMV coincide in the simple case of two time points. In some settings, such as a bivariate normal sample, restrictions are very natural to apply, because both the marginal distribution of the first measurement, as well as the conditional distribution of the second measurement given the first one, can be expressed as simple functions of the mean vector and the covariance matrix components. For categorical data in general, and the Dale model in particular, there is no easy transition from marginal to conditional distributions in terms of the

model parameters. In order to apply ACMV to the Dale model, we have to proceed in a different way.

First, the minimal approach is followed in the sense that a bivariate Dale model for the complete pattern is combined with a univariate logistic model for the incomplete pattern. From this approach $\hat{\boldsymbol{\beta}}_0^P$ and $\hat{\boldsymbol{\alpha}}_0^P$ follow and hence the underlying probabilities $\hat{\mu}_{ijk|2}$ and $\hat{\mu}_{ij|1}$ can be estimated. Then, ACMV implies that $\hat{\mu}_{ik|1j} \equiv \hat{\mu}_{ik|2j}$ and hence the partial count $Z_{ij|1}$ can be used to impute $Z^*_{ijk|1} = Z_{ij|1}\hat{\mu}_{ik|2j}$. From these completed counts and $Z^c_{ijk|2}$, one can estimate the parameters of interest, in our example a Dale model for both patterns, yielding $\hat{\boldsymbol{\beta}}^P$ and $\hat{\boldsymbol{\alpha}}^P \equiv \hat{\boldsymbol{\alpha}}_0^P$.

Above two-step procedure is clearly not restricted to the Dale model. Furthermore, extension to more than two measurement occasions is straightforward, certainly in the case of monotone dropout. Although parameter estimation is very elegant and computationally simple with the two-step procedure, precision estimation is less simple. Indeed, treating the filled-in table as if it represented observed data fails to reflect random variability in the unobserved counts. Strategies to determine confidence intervals will be discussed in Section 6.2.6.

6.2.5 Relative Merits of Both Families

It is worthwile to consider the reason why pattern-mixture models are tied to restrictions, whereas selection models apparantly are not. It is useful to start our discussion with the MAR case. For the selection model, such a mechanism entails $\phi^c_{id|jk} = \phi^c_{id|j}$. For a pattern-mixture model, it implies $\mu^c_{ik|1j} = \mu^c_{ik|2j}$ (Molenberghs, Michiels, Kenward and Diggle 1998). In other words, MAR naturally translates into assumptions about the dropout probabilities in a selection model, but into a restriction in the pattern-mixture section. Then, data to estimate $\phi^c_{id|jk}$ (in particular $\phi^c_{id|j}$) are available, but the data to estimate $\mu^c_{ik|1j}$ are not.

However, it is important to understand that both are different faces of the same coin and that in both cases this assumption is untestable. While this is clearly true for the pattern-mixture models, it is less obvious for the selection models, since wide classes of models for $\phi^c_{id|jk}$ are estimable. However, in order to correctly test for MAR, one would need to observe both measurements in both patterns, which

is by definition impossible. See also Glynn, Laird and Rubin (1986), as well as the discussion.

The same is true for non-random missingness mechanisms. For pattern-mixture models, MNAR mechanisms are reflected by different restrictions (e.g., protective restrictions, see Chapter 4). For selection models, MNAR is encompassed by models for $\phi^c_{id|jk}$ that depend explicitly on k. In Molenberghs and Goetghebeur (1997) it is seen how two non-random selection models can be supported by the observed data almost equally, but yield radically different interpretations for the unobserved data, in the sense that different models distribute an observed count Z_{i1j} in entirely different ways over the full data cells Z^c_{i1jk}.

An advantage of pattern-mixture models in the context of non-random dropout, quoted by Little (1995), is that no explicit model for the dropout process is needed, as long as the restrictions imposed are acceptable. However, this claim is slightly deceptive, since there is no symmetry between the $\boldsymbol{\phi}$ parameters in the two families. In a selection model, $\phi^c_{id|jk}$ contains all information about the dropout process, whereas the same information is spread out over ϕ_{id} and $\mu^c_{ijk|d}$ in a pattern-mixture model. This is seen through the fact that MAR is emanated by the ϕ's in the first case but by the μ's in the latter. Furthermore, the interdependence between dropout and measurement processes is modelled in $\phi^c_{id|jk}$ in the first case and in $\mu^c_{ijk|d}$ in the latter one. The pattern-mixture dropout probabilities ϕ_{id} can be seen as the "covariate dependent" part of the dropout mechanism.

Arguably, a framework has to be chosen based on the questions of scientific interest. For instance, in case one is interested in the population as a whole, a selection model might be the natural choice. However, investigators who would like to explore differences among subgroups that are identified by their response patterns, should consider fitting pattern-mixture models. The latter situation could be of interest to differentiate therapies between subgroups. For instance, if males would suffer more from dropout than females, one may want to establish sex dependent treatment protocols.

6.2.6 Precision Estimation with Pattern-Mixture Models

We propose two methods to calculate 95% confidence intervals: profile likelihood and multiple imputation.

Let us discuss profile likelihood (Clayton and Hills 1993, Welsh 1996) first. For each component β_i^P of the measurement parameter vector $\boldsymbol{\beta}^P$, the profile likelihood is constructed by keeping β_i^P fixed and maximizing the observed data log-likelihood

$$\ell(\boldsymbol{\beta}^P) = \sum_{i=1}^{N} \sum_{j=1}^{c} \left(\sum_{k=1}^{c} Z_{i2jk} \log \nu_{i2jk}(\boldsymbol{\beta}^P, \boldsymbol{\alpha}^P) + Z_{i1j} \log \nu_{i1j}(\boldsymbol{\beta}^P, \boldsymbol{\alpha}^P) \right)$$

with respect to the remaining parameters. In particular, lower and upper bounds β_{il}^P and β_{iu}^P of a 95% confidence interval for $\hat{\beta}_i^P$ are found by solving $2(\ell(\hat{\boldsymbol{\beta}}^P) - \ell(\hat{\boldsymbol{\beta}}_{(i)}^P)) = \chi_1^2(0.05)$, where $\hat{\boldsymbol{\beta}}_{(i)}^P$ is the constrained maximization over $\beta_i^P = \beta_{il}^P$ or $\beta_i^P = \beta_{iu}^P$ and $\chi_1^2(0.05)$ is the 95% quantile of the χ^2 distribution with a single degree of freedom. The advantage of profile likelihood is that it is able to reflect asymmetry in the log-likelihood function.

Alternatively, multiple imputation (see Section 3.2.2) can be used to construct an asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}^P$, from which asymptotic 95% confidence intervals readily follow.

1. Draw $\boldsymbol{\gamma}^*$ from the posterior distribution of $\boldsymbol{\gamma}$. In our case, the vector $\boldsymbol{\gamma}$ are just the parameters of interest $\boldsymbol{\beta}^P$, and we approximate the posterior distribution by a normal.

2. Draw $\boldsymbol{Z}_i^c$ from $f(\boldsymbol{Z}_i^c | \boldsymbol{Z}_i, \boldsymbol{\gamma}^*)$. This is most easily done by using a unifom random number generator (see Section 3.2.2) to divide Z_{i1j} over the cells Z_{i1jk}^c, $k = 1, \ldots, c$.

3. Use the completed data $\boldsymbol{Z}_i^c$ and the model to estimate the parameter of interest $\boldsymbol{\beta}^{P*}$ and its variance $\boldsymbol{\Sigma}(\boldsymbol{\beta}^{P*})$, called the within-imputation variance.

These three steps are repeated independently M times, resulting in $\boldsymbol{\beta}_m^{P*}$, $\boldsymbol{\Sigma}(\boldsymbol{\beta}_m^{P*})$, $m = 1, \ldots, M$.

Finally, we combine these estimates as described in Section 3.2.2. Since in our case the number of imputations is large, we can certainly rely on the corresponding normal approximation to obtain 95% confidence intervals.

These two methods do not need to give the same results for the variances. Apart from sampling variation, introduced through multiple imputation, and different reference distribution approximations, the main difference is that multiple imputation based confidence intervals are symmetric by construction, while profile likelihood confidence intervals are not.

6.3 Analysis of Fluvoxamine Data

We will analyse the data presented in Section 2.1. The observations at times 2 and 5 will be used. Since we will use the covariates as well, only 293 patients are included in the study. The outcomes can be found in Table 2.4.

6.3.1 Selection Modelling for Side Effects

Table 6.1 represents parameter estimates and asymptotic confidence intervals for a selection model including *age, sex* and *psychiatric antecedents*, both into the marginal measurement model as well as in the logistic model for dropout. Note that *age* is a continuous covariate, while the other two are dichotomous. To allow for MAR, the first response is also entered in the dropout model. The association is modelled in terms of a constant log odds ratio. The model leads to the marginal log odds of *no* side effects at both occasions, and the log of the probability of *no* dropout.

In the marginal model, *sex* and *antecedents* seem to have little effect, while *age* is borderline and its coefficients at both measurement occasions are very similar. Likewise, *age* and *antecedents* add little to the dropout model, and further *sex* and the outcome at the first occasion are borderline, albeit at different sides of the critical level. The association between both measurements, even with adjustment of the marginal regression for covariate effects, remains very high, with an odds ratio of $\exp(2.038) = 7.675$.

Table 6.1: Fluvoxamine Data, Side Effects: Selection Model (full)

Parameter	Estimate	Confidence Interval		Estimate	Confidence Interval	
		Lower	Upper		Lower	Upper
	Measurement Model					
	First Measurement			Last Measurement		
intercept	0.786	-0.083	1.654	1.432	0.397	2.467
age (/30)	-0.669	-1.218	-0.119	-0.676	-1.318	-0.034
sex	-0.318	-0.811	0.175	0.254	-0.337	0.846
antecedents	0.134	-0.366	0.633	-0.057	-0.649	0.536
	Association					
log odds ratio	2.038	1.335	2.740			
	Dropout Model					
intercept	1.583	0.571	2.595			
previous	-0.556	-1.119	0.007			
age (/30)	-0.261	-0.874	0.352			
sex	0.608	0.052	1.164			
antecedents	-0.254	-0.836	0.327			

Some simplification of the model is clearly necessary. A backward selection procedure was followed on the measurement and dropout processes separately. At each step, one or two parameters were removed based on a likelihood ratio test. Parameters were removed in the following order. For the measurement model: both *antecedents* effects and both *sex* effects were removed. Subsequently, the two *age* parameters were combined into a common *age* effect. For the dropout model: *age* and *antecedents* were removed. The result is shown in Table 6.2. From this model, it is seen that the probability of side effects is higher at the first measurement occasion than at the last one, and increases with *age*. In particular, for an increase of 1 year, the odds of side effects increases with a factor $\exp(0.664/30) = 1.022$, because *age* was divided by 30 for ease of display of the estimates. The probability of dropout is higher if side effects are observed at the first occasion, and is higher for males than

Table 6.2: Fluvoxamine Data, Side Effects: Selection Model (reduced)

Parameter	Estimate	Confidence Interval		Estimate	Confidence Interval	
		Lower	Upper		Lower	Upper
Measurement Model						
	First Measurement			Last Measurement		
intercept	0.661	-0.043	1.365	1.560	0.823	2.297
common age (/30)	-0.664	-1.141	-0.188			
Association						
log odds ratio	1.956	1.270	2.642			
Dropout Model						
intercept	1.085	0.547	1.624			
previous	-0.584	-1.140	-0.028			
sex	0.568	0.025	1.110			

for females. In particular, the dropout probabilities are 0.256 (0.161) for females with (without) previous side effects, and 0.377 (0.253) for males with (without) side effects. The association, as well as the other parameters, except for the intercept, are similar to the ones found in Table 6.1.

Figure 6.1 plots *age* versus the probability of side effects on the first, the last, and on either occasion (union probabilities). Each plot shows the response for males and females as predicted by the model, labelled marginal (male) and marginal (female). As *sex* has disappeared from the marginal model, both profiles obviously coincide. From the predicted probabilities $\hat{\nu}^c_{idjk}$, we can compute the response profile for the completers and dropout groups separately, also shown in these plots. Because *sex* is part of the dropout model, the curves for males and females separate in this case. Even though the marginal probabilities for males and females are the same, the chance of side effects is higher for males in both groups separately.

Figure 6.1: Fluvoxamine Data, Selection Model: Probabilities of Side Effects w.r.t. Age (a) at the First Occasion, (b) at the Last Occasion, (c) at Any Occasion

6.3.2 Pattern-Mixture Modelling for Side Effects

For the pattern-mixture approach, the parameter estimates and confidence intervals for the variables *age*, *sex* and *antecedents* can be found in Table 6.3. The model is parametrized as follows: intercepts and covariate effects are given for the complete observations, together with the differences between effects for incomplete and complete observations. The latter ones would be zero if the distribution among completers would equal the distribution among dropouts. This model is used for the first as well as for the last observation. A constant log odds ratio is assumed for

the association between both measurements. The confidence intervals are calculated using profile likelihood, and using multiple imputation. For the multiple imputation technique, the results are given for 100 imputations. We have also calculated the confidence intervals for 1000 and 4100 imputations; the differences were minor, but of course the computing time increased accordingly. Although some differences were anticipated, both methods to calculate confidence intervals gave approximately the same results. The same variables are used to fit the dropout model, and because the data needed to estimate this model are complete, we have calculated confidence intervals based on the asymptotic variance. Although multiple imputation is only performed to estimate the precision, we also display the corresponding parameter estimates as an extra indication for convergence of the algorithm.

Antecedents and *sex* have nearly no effect on the measurement model, but the *sex* parameter for the first measurement gives a borderline influence. *Age* has an effect on the measurement outcomes, but there is no difference between this effect for the complete and incomplete observations. The association between both measurements is very strong. The odds ratio is $\exp(2.038) = 7.675$. *Age* and *antecedents* have no effect on the dropout model, but *sex* has.

Again, we simplified our model using a backward selection procedure. The selection was based on a likelihood ratio test. For the measurement model, we dropped *antecedents*, the additional *age* effect for the incomplete observations, and all the *sex* effects. Finally, a joint *age* effect for both time points is assumed. In the dropout model, *antecedents* and *age* were removed. *Sex* was kept, although it is borderline. The final model can be found in Table 6.4.

From this model one can see that the probability of dropout is higher for males than for females: 0.253 and 0.168 respectively. The probability of having side effects is higher at the first occasion than at the last, and increases for those who did not show up at the last visit. This probability also increases with *age.* For an increase of 1 year, the odds of having side effects increases with 1.022. The association is similar to its value in the full model, found in Table 6.3.

Figure 6.2 plots *age* versus the probability of having side effects at the first, the last, and at either occasion. We have plotted the responses for males and females

Table 6.3: Fluvoxamine Data, Side Effects: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (full)

Parameter	Method	Estimate	Confidence Interval		Estimate	Confidence Interval	
			Lower	Upper		Lower	Upper
Measurement Model							
		First Measurement			Last Measurement		
	Complete Observations						
intercept	PL	1.296	0.289	2.339	1.664	0.616	2.767
	MI	1.296	0.268	2.325	1.663	0.596	2.731
age (/30)	PL	-0.849	-1.519	-0.203	-0.756	-1.440	-0.091
	MI	-0.849	-1.500	-0.198	-0.756	-1.414	-0.097
sex	PL	-0.593	-1.189	-0.007	0.127	-0.497	0.739
	MI	-0.593	-1.182	-0.004	0.127	-0.483	0.737
antecedents	PL	0.222	-0.353	0.805	-0.016	-0.634	0.594
	MI	0.222	-0.357	0.800	-0.016	-0.621	0.589
	Incomplete Minus Complete Observations						
intercept	PL	-2.151	-4.300	-0.084	-0.913	-4.376	3.204
	MI	-2.156	-4.224	-0.087	-1.018	-4.393	2.357
age (/30)	PL	0.869	-0.396	2.142	0.366	-1.845	2.435
	MI	0.871	-0.396	2.139	0.395	-1.503	2.292
sex	PL	0.879	-0.268	2.050	0.382	-1.413	2.236
	MI	0.879	-0.274	2.033	0.347	-1.477	2.171
antecedents	PL	-0.234	-1.428	0.986	-0.107	-2.271	1.802
	MI	-0.234	-1.439	0.970	-0.012	-1.858	1.834
	Association						
log odds ratio	PL	2.038	1.354	2.789			
	MI	2.065	1.346	2.784			
Dropout Model, Confidence Intervals Based on Asymptotic Variance (AV)							
intercept	AV	1.390	0.450	2.370			
age (/30)	AV	-0.349	-0.953	0.255			
sex	AV	0.559	0.010	1.108			
antecedents	AV	-0.232	-0.809	0.345			

for both the complete and incomplete groups, as predicted by the model. Also the marginal probabilities are calculated and plotted. Because *sex* is not part of the dropout model, the plots for males and females coincide if we look at the completers and dropouts separately. But if we look at the marginal probabilities, a difference

Table 6.4: Fluvoxamine Data, Side Effects: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (reduced)

Parameter	Method	Estimate	Confidence Interval		Estimate	Confidence Interval	
			Lower	Upper		Lower	Upper
Measurement Model							
		First Measurement			Last Measurement		
	Complete Observations						
intercept	PL	0.762	0.036	1.478	1.590	0.846	2.333
	MI	0.747	0.029	1.466	1.576	0.836	2.315
	Incomplete Minus Complete Observations						
intercept	PL	-0.499	-1.065	0.050	-0.268	-1.123	0.704
	MI	-0.499	-1.055	0.056	-0.275	-1.071	0.521
	Common Age Effect						
	PL	-0.650	-1.132	-0.162			
	MI	-0.639	-1.121	-0.158			
	Association						
log odds ratio	PL	1.977	1.291	2.682			
	MI	1.943	1.263	2.623			
Dropout Model, Confidence Intervals Based on Asymptotic Variance (AV)							
intercept	AV	0.766	0.353	1.179			
sex	AV	0.517	-0.021	1.056			

appears. Marginally we can see that the probability of side effects is higher for males than for females. This difference is due to the fact that *sex* is part of the dropout model.

Finally, note that the pattern-mixture model assumed a common odds ratio among completers and dropouts. This implies that the conditional distribution of the missing second measure follows the same conditional distribution given the first variable as do the complete variable. This ACMV/CCMV restriction, as discussed in Section 6.2.4, is equivalent to the MAR assumption in the selection model.

Figure 6.2: Fluvoxamine Data, Pattern-Mixture Model: Probabilities of Side Effects w.r.t. Age (a) at the First Occasion, (b) at the Last Occasion, (c) at Any Occasion

6.3.3 Comparison for Side Effects

Both reduced models include *age* as a predictor for side effects. For the selection model, this effect is the same at both measurement occasions. The same is true for the pattern-mixture model and although it could in principle differ for completers and dropouts, it is the same for both subgroups. Due to the latter fact, the estimates of *age* effects in both frameworks become comparable and their numerical values are indeed very close. By construction, the association parameters are also comparable; they are certainly of the same magnitude. As for the dropout models, they are

different because only in a selection model can one include measurements into the dropout part. The *sex* effect is similar in both models, but its effect is borderline.

Let us now compare Figures 6.1 and 6.2. Qualitatively, most of the conclusions are very similar. But because the frameworks differ in the choice of distributions that are modelled directly, slight differences are to be expected. For example, in our selection model, the marginal curves for males and females coincide, whereas the response pattern specific curves for males and females coincide in the pattern-mixture version. This can be explained as follows. In our selection model, the measurement distribution is independent of *sex,* but the conditional curves are based upon the probabilities

$$\nu^c_{jk|d} = \frac{\mu^c_{jk}\phi^c_{d|jk}}{\sum_{d=1}^{2}\mu^c_{jk}\phi^c_{d|jk}}$$

and since $\phi^c_{d|jk}$ depends on *sex,* the curves necessarily do too. A similar argument explains why the marginal curves are *sex* dependent in the pattern-mixture model. However, even for the curves that differ with *sex*, the discrepancy is very small.

A noteworthy feature is that the marginal probability of side effects is slightly higher for males than for females in the pattern-mixture model (Figure 6.1) and equal in the selection model, whereas the conditional probability of side effects given the non-response pattern (completers and dropouts) is higher for males than for females in the selection model (Figure 6.2), and equal in the pattern-mixture model. Of course, we should not forget that the differences between separating curves are not big, since *sex* disappeared from the measurement models in both frameworks.

It is important to note that the pattern-mixture model can yield valuable insight in its own right. Specifically, the probability of side effects, after adjusting for *age*, is higher in the dropout group than in the completers group, both at the first as well as at the last measurement occasion. For someone aged 30 say, the probabilities of side effects at the first measurement occasion for in the completers' group and the dropouts' groups are 0.4720 and 0.5956 respectively. At the last measurement occasion these probabilities are 0.2809 and 0.3380 respectively. These values can be obtained in a selection framework as well, but less straightforwardly. Another advantage of the pattern-mixture model is that the model building can be done for the different dropout groups separately. For example, if *sex* would be a prognostic

Table 6.5: Fluvoxamine Data, Therapeutic Effect: Selection Model (full)

Parameter	Estimate	Confidence Interval		Estimate	Confidence Interval	
		Lower	Upper		Lower	Upper
	Measurement Model					
	First Measurement			Last Measurement		
intercept	-3.282	-5.126	-1.438	0.638	-0.372	1.647
age (/30)	0.272	-0.812	1.355	0.169	-0.468	0.806
sex	0.883	-0.265	2.031	-0.243	-0.839	0.352
antecedents	-0.686	-1.656	0.284	-0.363	-0.948	0.222
	Association					
log odds ratio	2.009	-0.046	4.064			
	Dropout Model					
intercept	0.701	-0.653	2.055			
previous	0.725	-0.270	1.719			
age (/30)	-0.338	-0.944	0.269			
sex	0.603	0.048	1.158			
antecedents	-0.270	-0.852	0.312			

factor for side effects in the dropout group but not in the completers group, this is easily incorporated in the pattern-mixture analysis.

6.3.4 Selection Modelling for Therapeutic Effect

For the therapeutic effect, the same covariates as for the analysis of the side effects are included: *age*, *sex* and *psychiatric antecedents*. The parameter estimates and asymptotic confidence intervals are represented in Table 6.5. To allow for MAR, the first response is also entered in the dropout model. The association is again modelled in terms of a constant log odds ratio.

In the marginal measurement model, only the intercepts seem to have an effect on the therapeutic effect. The association between both measurements, even with adjustment of the marginal regression for covariate effects, remains very high, with an odds ratio of $\exp(2.009) = 7.456$. In the dropout model, only the *sex* effect seems

Table 6.6: Fluvoxamine Data, Therapeutic Effect: Selection Model (reduced)

Parameter	Estimate	Confidence Interval		Estimate	Confidence Interval	
		Lower	Upper		Lower	Upper
	Measurement Model					
	First Measurement			Last Measurement		
intercept	-2.669	-3.134	-2.204	0.469	0.198	0.740
	Association					
log odds ratio	2.016	0.114	3.918			
	Dropout Model					
intercept	1.085	0.821	1.349			

to be significant, although borderline.

But to come to more thorough conclusions, a backwards selection procedure was performed, based on the likelihood ratio test. The following parameters were removed: for the measurement model the *age*, *sex* and *antecedents* effects at both occasions; for the dropout model the *age*, *antecedents*, *previous* and *sex* effects. So only the intercepts and the association remain in the model. This result is shown in Table 6.6. From this model, it is seen that the probability of therapeutic effect is higher at the first measurement occasion (0.926) than at the last one (0.385). The dropout probability is 0.747, independent of the previous measurement. All these probabilities are independent from the patients' age, gender and psychiatric antecedents. The association is similar to the one found in Table 6.5.

6.3.5 Pattern-Mixture Modelling for Therapeutic Effect

Next, we analysed the therapeutic effect using a pattern-mixture model. The parameter estimates and confidence intervals for the variables *age, sex* and *antecedents* in the pattern-mixture model can be found in Table 6.7.

The model is parametrized as follows: intercepts and covariate effects are given for the complete observations, together with the differences between effects for incomplete and complete observations. The latter ones would be zero if the distribu-

Table 6.7: Fluvoxamine Data, Therapeutic Effect: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (full)

Parameter	Method	Estimate	Confidence Interval		Estimate	Confidence Interval	
			Lower	Upper		Lower	Upper
Measurement Model							
		First Measurement			Last Measurement		
	Complete Observations						
intercept	PL	-4.296	-7.291	-1.809	0.566	-0.588	1.743
	MI	-4.297	-6.644	-1.951	0.566	-0.447	1.580
age (/30)	PL	0.951	-0.608	2.514	0.213	-0.524	0.968
	MI	0.951	-0.376	2.278	0.213	-0.428	0.855
sex	PL	0.431	-1.017	2.245	-0.281	-0.978	0.397
	MI	0.431	-0.931	1.792	-0.281	-0.879	0.318
antecedents	PL	-0.415	-1.811	1.063	-0.334	-1.015	0.331
	MI	-0.415	-1.648	0.819	-0.334	-0.922	0.254
	Incomplete Minus Complete Observations						
intercept	PL	2.604	-2.394	7.109	0.323	-4.188	6.994
	MI	2.645	-1.366	6.655	0.493	-3.363	4.350
age (/30)	PL	-1.779	-4.664	0.855	-0.170	-3.171	2.852
	MI	-1.810	-4.251	0.632	-0.278	-2.595	2.039
sex	PL	1.365	-1.409	5.231	0.143	-2.591	2.612
	MI	1.372	-1.191	3.935	0.205	-1.647	2.057
antecedents	PL	-0.807	-3.351	1.575	-0.138	$-\infty$	2.387
	MI	-0.813	-2.931	1.305	-0.196	-2.338	1.946
	Association						
log odds ratio	PL	2.018	0.137	5.561			
	MI	2.132	0.020	4.243			
Dropout Model, Confidence Intervals Based on Asymptotic Variance (AV)							
intercept	AV	1.390	0.450	2.370			
age (/30)	AV	-0.349	-0.953	0.255			
sex	AV	0.559	0.010	1.108			
antecedents	AV	-0.232	-0.809	0.345			

tion among completers would equal the distribution among dropouts, i.e., MCAR. A similar model is used for the first as well as for the last observation. A constant log odds ratio is assumed for the association between both measurements. The confidence intervals are calculated using profile likelihood, and using multiple impu-

tation. Although multiple imputation is only performed to estimate the precision, the parameter estimates are produced as a by-product and hence they are displayed as an extra indication for convergence of the algorithm. Results for the two approaches are comparable except for the striking difference in the case of *antecedents* (last measurement, difference between both patterns), where profile likelihood yields an unbounded interval, reflecting that one tail of the likelihood levels off at a value close to the maximum. For the multiple imputation technique, the results are given for 100 imputations. As a check, we also calculated the confidence intervals for 10 and 1000 imputations, leading to negligible differences. These results can be found in Table 6.8.

The same variables are used to fit the dropout model, and because the data needed to estimate this model are complete, we have calculated confidence intervals based on the asymptotic variance.

In the measurement model, *antecedents*, *age* and *sex* have nearly no effect. The association between both measurements is very strong. The odds ratio is $\exp(2.018) = 7.523$. *Age* and *antecedents* have no effect on the dropout model, but *sex* has a (borderline) significant influence.

We reduced our model using a backward selection procedure. For the measurement model, we dropped *antecedents*, *age* and *sex* effects for the first and last observations, both for the complete observations and for the difference between the incomplete and complete observations. In the dropout model, *antecedents*, *age* and *sex* were removed, although the latter was borderline. The final model can be found in Table 6.9.

The probability of therapeutic effect is much higher at the first occasion than at the last, and decreases a little for those who did not show up at the last visit: for a person with two observations, the probabilities of therapeutic effect at the first and last observations are respectively 0.945 and 0.388; for a person who drops out after the first measurement, these probabilities are respectively 0.905 and 0.366. The dropout probability is 0.747. The association is similar to its value in the full model, found in Table 6.7.

Table 6.8: Fluvoxamine Data, Therapeutic Effect: Pattern-Mixture Model, Multiple Imputation (MI) (full)

Parameter	AV	MI(10)	MI(100)	MI(1000)
Completers, First Measurement				
intercept	-4.296(1.197)	-4.295(1.197)	-4.297(1.197)	-4.297(1.197)
age (/30)	0.951(0.677)	0.951(0.677)	0.951(0.677)	0.951(0.677)
sex	0.431(0.695)	0.431(0.695)	0.431(0.695)	0.431(0.695)
antecedents	-0.415(0.629)	-0.416(0.629)	-0.415(0.629)	-0.415(0.629)
Completers, Last Measurement				
intercept	0.566(0.517)	0.566(0.517)	0.566(0.517)	0.566(0.517)
age (/30)	0.213(0.327)	0.213(0.327)	0.213(0.327)	0.213(0.327)
sex	-0.281(0.305)	-0.281(0.305)	-0.281(0.305)	-0.281(0.305)
antecedents	-0.334(0.300)	-0.334(0.300)	-0.334(0.300)	-0.334(0.300)
Dropouts - Completers, First Measurement				
intercept	2.604(2.040)	2.611(2.061)	2.645(2.046)	2.629(2.051)
age (/30)	-1.779(1.236)	-1.764(1.250)	-1.810(1.246)	-1.802(1.246)
sex	1.365(1.303)	1.374(1.308)	1.372(1.308)	1.407(1.315)
antecedents	-0.807(1.071)	-0.868(1.080)	-0.813(1.081)	-0.849(1.081)
Dropouts - Completers, Last Measurement				
intercept	0.323(1.045)	-0.177(1.842)	0.493(1.968)	0.328(1.789)
age (/30)	-0.170(0.638)	-0.144(1.019)	-0.278(1.182)	-0.150(1.088)
sex	0.143(0.583)	0.081(1.085)	0.205(0.945)	0.112(0.958)
antecedents	-0.138(0.621)	0.289(1.193)	-0.196(1.093)	-0.155(1.052)
Association				
log odds ratio	2.018(0.853)	1.992(1.051)	2.132(1.077)	2.140(1.091)
Dropout Model				
intercept	1.390 (0.500)			
age (/30)	-0.349 (0.308)			
sex	0.559 (0.280)			
antecedents	-0.232 (0.294)			

Table 6.9: Fluvoxamine Data, Therapeutic Effect: Pattern-Mixture Model, Profile Likelihood (PL) and Multiple Imputation (MI) (reduced)

Parameter	Method	Estimate	Confidence Interval		Estimate	Confidence Interval	
			Lower	Upper		Lower	Upper
Measurement Model							
		First Measurement			Last Measurement		
	Complete Observations						
intercept	PL	-2.848	-3.588	-2.241	0.455	0.147	0.770
	MI	-2.848	-3.430	-2.266	0.455	0.183	0.727
	Incomplete Minus Complete Observations						
intercept	PL	0.589	-0.593	1.677	0.093	-1.021	1.540
	MI	0.589	-0.383	1.561	0.047	-0.843	0.937
	Association						
log odds ratio	PL	1.992	0.154	5.550			
	MI	2.099	0.059	4.139			
Dropout Model, Confidence Intervals Based on Asymptotic Variance (AV)							
intercept	AV	1.085	0.821	1.349			

6.3.6 Comparison for Therapeutic Effect

The reduced measurement models in both frameworks are similar. No effect of *age*, *sex* or *antecedents* remain in the model. Only the intercepts were kept. By construction, the association parameters are comparable as well. The dropout models are exactly the same. Once the first observation is taken out of the model, both dropout models are equal.

It is important to note that the pattern-mixture model can yield valuable insight in its own right. Specifically, the probability of therapeutic effect is higher in the dropout group than in the completers group, both at the first as well as at the last measurement occasion. Since these differences are not significant, we know there is no difference in therapeutic effect between the completers and the dropout group. These values can be obtained in a selection framework as well, but less straightforwardly. Another advantage of the pattern-mixture model is that the model building

can be done for the different dropout groups separately.

6.3.7 Different Missing Data Mechanisms

For the pattern-mixture model, we assumed MAR as missing data mechanism to complete the data. Restricting this further to MCAR, we can complete the data in essentially two ways: with or without covariance dependance. We have done this for the side effects, leading to the results displayed in Table 6.10.

Since the observed data remain the same, the dropout models coincide. Furthermore, the parts based on the observed data (first and second margin for the completers, first margin for the dropouts) are very alike. The only differences appear in the parts where the 'filled in' data is used: the second margin for the dropouts (or in this case, the difference in second margin between the dropouts and the completers), and the association. Under MAR, the association is assumed to be the same for both the completers' and the dropout group, but the MCAR-assumptions neglects this association, which brings the log odds ratio much closer to 1 (=no association). Also in the fourth part (difference in second margin), the results differ for the three methods: assuming 'raw' MCAR, this part only depends on the random number generator used to fill in values; if one allows dependence on the covariates, this fourth part has nearly no influence, since the dependence is already captured in the second margin for the completers, which was used to fill in the data.

6.4 Conclusion

While selection models are much more prominent in the literature than pattern-mixture models, this chapter has shown that fitting pattern-mixture models is no more complex than fitting selection models. Discordant views in this matter are presumably inspired by the discrepancy between the volume of research devoted to both frameworks. Further, identifying restrictions in the pattern-mixture case play a very similar role to the modelling assumptions in the selection case. Indeed, for instance assuming MAR implies a particular form for the dropout mechanism in selection models and dictates a set of restrictions in pattern-mixture modelling.

Table 6.10: Fluvoxamine Data, Side Effects: Different Pattern-Mixture Models (full)

Parameter	MCAR	MCAR(X)	MAR
Completers, First Measurement			
intercept	1.294 (0.525)	1.294 (0.525)	1.296 (0.525)
age (/30)	-0.848 (0.332)	-0.848 (0.332)	-0.849 (0.332)
sex	-0.594 (0.301)	-0.594 (0.301)	-0.593 (0.301)
antecedents	0.224 (0.295)	0.224 (0.295)	0.222 (0.295)
Completers, Last Measurement			
intercept	1.669 (0.545)	1.669 (0.545)	1.664 (0.545)
age (/30)	-0.761 (0.336)	-0.761 (0.336)	-0.756 (0.336)
sex	0.126 (0.312)	0.126 (0.311)	0.127 (0.311)
antecedents	-0.010 (0.309)	-0.010 (0.309)	-0.012 (0.309)
Dropouts - Completers, First Measurement			
intercept	-2.144 (1.055)	-2.162 (1.055)	-2.151 (1.055)
age (/30)	0.862 (0.647)	0.874 (0.647)	0.868 (0.647)
sex	0.876 (0.589)	0.881 (0.589)	0.879 (0.589)
antecedents	-0.229 (0.614)	-0.234 (0.614)	-0.234 (0.614)
Dropouts - Completers, Last Measurement			
intercept	-0.682 (1.035)	0.093 (1.088)	-3.145 (1.122)
age (/30)	0.286 (0.629)	-0.056 (0.653)	1.148 (0.674)
sex	-0.314 (0.572)	-0.016 (0.593)	-0.381 (0.617)
antecedents	-0.020 (0.599)	0.004 (0.623)	0.126 (0.653)
Association			
log odds ratio	1.492 (0.285)	1.479 (0.292)	2.038 (0.309)
Dropout Model			
intercept	1.390 (0.500)	1.390 (0.500)	1.390 (0.500)
age (/30)	-0.349 (0.308)	-0.349 (0.308)	-0.349 (0.308)
sex	0.559 (0.280)	0.559 (0.280)	0.559 (0.280)
antecedents	-0.232 (0.294)	-0.232 (0.294)	-0.232 (0.294)

We analysed categorical outcomes using a Dale model. Both frameworks lead to similar results, and hence a particular framework can be chosen, depending on the scientific interest. In case one is interested in the population as a whole, a selection model might be the natural choice. However, investigators who would like to explore

differences among subgroups that are identified by their response patterns, should consider fitting pattern-mixture models instead. The latter situation could be of interest to differentiate therapies between subgroups. For instance, if males would suffer more from dropout than females, one may want to establish sex-dependent treatment protocols.

Further, it has been argued that fitting both a selection model as well as a pattern-mixture model can be a valuable sensitivity analysis tool. In our example, the conclusions reached under both formalisms are virtually identical, so that more confidence can be put into them. This points to the use of pattern-mixture models to assess sensitivity of selection models. Even without applying restrictions, pattern-mixture models are useful to assess the fit of an MAR selection model. Indeed, should the MAR assumption be violated, then an ignorable selection model is invalid, but an unstructered pattern-mixture model will correctly reflect differences between models for different patterns.

A sensitivity analysis based on a selection model and a pattern-mixture model, but then for continuous data, can be found in Chapter 8.

Chapter 7

Pseudo-Likelihood Estimation for a Combined Selection and Pattern-Mixture Model

In this chapter we develop pseudo-likelihood methods for the estimation of parameters in a model that is specified in terms of both selection modelling and pattern-mixture modelling quantities (Molenberghs, Michiels and Kenward 1998). When scientific interest focuses on both the structure of the non-response mechanism and the behavior of subjects given their response pattern, it is natural to study both of the quantities $f(\boldsymbol{r}|\boldsymbol{y})$ and $f(\boldsymbol{y}|\boldsymbol{r})$ simultaneously. This idea was formulated also by Holland (1986) and Wainer (1989), where essentially, apart from the observable parts of $f(\boldsymbol{y}|\boldsymbol{r})$, one also models $f(\boldsymbol{r}|\boldsymbol{y})$. Clearly, $f(\boldsymbol{r}|\boldsymbol{y})$ and $f(\boldsymbol{y}|\boldsymbol{r})$ cannot be combined straightforwardly into a likelihood function, except in the trivial MCAR case. When both components are derived from the same joint distribution, it follows from Gelman and Speed (1993) that they uniquely determine this distribution. We will consider two cases: (1) the model is specified directly from a joint model for the measurement and dropout processes; (2) conditional models for the measurement process given dropout and vice versa are specified directly. In the latter case, compatibility constraints to ensure the existence of a joint density are derived.

The goal of this chapter is to present methods for statistical inference in such conditionally specified distributions, subject to their non-response pattern. Note

that conditionally specified distributions are encountered in other areas of statistics as well. Standard applications are found in spatial statistics (Cressie 1991; Geman and Geman 1984), where the Hammersley-Clifford theorem (Besag 1974) is used to ensure existence of a valid probability model. Arnold, Castillo and Sarabia (1992) give a comprehensive treatment of conditionally specified distributions. But both of their constraints are of no use if one wants to specify the conditional models directly. Therefore new compatibility constraints are defined. We will focus on pseudo-likelihood estimation (Arnold and Strauss 1991; Geys, Molenberghs and Ryan 1997, 1999), a method that can operate directly on the conditional distributions.

In Section 7.1, pseudo-likelihood is reviewed and tailored to the needs of incomplete data problems. Section 7.2 illustrates these ideas using a trivariate loglinear model, in which case the results from the pseudo-likelihood are exactly the maximum likelihood estimates, as is shown in Theorem 3. We will show in Section 7.3 how progress is still possible when the conditional distributions are not compatible (i.e., do not necessarily correspond to a joint probability model). These methods are used in Section 7.4, where the fluvoxamine data introduced in Chapter 2 are analysed.

7.1 Pseudo-Likelihood

7.1.1 Definition and Properties

The full data log-likelihood can be written as

$$\ell(\boldsymbol{\theta}) \propto \sum_{s=1}^{N} \ln f(\boldsymbol{y}_s, \boldsymbol{r}_s | \boldsymbol{X}_s, \boldsymbol{\theta}).$$

As in Section 3.1.2, the full density can be written as

$$f(\boldsymbol{y}_s, \boldsymbol{r}_s | \boldsymbol{X}_s, \boldsymbol{\theta}) = f(\boldsymbol{y}_s | \boldsymbol{X}_s^{\beta}, \boldsymbol{\beta}) f(\boldsymbol{r}_s | \boldsymbol{y}_s, \boldsymbol{X}_s^{\alpha}, \boldsymbol{\alpha}) \tag{7.1}$$

in a selection modelling framework, and for a pattern-mixtuire model as

$$f(\boldsymbol{y}_s, \boldsymbol{r}_s | \boldsymbol{X}_s, \boldsymbol{\theta}) = f(\boldsymbol{y}_s | \boldsymbol{r}_s, \boldsymbol{X}_s^{\beta}, \boldsymbol{\beta}) f(\boldsymbol{r}_s | \boldsymbol{X}_s^{\alpha}, \boldsymbol{\alpha}). \tag{7.2}$$

One is often interested in both the dependence of the missing data process on the responses $f(\boldsymbol{r}_s|\boldsymbol{y}_s, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha})$ as well as in the pattern-specific average profiles $f(\boldsymbol{y}_s|\boldsymbol{r}_s, \boldsymbol{X}_s^\beta, \boldsymbol{\beta})$. Using the joint distribution, a choice has to be made, implying that one of these densities is described in terms of simple parameters, while the other has to be calculated using, for example, Bayes theorem. Therefore, pseudo-likelihood will be a useful alternative for full likelihood estimation. In our case, the log-pseudo-likelihood can be written as

$$p\ell(\boldsymbol{\theta}) \propto \sum_{i=s}^{N} \ln \left[f(\boldsymbol{r}_s|\boldsymbol{y}_s, \boldsymbol{X}_s^\alpha, \boldsymbol{\alpha}) f(\boldsymbol{y}_s|\boldsymbol{r}_s, \boldsymbol{X}_s^\beta, \boldsymbol{\beta}) \right].$$

Not all choices for the conditional models lead to a pseudo-likelihood that corresponds to a probability model. Arnold, Castillo and Sarabia (1992) discuss necessary and sufficient conditions in the case of two variables: to ensure the existence of an underlying likelihood, it is necessary to be able to write the ratio of the two components of the pseudo-likelihood as the product of two functions where the first one depends solely on the first component, and the other solely on the second component. In our situation, these two components are replaced by the measurement vector and the vector of dropout indicators respectively:

$$\frac{f(\boldsymbol{y}|\boldsymbol{r})}{f(\boldsymbol{r}|\boldsymbol{y})} = u(\boldsymbol{y})v(\boldsymbol{r}).$$

These constraints are clearly necessary. To indicate that they are sufficient note that, given the functions u and v, the marginal density for $\boldsymbol{y}$ can be expressed as:

$$\frac{u(\boldsymbol{y})}{\int\int f(\boldsymbol{r}|\boldsymbol{y})u(\boldsymbol{y})d\boldsymbol{r}d\boldsymbol{y}},$$

with a similar expression for the marginal density of $\boldsymbol{r}$. We will refer to these constraints as the *compatibility constraints*.

Arnold and Strauss (1991) have shown that such a pseudo-likelihood yields consistent point estimators. To estimate precision, purely model based standard errors are unacceptable, since they do not take into account that some information is used more than once and that, due to the compatibility constraints, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are not necessarily functionally independent. This results in an underestimation of the variance, and therefore a correction is needed to get consistent standard errors. This correction is made by calculating the robust variance estimator

(Michiels and Molenberghs 1995, 1997; Geys, Molenberghs and Ryan 1997, 1999): let $\boldsymbol{\lambda}$ be the union of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Write the log-pseudo-likelihood as $p\ell(\boldsymbol{\lambda}) = \sum_{s=1}^{N} p\ell_s(\boldsymbol{\lambda})$, where $p\ell_s$ is the contribution of the s-th subject to the log-pseudo-likelihood. Then $\sqrt{N}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ converges in distribution to the normal distribution $\boldsymbol{N}(0, \boldsymbol{J}(\boldsymbol{\lambda})^{-1}\boldsymbol{K}(\boldsymbol{\lambda})\boldsymbol{J}(\boldsymbol{\lambda})^{-1})$, where

$$\boldsymbol{J}(\boldsymbol{\lambda}) = \boldsymbol{E}\left(\frac{\partial^2 p\ell(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right)$$

and

$$\boldsymbol{K}(\boldsymbol{\lambda}) = \sum_{s=1}^{N} \boldsymbol{E}\left(\left(\frac{\partial p\ell_s(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right)\left(\frac{\partial p\ell_s(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right)'\right).$$

Similar ideas have been used in the context of generalized estimating equations (Liang and Zeger 1986). We have used this robust estimator also in Section 4.4.

7.1.2 Missing Data Mechanisms

Let us adapt the taxonomy of missing data mechanisms to the case of pseudo-likelihood, still restricting attention to dropout. The notation of these models is similar to the one in the previous chapter. If we assume MCAR, the pseudo-likelihood reduces to $f(d)f(\boldsymbol{y})$, and hence to the likelihood. MAR implies $f(d|\boldsymbol{y}) = f(d|\boldsymbol{y}^o)$. As shown in Chapter 5, this is equivalent to available case missing value restrictions (ACMV), implying that, for any $j \leq t$:

$$\begin{aligned} f(y_{s,t+1}|y_{s1},\ldots,y_{st},d_s=j) &= f(y_{s,t+1}|y_{s1},\ldots,y_{st},d_s\geq t+1) \qquad (7.3)\\ &= f(y_{s,t+1}|y_{s1},\ldots,y_{st}). \qquad (7.4)\end{aligned}$$

Using this assumption, the pseudo-likelihood reduces to

$$\begin{aligned} & f(d|\boldsymbol{y})f(\boldsymbol{y}|d)\\ &= f(d|\boldsymbol{y})\prod_{s=1}^{N}\left[\prod_{j=1}^{n}\left(f(y_{s1},\ldots,y_{sj}|d_s=j)\prod_{t=j}^{n-1}f(y_{s,t+1}|y_{s1},\ldots,y_{st},d_s=j)\right)\right]\\ &= f(d|\boldsymbol{y}^o)\prod_{s=1}^{N}\left[\prod_{j=1}^{n}\left(f(y_{s1},\ldots,y_{sj}|d_s=j)\prod_{t=j}^{n-1}f(y_{s,t+1}|y_{s1},\ldots,y_{st},d_s\geq t+1)\right)\right]. \qquad (7.5)\end{aligned}$$

All quantities in (7.5) can be calculated from the observed data.

7.2 A Trivariate Loglinear Model

To illustrate these concepts, consider the trivariate loglinear model (Cox 1972)

$$f(y_1, y_2, d) = \frac{1}{C}\exp(\alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 d + \alpha_4 y_1 y_2 + \alpha_5 y_1 d + \alpha_6 y_2 d + \alpha_7 y_1 y_2 d), \tag{7.6}$$

where C is the normalizing constant. For simplicity, subscripts s are suppressed. Assume the first outcome, Y_1, is always observed, and the second, Y_2, is possibly missing. The dropout indicator D is 1 if Y_2 is observed, and 0 otherwise. Both Y_1 and Y_2 are binary. Based on this model, we can calculate the two conditional densities needed for the pseudo-likelihood

$$f(y_1, y_2|d) = \frac{\exp(\alpha_1 y_1 + \alpha_2 y_2 + \alpha_4 y_1 y_2 + \alpha_5 y_1 d + \alpha_6 y_2 d + \alpha_7 y_1 y_2 d)}{1 + \exp(\alpha_1 + \alpha_5 d) + \exp(\alpha_2 + \alpha_6 d) + \exp(\alpha_1 + \alpha_2 + \alpha_4 + \alpha_5 d + \alpha_6 d + \alpha_7 d)}, \tag{7.7}$$

$$f(d|y_1, y_2) = \frac{\exp(\alpha_3 d + \alpha_5 y_1 d + \alpha_6 y_2 d + \alpha_7 y_1 y_2 d)}{1 + \exp(\alpha_3 + \alpha_5 y_1 + \alpha_6 y_2 + \alpha_7 y_1 y_2)}. \tag{7.8}$$

Note that, while these full conditionals have a simple form, the marginals such as $f(y_1, y_2)$ or $f(d)$ do not, rendering the construction of both selection and pattern-mixture models starting from (7.6) complicated. Since (7.7) and (7.8) stem from the same joint distribution, the compatibility constraints are automatically fulfilled.

Depending on the missing data mechanism, some of the parameters in (7.7) and (7.8) will be zero. Under the MAR/ACMV assumptions, the following conditions have to be satisfied: $f(y_2|y_1, d) = f(y_2|y_1)$ and $f(d|y_1, y_2) = f(d|y_1)$. From

$$f(y_2|y_1, d) = \frac{\exp(\alpha_2 y_2 + \alpha_4 y_1 y_2 + \alpha_6 y_2 d + \alpha_7 y_1 y_2 d)}{1 + \exp(\alpha_2 + \alpha_4 y_1 + \alpha_6 d + \alpha_7 y_1 d)}$$

and (7.8) it follows that both α_6 and α_7 need to be 0. Then, the pseudo-likelihood components reduce to

$$\begin{aligned} f(y_1, y_2|d) &= \frac{\exp(\alpha_1 y_1 + \alpha_2 y_2 + \alpha_4 y_1 y_2 + \alpha_5 y_1 d)}{1 + \exp(\alpha_1 + \alpha_5 d) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2 + \alpha_4 + \alpha_5 d)}, \\ f(d|y_1, y_2) &= \frac{\exp(\alpha_3 d + \alpha_5 y_1 d)}{1 + \exp(\alpha_3 + \alpha_5 y_1)}. \end{aligned}$$

We consider four sets of data to illustrate several possible cases. They are displayed in Table 7.1.

Table 7.1: Four Sets of Artificial Data. Each time a contingency table for the completers ($Y_1 = 0/1, Y_2 = 0/1$), and an additional contingency table for the dropouts ($Y_1 = 0/1$) is given.

(a)	50	25	25	
	25	50	25	
(b)	50	25	50	
	25	50	25	
(c)	50	25	75	
	25	50	25	
(d)	80	10	60	
	40	20	90	

We calculated the parameter values for the pseudo-likelihood model, together with their robust standard errors. These results are shown in Table 7.2. For reference, we have also included the results of the likelihood analysis. The first three sets of artificial data have the same counts for the completers, which is reflected in the equality of the estimates for α_2 and α_4.

From these results, one can observe that the parameter estimates and standard errors for the pseudo-likelihood and the likelihood models are exactly the same. This result is not specific to our datasets. It holds more generally for the trivariate loglinear model under MAR/ACMV:

Table 7.2: Parameter Estimates (Standard Errors) for the Artificial Data from Table 7.1, based on a Loglinear Model

Data	Param.	Pseudo-Likelihood	Likelihood
(a)	α_1	-0.6931(0.3367)	-0.6931(0.3367)
	α_2	-0.6931(0.2449)	-0.6931(0.2449)
	α_3	1.0986(0.2309)	1.0986(0.2309)
	α_4	1.3863(0.3464)	1.3863(0.3464)
	α_5	0.0000(0.3266)	0.0000(0.3266)
(b)	α_1	-1.3863(0.3055)	-1.3863(0.3055)
	α_2	-0.6931(0.2449)	-0.6931(0.2449)
	α_3	0.4055(0.1826)	0.4055(0.1826)
	α_4	1.3863(0.3464)	1.3863(0.3464)
	α_5	0.6931(0.2944)	0.6931(0.2944)
(c)	α_1	-1.7918(0.2944)	-1.7918(0.2944)
	α_2	-0.6931(0.2449)	-0.6931(0.2449)
	α_3	0.0000(0.1633)	0.0000(0.1633)
	α_4	1.3863(0.3464)	1.3863(0.3464)
	α_5	1.0986(0.2828)	1.0986(0.2828)
(d)	α_1	0.1178(0.1936)	0.1178(0.1936)
	α_2	-2.0794(0.3354)	-2.0794(0.3354)
	α_3	0.4055(0.1667)	0.4055(0.1667)
	α_4	1.3863(0.4430)	1.3863(0.4330)
	α_5	-0.8109(0.2357)	-0.8109(0.2357)

Theorem 3 *For a trivariate loglinear model $f(y_1, y_2, d)$, assuming missingness is random (MAR/ACMV), the estimates obtained from the score-equations based on the pseudo-likelihood $f(y_1, y_2|d)f(d|y_1, y_2)$ are equal to the maximum likelihood estimates, obtained from the score-equations from the likelihood.*

Proof of Theorem 3

Denote the observed data as n_{ijd}, $i, j, d \in \{0, 1, +\}$, where a "+" indicates one has to marginalize over that index. Let i indicate the result of the first, fully observed outcome; j indicates the result of the second outcome and d indicates the dropout value: $d = 1$ corresponds to an observation that has been observed twice; $d = 0$ indicates that subject dropped out after the first observation. So the observed data consists of six counts $(n_{001}, n_{011}, n_{101}, n_{111}, n_{0+0}, n_{1+0})$.

Using the following notation

$$\begin{aligned}
N_1 &= 1 + e^{\alpha_2} + e^{\alpha_1} + e^{\alpha_1+\alpha_2+\alpha_4}, \\
N_2 &= 1 + e^{\alpha_2} + e^{\alpha_1+\alpha_5} + e^{\alpha_1+\alpha_2+\alpha_4+\alpha_5}, \\
N_3 &= 1 + e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3} + e^{\alpha_1+\alpha_2+\alpha_4} + e^{\alpha_1+\alpha_3+\alpha_5} + e^{\alpha_2+\alpha_3} + e^{\alpha_1+\alpha_2+\alpha_3+\alpha_4+\alpha_5},
\end{aligned}$$

the pseudo-likelihood score-equations can be written as:

$$\begin{aligned}
n_{1++} &= n_{++0}\frac{e^{\alpha_1}(1+e^{\alpha_2+\alpha_4})}{N_1} \\
&\quad + n_{++1}\frac{e^{\alpha_1+\alpha_5}(1+e^{\alpha_2+\alpha_4})}{N_2}, \qquad (7.9)
\end{aligned}$$

$$\begin{aligned}
n_{+11} + n_{0+0}\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + n_{1+0}\frac{e^{\alpha_2+\alpha_4}}{1+e^{\alpha_2+\alpha_4}} &= n_{++0}\frac{e^{\alpha_2}(1+e^{\alpha_1+\alpha_4})}{N_1} \\
&\quad + n_{++1}\frac{e^{\alpha_2}(1+e^{\alpha_1+\alpha_4+\alpha_5})}{N_2}, \qquad (7.10)
\end{aligned}$$

$$n_{++1} = n_{0++}\frac{e^{\alpha_3}}{1+e^{\alpha_3}} + n_{1++}\frac{e^{\alpha_3+\alpha_5}}{1+e^{\alpha_3+\alpha_5}}, \qquad (7.11)$$

$$n_{111} + n_{1+0}\frac{e^{\alpha_2+\alpha_4}}{1+e^{\alpha_2+\alpha_4}} = n_{++0}\frac{e^{\alpha_1+\alpha_2+\alpha_4}}{N_1} + n_{++1}\frac{e^{\alpha_1+\alpha_2+\alpha_4+\alpha_5}}{N_2}, \qquad (7.12)$$

$$2n_{1+1} = n_{1++}\frac{e^{\alpha_3+\alpha_5}}{1+e^{\alpha_3+\alpha_5}} + n_{++1}\frac{e^{\alpha_1+\alpha_5}(1+e^{\alpha_2+\alpha_4})}{N_2} \qquad (7.13)$$

Based on the same notation, the likelihood score-equations can be written as:

$$n_{1++} = n_{+++}\frac{e^{\alpha_1}(1+e^{\alpha_2+\alpha_4})(1+e^{\alpha_3+\alpha_5})}{N_3}, \tag{7.14}$$

$$n_{+11} + n_{0+0}\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + n_{1+0}\frac{e^{\alpha_2+\alpha_4}}{1+e^{\alpha_2+\alpha_4}} = n_{+++}\frac{e^{\alpha_2}(1+e^{\alpha_1+\alpha_4}+e^{\alpha_3}+e^{\alpha_1+\alpha_3+\alpha_4+\alpha_5})}{N_3}, \tag{7.15}$$

$$n_{++1} = n_{+++}\frac{e^{\alpha_3}(1+e^{\alpha_1+\alpha_5}+e^{\alpha_2}+e^{\alpha_1+\alpha_2+\alpha_4+\alpha_5})}{N_3}, \tag{7.16}$$

$$n_{111} + n_{1+0}\frac{e^{\alpha_2+\alpha_4}}{1+e^{\alpha_2+\alpha_4}} = n_{+++}\frac{e^{\alpha_1+\alpha_2+\alpha_4}(1+e^{\alpha_3+\alpha_5})}{N_3}, \tag{7.17}$$

$$n_{1+1} = n_{+++}\frac{e^{\alpha_1+\alpha_3+\alpha_5}(1+e^{\alpha_2+\alpha_4})}{N_3}. \tag{7.18}$$

We will prove that, if the pseudo-likelihood score-equations (7.9)–(7.13) are fulfilled, the likelihood score-equations (7.14)–(7.18) are too.
Equation (7.11) leads to

$$n_{++0} = n_{0++}\frac{1}{1+e^{\alpha_3}} + n_{1++}\frac{1}{1+e^{\alpha_3+\alpha_5}}. \tag{7.19}$$

Using (7.11) and (7.19), (7.9) can be rewritten as

$$n_{0++}e^{\alpha_1}(1+e^{\alpha_2+\alpha_4})(1+e^{\alpha_3+\alpha_5})(N_2+e^{\alpha_3+\alpha_5}N_1) = n_{1++}(1+e^{\alpha_3})A,$$

where

$$\begin{aligned} A &= [(1+e^{\alpha_3+\alpha_5})N_1N_2 - e^{\alpha_1}(1+e^{\alpha_2+\alpha_4})(N_2+e^{\alpha_3+2\alpha_5}N_1)] \\ &= (1+e^{\alpha_2})(N_2+e^{\alpha_3+\alpha_5}N_1). \end{aligned}$$

This leads to

$$n_{1++}(1+e^{\alpha_3})(1+e^{\alpha_2}) = n_{0++}e^{\alpha_1}(1+e^{\alpha_2+\alpha_4})(1+e^{\alpha_3+\alpha_5}). \tag{7.20}$$

Based on (7.20), we find that

$$\frac{n_{0++}}{1+e^{\alpha_3}} = n_{+++}\frac{1+e^{\alpha_2}}{N_3} \tag{7.21}$$

$$\frac{n_{1++}}{1+e^{\alpha_3+\alpha_5}} = n_{+++}\frac{e^{\alpha_1}(1+e^{\alpha_2+\alpha_4})}{N_3}. \tag{7.22}$$

Plugging (7.21) and (7.22) into (7.11), we obtain (7.16). Since $N_3 = N_1 + e^{\alpha_3} N_2$, we find, using (7.16), that

$$\frac{n_{++0}}{N_1} = \frac{n_{+++}}{N_3}, \tag{7.23}$$

$$\frac{n_{++1}}{N_2} = \frac{n_{+++} e^{\alpha_3}}{N_3}. \tag{7.24}$$

Based on these two equations, (7.14), (7.15) and (7.17) follow immediately from (7.9), (7.10) and (7.12) respectively. Finally, (7.18) follows from (7.13), (7.22) and (7.24).

This implies that the pseudo-likelihood estimator satisfies the likelihood equations and hence coincides with the maximum likelihood estimator. □

The pseudo-likelihood model we used is derived from the joint distribution. This case will turn out to be convenient when the joint distribution is complex to evaluate (e.g., due to a complicated normalizing constant), but the conditionals are not. By using pseudo-likelihood methods instead of likelihood methods, the joint distribution is avoided or restricted to a single evaluation (e.g., to calculate the joint probabilities using the estimated parameters). In addition, when the conditionals $f(\boldsymbol{y}|d)$ and $f(d|\boldsymbol{y})$ are of interest, the pseudo-likelihood is well-motivated. Another advantage of the pseudo-likelihood theory can be found in the next section, where we consider the case when conditionals are constructed directly, without necessarily starting from a joint distribution.

7.3 No Underlying Joint Density

Suppose we do not want to address the model for $f(y_1, y_2, d)$ directly. Under the MAR/ACMV assumption, the pseudo-likelihood reduces to $f(d|y_1)f(y_1|d)f(y_2|y_1)$. Thus, we have to define models for each of these three components. Since we work with binary data, a natural choice is to assume logistic models. Furthermore, these

models yield an easy interpretation for their parameters. Thus, suppose:

$$f(d|y_1) = \frac{\exp[(\alpha_3 + \alpha_5 y_1)d]}{1 + \exp(\alpha_3 + \alpha_5 y_1)}, \tag{7.25}$$

$$f(y_1|d) = \frac{\exp[(\alpha_1 + \alpha_6 d)y_1]}{1 + \exp(\alpha_1 + \alpha_6 d)}, \tag{7.26}$$

$$f(y_2|y_1) = \frac{\exp[(\alpha_2 + \alpha_4 y_1)y_2]}{1 + \exp(\alpha_2 + \alpha_4 y_1)}. \tag{7.27}$$

Then, the compatibility constraint

$$\frac{f(d|y_1, y_2)}{f(y_1, y_2|d)} = \frac{f(d|y_1)}{f(y_1|d)f(y_2|y_1)} = u(y_1, y_2)v(d)$$

implies $\alpha_5 = \alpha_6$. The Hammersley-Clifford constraints for existence of a valid joint likelihood (Besag 1974), or the compatibility constraints as defined by Arnold, Castillo and Sarabia (1992), are of no use here since they assume three full conditional densities. For our pseudo-likelihood, only the components $f(d|y_1)$ and $f(y_2|y_1)$ are full conditionals, whereas $f(y_1|d)$ is marginalized over y_2. We use the full conditionals $f(d|y_1, y_2)$ and $f(y_1, y_2|d)$, leading to our constraints.

After specifying these conditional models, $f(d)$ and $f(y_1, y_2)$ can be deduced from the compatibility constraints and they can be used to fit a pattern-mixture or a selection model respectively:

$$f(d) = \frac{\exp(\alpha_3 d)[1 + \exp(\alpha_1 + \alpha_5 d)]}{1 + \exp(\alpha_1) + \exp(\alpha_3) + \exp(\alpha_1 + \alpha_3 + \alpha_5)},$$

$$f(y_1, y_2) = \frac{\exp(\alpha_1 y_1)[1 + \exp(\alpha_3 + \alpha_5 y_1)]}{1 + \exp(\alpha_1) + \exp(\alpha_3) + \exp(\alpha_1 + \alpha_3 + \alpha_5)} \cdot \frac{\exp[(\alpha_2 + \alpha_4 y_1)y_2]}{1 + \exp(\alpha_2 + \alpha_4 y_1)}.$$

However, these marginal models will often have complicated expressions, suggesting that pseudo-likelihood will be the preferred technique.

The results from fitting Model (7.25)–(7.27) under the constraint $\alpha_5 = \alpha_6$ to the data from Table 7.1 are displayed in Table 7.3. The results are the same as in Table 7.2, except for α_1, which has a different interpretation. See also Table 7.4.

As one can observe from this table, α_4 and α_5 have exactly the same interpretation in both models. This does not hold for the other parameters, although, under the MAR/ACMV assumption, α_2 and α_3 are equal. The only difference is α_1, which

Table 7.3: Parameter Estimates (Standard Errors) for the Artificial Data from Table 7.1, based on a Pseudo-Likelihood containing Three Logistic Models

	(a)	(b)	(c)	(d)
α_1	0.0000(0.2828)	-0.6931(0.2449)	-1.0986(0.2309)	0.4055(0.1667)
α_2	-0.6931(0.2449)	-0.6931(0.2449)	-0.6931(0.2449)	-2.0794(0.3354)
α_3	1.0986(0.2309)	0.4055(0.1826)	0.0000(0.1633)	0.4055(0.1667)
α_4	1.3863(0.3464)	1.3863(0.3464)	1.3863(0.3464)	1.3863(0.4430)
$\alpha_5 \equiv \alpha_6$	0.0000(0.3266)	0.6931(0.2944)	1.0986(0.2828)	-0.8109(0.2357)

Table 7.4: Interpretation of the Parameters (logits of the means of the random variables for α_1, α_2, and α_3, and log odds ratios for α_4 and α_5)

	Trivariate Loglinear Model	Three Logistic Models
α_1	$Y_1\|(y_2, d) = (0, 0)$	$Y_1\|d = 0$
α_2	$Y_2\|(y_1, d) = (0, 0)$	$Y_2\|y_1 = 0$
α_3	$D\|(y_1, y_2) = (0, 0)$	$D\|y_1 = 0$
α_4	Y_1Y_2	Y_1Y_2
α_5	Y_1D	Y_1D

in the loglinear model represents the conditional logit of Y_1, given Y_2 and D are zero, and for the three logistic models gives the conditional effect of Y_1, given that only D is zero.

At first sight, it seems somewhat strange that two different models yield such similar results. Let us investigate the similarity between both approaches more closely. We can rewrite the loglinear model from Section 7.2 as Equations (7.25), (7.27) and

$$f(y_1|d) = \frac{\exp[(\alpha_1 + \alpha_5 d)y_1][1 + \exp(\alpha_2 + \alpha_4 y_1)]}{1 + \exp(\alpha_1 + \alpha_5 d) + \exp(\alpha_2) + \exp(\alpha_1 + \alpha_2 + \alpha_4 + \alpha_5 d)}. \tag{7.28}$$

Based on (7.25) and (7.27), $\alpha_2, \alpha_3, \alpha_4$ and α_5 can be estimated. Models (7.28) and

(7.26) are only needed to estimate α_1. This last component is different for both models, leading to a different estimate for α_1. It has to be noted that, although α_2 seems to be equal in both settings, in fact its interpretation is different. In the loglinear model, α_2 describes the logit of Y_2 given $Y_1 = 0$ and $D = 1$, whereas for the logistic models, α_2 captures the logit of Y_2 given $Y_1 = 0$. But due to the MAR/ACMV assumption, both values are equal. A similar argument holds for α_3.

In fact, due to the compatibility constraints, we did not have to reduce the model to the missing data mechanisms the way we did. If one includes the missing data mechanism in only one component, $f(y_1, y_2|d)$ say, then the compatibility constraints reduce the other component, $f(d|y_1, y_2)$, to the compatible missing data mechanism.

As suggested by Louise Ryan (private communication), ignoring the compatibility constraints when fitting the model can be used as a sensitivity analysis. Indeed, by formally considering the difference between parameters that should be equal, the sensitivity of the model to the assumptions made can be assessed, at least in part.

In the case of Model (7.25)–(7.27), this procedure would ignore the constraint $\alpha_5 \equiv \alpha_6$. Parameter estimates and standard errors found by fitting the model to the data of Table 7.1, are exactly the same as those displayed in Table 7.2. In particular, $\hat{\alpha}_5 = \hat{\alpha}_6$, and is also equal to the value found in Table 7.2. This implies, once again, that the constraint was correctly chosen. Furthermore, the estimates of precision are exactly the same, indicating that both models use exactly the same information to estimate the (Y_1, D) interaction. Indeed, the fact that the information about this interaction is used twice in Table 7.2, is properly corrected by means of the robust variance estimator.

7.4 Fluvoxamine Data

Until now, we have considered only artificial data consisting of a single contingency table with supplemental margins. We now turn our attention to a real set of data with a repeated categorical outcome and individual level covariates. We will use the data introduced in Section 2.1, only looking at times 2 and 5, where the outcome is dichotomized. This comes down to the side effects from Table 2.4.

We included the covariates as main effects in the three logistic models (7.25), (7.26) and (7.27):

$$\begin{aligned} f(d|y_1) &= \frac{\exp[(\alpha_3 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{antecedents} + \alpha_5 y_1)d]}{1 + \exp(\alpha_3 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{antecedents} + \alpha_5 y_1)}, \\ f(y_1|d) &= \frac{\exp[(\alpha_1 + \gamma_1 \text{age} + \gamma_2 \text{sex} + \gamma_3 \text{antecedents} + \alpha_6 d)y_1]}{1 + \exp(\alpha_1 + \gamma_1 \text{age} + \gamma_2 \text{sex} + \gamma_3 \text{antecedents} + \alpha_6 d)}, \\ f(y_2|y_1) &= \frac{\exp[(\alpha_2 + \delta_1 \text{age} + \delta_2 \text{sex} + \delta_3 \text{antecedents} + \alpha_4 y_1)y_2]}{1 + \exp(\alpha_2 + \delta_1 \text{age} + \delta_2 \text{sex} + \delta_3 \text{antecedents} + \alpha_4 y_1)}. \end{aligned}$$

Parameter estimates and standard errors are displayed in Table 7.5. The fit of a more parsimonious model, found with a backward selection procedure, is presented in Table 7.6.

Since a positive *sex* effect remains in the model for $f(y_2|y_1)$, this means that the probability of having side effects at the second time point is higher for males than for females, both for those with and without side effects at the first time point. In particular, for someone without side effects at the first time point, the probability of side effects at the second time is 0.268 for females, and 0.441 for males. For someone with side effects at the first time, these probabilities are increased to 0.747 and 0.864 respectively. Furthermore, side effects at the first time increase the dropout probability (0.398 versus 0.147), thereby ruling out completely random dropout. These results are in agreement with the results found in Chapter 6.

7.5 Conclusion

While full likelihood is commonly used to analyse incomplete data, a choice is typically made between a selection or a pattern-mixture model. This forces one to choose between $f(\boldsymbol{y}|d)$ and $f(d|\boldsymbol{y})$. In many cases, both relationships will be of interest. If a choice is not preferable, an alternative procedure is pseudo-likelihood. A loglinear model, such as (7.6), can be rewritten in a selection model or pattern-mixture model way, but although the calculation of $f(\boldsymbol{y}|d)$ or $f(d|\boldsymbol{y})$ is straightforward, $f(d)$ and $f(\boldsymbol{y})$ have no easy form. Thus, using pseudo-likelihood theory has not only the advantage of easy forms for the models for $f(\boldsymbol{y}|d)$ and $f(d|\boldsymbol{y})$, but also no choice is implied.

Table 7.5: Fluvoxamine Data, Side Effects (full)

Parameter	Estimate (s.e.)
α_1	-0.048(0.507)
α_2	-1.737(0.602)
α_3	1.808(0.543)
α_4	2.031(0.361)
$\alpha_5 \equiv \alpha_6$	-1.307(0.291)
β_1	-0.165(0.349)
β_2	0.507(0.289)
β_3	-0.224(0.305)
γ_1	0.553(0.297)
γ_2	-0.156(0.261)
γ_3	0.014(0.263)
δ_1	0.652(0.350)
δ_2	0.784(0.333)
δ_3	-0.262(0.324)

Table 7.6: Fluvoxamine Data, Side Effects (reduced)

Parameter	Estimate (s.e.)
α_1	0.673(0.246)
α_2	-1.004(0.294)
α_3	1.758(0.217)
α_4	2.085(0.361)
$\alpha_5 \equiv \alpha_6$	-1.346(0.284)
δ_2	0.767(0.329)

Another major advantage of pseudo-likelihood is that, rather than deriving both full conditionals from a joint model, we have shown that the conditional densities can be specified directly, leaving the investigator the choice of models with a more convenient form. But since not all choices for both conditional densities are compatible, constraints are needed. To this end, we have generalized the compatibility constraints from Arnold, Castillo and Sarabia (1992). This procedure has the advantage of enabling the specification of simple models for all conditional densities involved. We have established a situation where the joint density compatible with the conditionals specified is different from but similar to the trivariate loglinear model. In particular, most of the parameters retain their interpretation.

We have applied this method to the fluvoxamine data (see Chapter 2). The conclusions from the pseudo-likelihood analysis are similar to the ones obtained from previous likelihood based analyses.

Chapter 8

Sensitivity Analysis for a Longitudinal Quality of Life Measure in a Cancer Trial

In this chapter, we advocate the use of pattern-mixture models as a tool to assess sensitivity of a selection model to the modelling assumptions, or vice versa, for continuous data (Michiels *et al.* 1998). It complements Chapter 6, where a sensitivity analysis based on selection models and pattern-mixture models for categorical data is performed. Explicitly, it will be argued that extra confidence in the conclusion can be gained if two analyses, one within each framework, coincide in key aspects, such as covariate dependencies, strength of association between outcomes, etc. We will outline ways to fit both selection and pattern-mixture models, based on linear mixed models for the measurement process. Virtually all models will be fitted using standard statistical software.

In the first section, the mixed model is introduced. The data used throughout this chapter come from the Vorozole Study, described in Section 2.2. A first exploratory analysis is given in Section 8.2. Then the model is analysed using a selection model (Section 8.3), and a pattern-mixture model (Section 8.4).

8.1 A Repeated-Measures Model

A model for repeated measurements, incorporating random effects (Laird and Ware 1982) and serial correlation (Diggle 1988) can be written as

$$\boldsymbol{Y}_s = Z_s^\beta \boldsymbol{\beta} + Z_s \boldsymbol{b}_s + \boldsymbol{W}_s + \boldsymbol{\varepsilon}_s \tag{8.1}$$

where $\boldsymbol{Y}_s$ is the n_s dimensional response vector for subject s with components Y_{sj}, $1 \leq s \leq N$, N is the number of subjects, Z_s^β and Z_s are $(n_s \times p)$ and $(n_s \times q)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is the p dimensional vector containing the fixed effects, $\boldsymbol{b}_s \sim N(\boldsymbol{0}, D)$ is the q dimensional vector containing the random effects, $\boldsymbol{W}_s \sim N(\boldsymbol{0}, \tau^2 H_s)$ is a vector of n_s realizations of a Gaussian stochastic process and $\boldsymbol{\varepsilon}_s \sim N(\boldsymbol{0}, \sigma^2 I_{n_s})$ is a n_s dimensional vector of uncorrelated error terms. Further, D is a general $(q \times q)$ covariance matrix, τ^2 is the serial variance, H_s is the serial correlation matrix, usually modelled in terms of one or a few parameters, and σ^2 is the measurement error variance. Often, the serial and measurement error processes are combined to yield a single residual variance matrix $\Sigma_s = \tau^2 H_s + \sigma^2 I_{n_s}$.

Two popular choices to capture serial correlation is by means of exponential or Gaussian decay. An exponential process is based on writing the correlation between two residuals at times t_{sj} and t_{sk} as

$$\text{Corr}(t_{sj}, t_{sk}) = \exp\left(-\frac{|t_{sj} - t_{sk}|}{\phi}\right) = \rho^{|t_{sj} - t_{sk}|}, \tag{8.2}$$

for some value of $\phi > 0$, where $\rho = \exp(-1/\phi)$. The Gaussian counterpart is

$$\text{Corr}(t_{sj}, t_{sk}) = \exp\left(-\frac{(t_{sj} - t_{sk})^2}{\phi^2}\right) = \rho^{(t_{sj} - t_{sk})^2}, \tag{8.3}$$

for some value of $\phi > 0$, where $\rho = \exp(-1/\phi^2)$.

It follows from (8.1) that, conditional on the random effect $\boldsymbol{b}_s$, $\boldsymbol{Y}_s$ is normally distributed with mean vector $Z_s^\beta \boldsymbol{\beta} + Z_s \boldsymbol{b}_s$ and with covariance matrix Σ_s. Define $V_s = Z_s D Z_s' + \Sigma_s$. Then the marginal distribution of $\boldsymbol{Y}_s$ is

$$\boldsymbol{Y}_s \sim N(Z_s^\beta \boldsymbol{\beta}, V_s). \tag{8.4}$$

The most popular approaches to parameter estimation are maximum likelihood and restricted maximum likelihood (Verbeke and Molenberghs 1997, Section 3.4).

8.2 Exploratory Analysis

Most books on longitudinal data discuss exploratory analysis. See, for example, Diggle, Liang and Zeger (1994). However, most effort is spent on model building and formal aspects of inference. In this section, we present a selected set of plots to underpin the model building. We distinguish between two modes of display: (1) plots averaged over (sub)populations and (2) individual profile plots. Both ways are used to present three fundamental aspects of the longitudinal structure: (1) the average evolution; (2) the variance function, (3) the correlation structure. Each of those will be discussed in turn. In addition, the variogram will be discussed. The data come from the Vorozole Study, introduced in Section 2.2. We will, apart from treatment, correct for dominant site of the disease as well as the baseline value for each patient. Most missing values in this study are due to dropout. The few intermittent missing values are treated as MCAR. The dropout process itself will be explored in further sections.

8.2.1 The Average Evolution

The average evolution describes how the profile for a number of relevant subpopulations (or the population as a whole), evolves over time. The results of this exploration will be useful in order to choose a fixed-effects structure for the linear mixed model.

The individual profiles are displayed in Figure 8.1, while the mean profiles per treatment arm, as well as their 95% confidence intervals, are plotted in Figure 8.2. The average profiles indicate an increase over time which is slightly stronger for the vorozole group until month 14, and afterwards, the megestrol acetate group obtains a slightly higher FLIC-score. As can be seen from the confidence intervals, these differences do not seem to be significant.

The individual profiles augment the averaged plot with a suggestion of the variability seen within the data. The thinning of the data towards the later study times suggests that trends at later times should be treated with caution. While these plots also give us some indications about the variability at given times and even about

Figure 8.1: Vorozole Study: Individual Profiles for Change (Raw, Detrended, and Standardized)

the correlation between measurements of the same individual, it is easier to base such considerations on residual profiles and standardized residual profiles.

8.2.2 The Variance Structure

In addition to the average evolution, the evolution of the variance is important to build an appropriate longitudinal model. Clearly, one has to correct the measurements for the fixed-effects structure and hence detrended residuals have to be used. These detrended residuals are merely the outcome values (change in FLIC-score), from which the mean change is subtracted, calculated at each time point separately. Again, two plots are of interest. The first one pictures the average evolution of the variance as function of time, the second one merely produces the individual residual plots. The detrended profiles are displayed in Figure 8.1, while the corresponding variance function is plotted in Figure 8.3.

Figure 8.2: Vorozole Study: Mean Profiles and 95% Confidence Intervals

The variance function seems to be relatively stable, except for a sharp decline near the end, due to the large amount of dropout, and hence a constant variance model could be a plausible starting point. The individual detrended profiles show subjects' tendency, most clearly in the vorozole group, to decrease right before they leave the study. In addition, also the detrended profiles suggest that the variance would decrease over time.

8.2.3 The Correlation Structure

The correlation structure describes how measurements within a subject correlate. The correlation function depends on a pair of times and only under the assumption of stationarity does this pair of times simplify to the time lag only. This is important since many exploratory and modelling tools are based on this assumption. A plot of standardized residuals is useful in this respect (see Figure 8.1). The picture is not radically different from the previous individual plots, which can be explained

Figure 8.3: Vorozole Study: Variance Function

by the relative flatness of both mean profile and variance functions. If one or both structures is varying with time, the standardized residuals will contribute useful additional information.

A different way of displaying the variance structure is using a scatterplot matrix, such as in Figure 8.4. The off-diagonal elements picture scatterplots of standardized residuals obtained from pairs of measurement occasions. The decay of correlation with time is studied by considering the evolution of the scatters with increasing distance to the main diagonal. Stationarity on the other hand implies that the scatterplots remain similar within diagonal bands if measurement occasions are approximately equally spaced. In addition to the scatterplots, we place histograms on the diagonal, capturing the variance structure including such features as skewness. Since the axes are given the same scales, it is very easy to capture the attrition rate as well.

Figure 8.4: Vorozole Study: Scatterplot Matrix

8.2.4 The Variogram

Model (8.1) distinguishes between three components of variability. The first one groups traditional random effects (as in a random-effects ANOVA model) and random coefficients (Longford 1993). It stems from inter-individual variability, i.e., heterogeneity between individual profiles. The second component, serial association, is present when residuals close to each other in time are more similar than residuals further apart. This notion is well-known from the time-series literature (Ripley 1981, Diggle 1983, Cressie 1991). Finally, on top of the other two components, there is potentially also measurement error. This results from the fact that for delicate measurements (e.g., laboratory assays), even immediate replication will not be able to avoid considerable variation. In longitudinal data, these three components of variability can be distinguished by virtue of both *replication* as well as a clear *distance* concept (time).

Diggle (1990) and Diggle, Liang and Zeger (1994) promote the so-called semi-

variogram to picture the variance components. It is easily estimated even with irregular observation times (but might require some amount of smoothing). Given a stationary mean-zero stochastic process $Y(t)$ with constant variance, the variogram is defined as

$$V(u) = \frac{1}{2}E\left\{[Y(t) - Y(t-u)]^2\right\}.$$

Specializing (8.1) to random intercept only, D simplifies to a scalar, δ^2 say, and it is easy to show (Diggle 1990) that the variogram equals

$$V(u) = \sigma^2 + \tau^2(1 - \rho(u)),$$

where $u = t_{sj} - t_{sk}$ is the time lag between both measurements and $\rho(u)$ is the serial correlation between two measurements with the specified lag, calculated for example from (8.2) or (8.3). Note that $V(0) = \sigma^2$ and $V(\infty) = \sigma^2 + \tau^2$. Plotting the process variance,

$$\text{Var}(Y_{sj}) = \delta^2 + \sigma^2 + \tau^2,$$

as a horizontal line and the variogram as a curve, the three components of variability are easy to retrieve. The measurement error is $V(0)$, the random intercept variance is the difference between the process variance and $V(\infty)$, and the variance of the serial process is seen as the band, occupied by the variogram, which increases from $V(0)$ to $V(\infty)$. With irregularly spaced data, it is usually necessary to smooth the variogram. The shape of the variogram conveys information about the structure of the serial correlation function.

The variogram for this study is given in Figure 8.5, where the three components of variability are seen to be roughly of the same magnitude.

8.3 A Selection Model Formulation

First, a linear mixed model for the measurements of the form (8.1) is assumed.

Secondly, we will model the dropout mechanism. We assume that incompleteness is due to dropout only, and that the first measurement Y_{s1} is obtained for everyone. The model for the dropout process is based on a logistic regression for the probability of dropout at occasion j, given the subject is still in the study. We denote this

Figure 8.5: Vorozole Study: Variogram

probability by $g(h_{sj}, y_{sj})$, in which h_{sj} is a vector containing all responses observed up to but not including occasion j, as well as relevant covariates X^{α}_{sk}. We then assume that $g(h_{sj}, y_{sj})$ satisfies

$$\text{logit}[g(h_{sj}, y_{sj})] = \text{logit}\left[P(D_s = j | D_s \geq j, \boldsymbol{y}_s, \boldsymbol{X}^{\alpha}_s)\right] = h_{sj}\boldsymbol{\alpha}_0 + y_{sj}\alpha_d \quad s = 1, \ldots, N, \tag{8.5}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \alpha_d)'$. When α_d equals zero, the dropout model is random, and all parameters can be estimated using standard software since the measurement model for which we use a linear mixed model and the dropout model, assumed to follow a logistic regression, can then be fitted separately. If $\alpha_d \neq 0$, the dropout process is assumed to be non-random.

Model (8.5) is now used to construct the dropout process:

$$f(d_s|\boldsymbol{y}_s, \boldsymbol{X}_s^{\alpha}, \boldsymbol{\alpha}) = \begin{cases} \prod_{j=2}^{n_s}[1 - g(h_{sj}, y_{sj})] & \text{for } d_s = n_s + 1, \\ \prod_{j=2}^{d-1}[1 - g(h_{sj}, y_{sj})]g(h_{sd}, y_{sd}) & \text{for } d_s = d \leq n_s. \end{cases} \tag{8.6}$$

Several authors point to the sensitivity of this model to assumptions about the dropout process which are fundamentally not verifiable. See the discussion to Diggle and Kenward (1994) and Verbeke *et al.* (1998).

Application to the Vorozole Study

Since we are modelling change versus baseline, all models are forced to pass through the origin. This is done by omitting the main covariate effects, and looking only at interactions of these covariates with time. The following covariates were considered for the measurement model: baseline value, treatment, dominant site, and time in months. Second order interactions were considered as well. For design reasons, treatment was kept in the model in spite of its non-significance. An F test for treatment effect produces a p value of 0.5822. Apart from baseline, no other time-stationary covariates were kept. A quadratic time effect provided an adequate description of the time trend. Based on the variogram, we confined the random-effects structure to random intercepts, and supplemented this with a spatial Gaussian process and measurement error. The final model is presented in Table 8.1. The fitted variance structure is represented by means of the fitted variogram in Figure 8.5. The total correlation between two measurements, one month apart, equals 0.696. The residual correlation, which remains after accounting for the random effects, is still equal to 0.491. The serial correlation, obtained by further ignoring the measurement error, equals $\rho = \exp(-1/7.22^2) = 0.981$.

The fitted profiles are displayed in Figure 8.6 and Figure 8.7. For the latter, the random effects were taken into account when calculating the predicted values, for the first they were not. For each treatment group, we obtain three sets of profiles. The fitted complete profile is the average curve that would be obtained,

Table 8.1: Vorozole Study: Selection Model

Effect	Estimate (s.e.)
Fixed-Effect Parameters:	
time	7.78 (1.05)
time∗baseline	-0.065 (0.009)
time∗treatment	0.086 (0.157)
time^2	-0.30 (0.06)
time^2∗baseline	0.0024 (0.0005)
Variance Parameters:	
random intercept (δ^2)	105.42
serial variance (τ^2)	77.96
serial association (ϕ)	7.22
measurement error (σ^2)	77.83

had all individuals been completely observed. If we use only those predicted values that correspond to occasions at which an observation was made, then the fitted incomplete profiles are obtained. The latter are somewhat above the former when the random effects are included, and somewhat below when they are not, suggesting that individuals with lower measurements are more likely to disappear from the study. In addition, while the fitted complete curves are very close (the treatment effect was not significant), the fitted incomplete curves are not, suggesting that there is more dropout in the standard arm than in the treatment arm. This is in agreement with the dropout rate, displayed in Figure 8.8. Finally, the observed curves, based on the measurements available at each time point, are displayed. These lie highest, but this should be viewed with the standard errors of the observed means in mind, which lie around 18.5 (see also Figure 8.2).

Next, we will study factors which influence dropout. A logistic regression model, described by (8.5) and (8.6) is used. To start, we restrict attention to MAR processes, whence $\alpha_d = 0$. The first model includes treatment, dominant site, baseline,

Figure 8.6: Vorozole Study: Fitted Profiles (averaging the predicted means for the incomplete and complete measurement sequences, without random effects)

and the previous measurement but only the last two are significant, producing

$$\text{logit}[g(h_{sj})] = 0.080(0.341) - 0.014(0.003)\text{base}_s - 0.033(0.004)y_{s,j-1}\,. \tag{8.7}$$

Diggle and Kenward (1994) and Molenberghs, Kenward and Lesaffre (1997) considered non-random versions of this model by including the current, possibly unobserved measurement, such as in (8.5). This requires more elaborate fitting algorithms, since the missing data process is then non-ignorable, and hence the full density needs to be used. Diggle and Kenward used the simplex algorithm (Nelder and Mead 1965), while Molenberghs, Kenward and Lesaffre fitted their models with the EM algorithm (Dempster, Laird and Rubin 1977). The algorithm of Diggle and Kenward is implemented in Oswald (Smith, Robertson and Diggle 1996). With larger datasets such as this one, convergence can be painstakingly difficult and one has to worry about apparant convergence. Therefore, we first proceed in an al-

Figure 8.7: Vorozole Study: Fitted Profiles (averaging the predicted means for the incomplete and complete measurement sequences, including the random effects)

ternative way. Both Diggle and Kenward (1994) and Molenberghs, Kenward and Lesaffre (1997) observed that in informative models, dropout tends to depend on the increment, i.e., the difference between the current and previous measurements $y_{sj} - y_{s,j-1}$. Clearly, a very similar quantity is obtained as $y_{s,j-1} - y_{s,j-2}$, but a major advantage of such a model is that it fits within the MAR framework. In our case, we obtain

$$\begin{aligned} \text{logit}[g(h_{sj})] &= 0.033(0.401) - 0.013(0.003)\text{base}_s + 0.012(0.006)y_{s,j-2} \\ &\quad -0.035(0.005)y_{s,j-1} \\ &= 0.033(0.401) - 0.013(0.003)\text{base}_s - 0.023(0.005)\frac{y_{s,j-2} + y_{s,j-1}}{2} \\ &\quad -0.047(0.010)\frac{y_{s,j-1} - y_{s,j-2}}{2} \end{aligned} \tag{8.8}$$

indicating that both size and increment are significant predictors for dropout. We conclude that dropout increases with a decrease in baseline, in overall level of the

Figure 8.8: Vorozole Study: Observed Dropout per Treatment Arm

outcome variable, as well as with a decreasing evolution in the outcome.

Using Oswald (Smith, Robertson and Diggle 1996), both dropout models (8.7) and (8.8) can be compared with their non-random counterparts, where y_{sj} is added to the linear predictor. The first one becomes

$$\text{logit}[g(h_{sj}, y_{sj})] = 0.53 - 0.015\text{base}_s - 0.076y_{s,j-1} + 0.057y_{sj} \tag{8.9}$$

while the second one becomes

$$\text{logit}[g(h_{sj}, y_{sj})] = 1.38 - 0.021\text{base}_s - 0.0027y_{s,j-2} - 0.064y_{s,j-1} + 0.035y_{sj}. \tag{8.10}$$

Formal testing of dropout models (8.9) versus (8.7) and (8.10) versus (8.8) are possible in principle, but will not be carried out for two reasons. First, the likelihood function tends to be very flat for non-random dropout models and therefore the determination of the likelihood ratio is often computationally non-trivial. More fundamentally, Rubin (1994), Little (1994b), Laird (1994), and Molenberghs, Kenward and Lesaffre (1997) point out that formal testing for non-random dropout

faces philosophical objections. Indeed, non-random dropout models are identified only due to strong but unverifiable assumptions. Hogan and Laird (1997) suggest pattern-mixture models as a viable alternative.

8.4 A Pattern-Mixture Model Formulation

Recall from Section 3.1.2 that a pattern-mixture model is based on the following factorization:

$$f(\boldsymbol{y}_s, d_s|\boldsymbol{X}_s, \boldsymbol{\theta}) = f(\boldsymbol{y}_s|d_s, \boldsymbol{X}_s^{\beta}, \boldsymbol{\beta}) f(d_s|\boldsymbol{X}_s^{\alpha}, \boldsymbol{\alpha}). \tag{8.11}$$

The dropout process (8.6) thus simplifies to $f(d_s|\boldsymbol{X}_s^{\alpha}, \boldsymbol{\alpha})$ which is a, possibly covariate-corrected, model for the probability to belong to a particular pattern. Its components, $g(h_{sj})$, containing only covariates now, describe the dropout rate at each occasion.

The measurement model has to reflect dependence on dropout. In its most general form, this implies that (8.1) is replaced by

$$\begin{cases} \boldsymbol{Y}_s = Z_s^{\beta}\boldsymbol{\beta}(d_s) + Z_s\boldsymbol{b}_s + \boldsymbol{\varepsilon}_s \\ \\ \boldsymbol{b}_s \sim N(\boldsymbol{0}, D(d_s)), \\ \\ \boldsymbol{\varepsilon}_s \sim N(\boldsymbol{0}, \Sigma_s(d_s)). \end{cases} \tag{8.12}$$

Thus, the fixed effects as well as the covariance parameters are allowed to change with dropout pattern and a priori no restrictions are placed on the structure of this change.

Model (8.12) contains underidentified members since it describes the full set of measurements in pattern d_s, even though there are no measurements after occasion $d_s - 1$. Little (1993, 1994a) advocated the use of identifiying restrictions which works well in relatively simple settings. Molenberghs, Michiels, Kenward and Diggle (1998) proposed a particular set of restrictions for the monotone case which correspond to MAR. To avoid this problem, simplified (identified) models can be considered.

The advantage is that the number of parameters decreases, which is generally an issue with pattern-mixture models. Hogan and Laird (1997) noted that in order to estimate the large number of parameters in general models, one has to make the awkward requirement that each dropout pattern is sufficiently "filled", in other words one has to require large numbers of dropouts. Note however that simplified models, qualified as "assumption rich" by Sheiner, Beale and Dunne (1997), are also making untestable assumptions and therefore illustrate that even pattern-mixture models do not provide a free lunch. A main advantage however is that the need of assumptions and their implications are more obvious. For example, it is not possible to assume an unstructured time trend in incomplete patterns, except if one restricts attention to the time range from onset until dropout. In contrast, assuming a linear time trend allows estimation in all patterns containing at least two measurements.

In general, we distinguish between two types of simplifications to identify pattern-mixture models. First, trends can be restricted to functional forms supported by the information available within a pattern. The linear time trend discussed earlier is an example. Secondly, one can let the parameters vary across patterns in a parametric way. Thus, rather than estimating a separate time trend in each pattern, one could asssume that the time evolution is unstructured in each pattern, but parallel across patterns. The available data can be used to assess whether such simplifications are supported within the time ranges for which there is information. Using the so-obtained profiles past the time of dropout still requires extrapolation.

Application to Vorozole Study

In analogy with the exploration in the selection model context, it is natural to explore the data from a pattern-mixture point of view. To this end, plots per dropout pattern can be constructed. Figures 8.9 and 8.10 display the individual and averaged profiles per pattern.

Figure 8.10 clearly shows that pattern-specific profiles are of a quadratic nature with in most cases a sharp decline prior to dropout. Note that this is in line with the fitted dropout mechanism (8.8). Therefore, this feature needs to be reflected in the pattern-mixture model. In analogy with our selection model, the profiles are

Figure 8.9: Vorozole Study: Individual Profiles per Dropout Pattern

Figure 8.10: Vorozole Study: Mean Profiles per Dropout Pattern

forced to pass through the origin. This is done by allowing only time as main effect in the model, and adding interactions of other covariables with time.

The most complex pattern-mixture model we consider includes a different parameter vector for each of the observed patterns. We then proceed by backward selection in order to simplify the model. First, we found that the covariance structure is common to all patterns, encompassing random intercept, a serial exponential process, and measurement error.

For the fixed effects we proceeded as follows. A backward selection procedure, starting from a model that includes a main effect of time and time^2, as well as interactions of time with baseline value, treatment effect, dominant site and pattern, and the interaction of pattern with time^2. This procedure revealed main effects of time and time^2, as well as interactions of time with baseline value, treatment effect, and pattern, and the interaction of pattern with time^2. This reduced model can be found in Table 8.2. As was the case with the selection model in Table 8.1,

Figure 8.11: Vorozole Study: Fitted Selection and First Pattern-Mixture Model

treatment effect is non-significant. Indeed, a single degree of freedom F test yields a p value of 0.6868. Note that such a test is possible since treatment effect does not interact with pattern, in contrast to the model which we will describe later. The fitted profiles are displayed in Figure 8.11. We observe that the profiles for both arms are very similar. This is due to the fact that treatment effect is not significant but perhaps also because we did not allow a more complex treatment effect. For example, we might consider an interaction of treatment with the square of time and, more importantly, an treatment effect which is pattern-specific. Some evidence for such an interaction is seen in Figure 8.10.

Our second, expanded model, allowed for up to cubic time effects, the interaction of time with dropout pattern, dominant site, baseline value and treatment, as well as their two- and three-way interactions. After a backward selection procedure, the effects included are time and time^2, the two-way interaction of time and dropout pattern, as well as three factor interactions of time and dropout pattern

Table 8.2: Vorozole Study: First Pattern-Mixture Model

Fixed-Effect Parameters (Estimate (s.e.)):				
Pattern	Time	Time∗Base	Time^2	Time∗Group (0)
main	4.671 (0.844)	-0.031 (0.004)	-0.034 (0.029)	-0.067 (0.166)
3	-8.856 (2.739)			
4	-0.796 (2.958)		-1.918 (1.269)	
5	-1.959 (1.794)		-0.145 (0.365)	
6	1.600 (1.441)		-0.541 (0.197)	
7	0.292 (1.295)		-0.107 (0.133)	
8	1.366 (1.035)		-0.181 (0.080)	
9	1.430 (1.045)		-0.132 (0.071)	
10	1.176 (1.025)		-0.118 (0.061)	
11	0.735 (0.934)		-0.083 (0.049)	
12	0.797 (1.078)		-0.078 (0.055)	
13	0.274 (0.989)		-0.023 (0.046)	
14	0.544 (1.087)		-0.026 (0.049)	
15				
Variance Parameters:				
Random intercept (δ^2)	78.45			
Serial variance (τ^2)	95.38			
Serial association (ϕ)	8.85			
Measurement error (σ^2)	73.77			

Table 8.3: Vorozole Study: Second Pattern-Mixture Model

Fixed-Effect Parameters (Estimate (s.e.)):				
Pattern	Time	Time∗Base	$Time^2$	$Time^2$∗Base
main	5.468 (5.089)	-0.034 (0.040)	-0.271 (0.206)	0.002 (0.002)
3	7.616 (21.908)	-0.119 (0.175)		
4	44.097 (17.489)	-0.440 (0.148)	-18.632 (7.491)	0.1458 (0.0644)
5	22.471 (10.907)	-0.218 (0.089)	-5.871 (2.143)	0.0484 (0.0178)
6	10.578 (9.833)	-0.055 (0.079)	-1.429 (1.276)	0.0080 (0.0107)
7	14.691 (8.424)	-0.123 (0.069)	-1.571 (0.814)	0.0127 (0.0069)
8	7.527 (6.401)	-0.061 (0.052)	-0.827 (0.431)	0.0058 (0.0036)
9	-12.631 (7.367)	0.086 (0.058)	0.653 (0.454)	-0.0065 (0.0038)
10	14.827 (6.467)	-0.126 (0.053)	-0.697 (0.343)	0.0052 (0.0029)
11	5.667 (6.050)	-0.049 (0.049)	-0.315 (0.288)	0.0021 (0.0023)
12	12.418 (6.473)	-0.093 (0.051)	-0.273 (0.296)	0.0016 (0.0024)
13	1.934 (6.551)	-0.022 (0.053)	-0.049 (0.289)	0.0003 (0.0024)
14	6.303 (6.426)	-0.052 (0.050)	-0.182 (0.259)	0.0015 (0.0021)
15				
Pattern	Time∗Group (0)	Time∗Domsite (1)	Time∗Domsite (2)	Time∗Domsite (3)
main		-0.873 (1.073)	0.941 (0.845)	0.023 (0.576)
3	0.445 (5.095)	-5.822 (17.401)	-9.320 (9.429)	1.431 (9.878)
4	0.867 (1.552)	2.024 (3.847)	4.393 (2.690)	5.681 (2.642)
5	-1.312 (0.808)	2.937 (2.596)	0.940 (1.697)	1.414 (1.633)
6	-0.249 (0.686)	-1.378 (2.699)	-4.366 (2.367)	-3.237 (2.289)
7	-0.184 (0.678)	-0.547 (1.917)	-1.099 (1.456)	-1.015 (1.344)
8	0.527 (0.448)	1.302 (1.130)	-0.914 (0.811)	
9	0.782 (0.502)	3.881 (1.485)	1.733 (1.226)	4.548 (1.218)
10	-0.809 (0.464)	2.359 (1.241)	-0.436 (0.843)	
11	-0.080 (0.443)	1.138 (1.128)	-0.326 (0.753)	
12	0.331 (0.579)		-3.595 (0.996)	
13	-0.679 (0.492)	0.317 (1.152)	0.182 (0.825)	
14	0.433 (0.688)		-1.694 (0.972)	
15	-1.323 (0.706)			
Variance Parameters:				
Random intercept (δ^2)		98.93		
Serial variance (τ^2)		38.86		
Serial association (ϕ)		6.10		
Measurement error (σ^2)		73.65		

Figure 8.12: Vorozole Study: Fitted Selection and Second Pattern-Mixture Model

with (1) baseline, (2) group, and (3) dominant site. Finally, time^2 interacts with dropout pattern and with the interaction of baseline and dropout pattern. No cubic time effects were necessary, which is in agreement with the observed profiles in Figure 8.10. The parameter estimates of this model are displayed in Table 8.3. The model is graphically represented in Figure 8.12.

Because a pattern-specific parameter has been included, we have several options for the assessment of treatment. Since there are 13 patterns, one can test the global hypothesis, based on 13 degrees of freedom, of no treatment effect. We obtain $F = 1.25$, producing $p = 0.2403$, indicating that there is no overall treatment effect. Each of the treatment effects separately is at a non-significant level. Alternatively, the marginal effect of treatment can be calculated, which is the weighted average of the pattern-specific treatment effects, with weights given by the probability of occurrence of the various patterns. Its standard error is calculated using a straightforward application of the delta method. This effect is equal to $-0.286(0.288)$ producing a

p value of 0.3206, which is still non-significant.

In summary, we obtain a differential assessment of treatment effect. We obtained highly non-significant results from the selection model and from the first pattern-mixture models. The 13 degrees of freedom assessment in the second pattern-mixture models produced a smaller but still non-significant p value, and also the marginal assessment in the second pattern-mixture model yielded a non-significant p value.

8.5 Conclusion

In this chapter we have concentrated on total FLIC (i.e., change of the score versus baseline), a quality of life score measured in a multi-centric two arm study in postmenopausal women suffering from metastatic breast cancer. Since virtually all patients were followed up until disease progression or death, the amount of dropout is large. A very large group of patients drops out after only a few months.

While classically only selection models are fitted, pattern-mixture models can be seen as a viable alternative. The average profile in the selection model depends on the baseline value, as well as on time. The latter effect is mildly quadratic. There is no evidence for a treatment difference. However, it should be noted that the average profile found is the one that *would* have been observed, had no subjects dropped out, and under the additional assumption that the MAR assumption is correct. Fitting non-random dropout models, in the sense of Diggle and Kenward (1994) is possible, but computationally difficult for a fairly large trial like this one. A separate study of the dropout mechanism revealed that dropout increases with three elements: (1) an unfavourable baseline score, (2) an unfavourable value at the previous month, as well as (3) an unfavourable change in value from the penultimate to the last obtained value.

A pattern-mixture model is fitted by allowing at first a completely separate parameter vector for each observed dropout pattern, which is then simplified by using standard model selection procedures, by considering whether effects are common to all patterns. A first pattern-mixture model features a common treatment effect, of

which the assessment is then straightforward. A second model includes a separate treatment effect for each dropout pattern. This leads to two distinct test. The first one tests for equality of the whole treatment vector to be zero. The second one first calculates the marginal treatment effect from the vector of effects, by composing a weighted sum, where the weights are the multinomially estimated probabilities of the various patterns. In all cases, there is no treatment effect. However, a graphical display of the fitted profiles per pattern is enlightening, since it clearly confirms the trend detected in the selection models, that patients tend to drop out when their quality of life score is declining. Since this feature is usually coupled to an imminent progression or death, it should not come as a surprise. An important advantage of pattern-mixture models is that fitting them is more straightforward than non-random selection models. The additional calculations needed for the marginal treatment effect and its associated precision can be done straightforwardly using the delta method.

Chapter 9

Concluding Remarks and Further Research

Since in many longitudinal studies missing data appear, it is necessary to explore methods to take this missingness into account. Most theory established for this problem is based on selection models, where a marginal measurement model is combined with the missing data model, conditional on the measurements. Selection models fit naturally into the different missing data mechanisms defined by Little and Rubin (1987). They divide the assumptions into basically three groups: MCAR, where independence between the measurements and the missingness process is assumed, MAR, where missingness can depend on the observed data, and MNAR, where missingness can also depend on the unobserved measurements. A special case of MNAR is the protective estimator, where one assumes missingness to depend on the unobserved, but not on the observed measurements. Brown (1990) defined the protective estimator for normal data. In Chapter 4, we translated the protective estimator to a version for categorical data, adding methods to estimate the precision. In some cases, this protective assumption can be much more realistic than MAR, for example when the measurements are taken relatively far apart in time, and there is sufficient washout.

In contrast to selection models, another factorization of the full density is possible, leading to pattern-mixture models. Here, for all the different missing data patterns, different models need to be fitted. It is clear that, if one assumes similar

models for different patterns, not all these models are identifiable, due to the missingness. Therefore, identifying restrictions are needed (Little 1993, 1995). These restrictions are the counterpart of the modelling assumptions used in selection modelling. But if one uses pattern-mixture models, it is clear what information is missing, and what assumptions about the missingness process are made. We established the ACMV restrictions (Chapter 5), a counterpart for the MAR assumption in case of monotone dropout. Due to this assumption, we can, in the case of dropout, make the same subdivision for pattern-mixture models as for selection models. In case of non-monotone missingness, the ACMV restrictions are not equivalent to MAR, and further research is therefore needed to formulate plausible restrictions.

Now that the same assumption can be used in a selection model and a pattern-mixture model, different methods combining both factorizations can be explored. First, we created a pseudo-likelihood, combining the interesting parts of both models: the pattern-specific measurement models on the one hand, and the dropout model, conditioned on the outcomes on the other (Chapter 7). This method has the advantage that interesting models can be chosen for both parts (with mild conditions on the compatibility), leading to results that are easy to interpret, or that are of direct interest to the investigator.

A second reason for combining selection models and pattern-mixture models is sensitivity analysis. If a selection model analysis and a pattern-mixture model analysis lead to similar conclusions, more confidence can be given to these results. We have carried out such a sensitivity analysis for categorical data assuming a Dale model (Chapter 6) and for continuous data assuming mixed models (Chapter 8).

Pattern-mixture models are in fact much more honest than selection models, because one can easily see which information is missing. Furthermore can it be of interest to have the measurement model parameters for different missing data patterns, values that demand quite some calculation after a selection model analysis. Still most analyses containing missing data fix on selection models. Therefore, pattern-mixture models should be included more in a comprehensive study, leading to a more thorough sensitivity analysis. Finally, pattern-mixture models deserve to be studied as sensitivity tools in their own right.

Samenvatting

In veel studies komt men het probleem tegen dat sommige waarden ontbreken. Vooral in longitudinale studies, waar gegevens op regelmatige tijdstippen opgemeten worden, moet met dit probleem rekening gehouden worden. Een speciaal geval van ontbrekende waarden is dropout, waar men per subject in de studie een aantal opeenvolgende metingen heeft, waarna verdere metingen ontbreken. Little en Rubin (1987) gaven een opsplitsing in de veronderstellingen omtrent het ontbreken van waarden. Ze gingen hiervoor uit van een *selectiemodel*, waar de gezamelijke dichtheidsfunctie van zowel het meetproces als het proces dat aangeeft welke waarden ontbreken, wordt opgesplitst in het product van het meetproces met het proces van de ontbrekende waarden, geconditioneerd op de meetresultaten. Little en Rubin spitsten de assumpties op in drie groepen:

MCAR: Het ontbreken van gegevens heeft niets met de metingen te maken.

MAR: Het ontbreken van gegevens kan verklaard worden door de metingen die men geobserveerd heeft.

MNAR: Ook de niet geobserveerde metingen zijn nodig om het ontbreken van gegevens te verklaren.

Brown (1990) introduceerde de *protective estimator*, een speciaal geval van MNAR, waarbij het ontbreken van een gegeven op een bepaald tijdstip enkel afhangt van dit ontbrekende gegeven zelf. Deze aanname kan bij voorbeeld heel logisch zijn als de tijd tussen verschillende metingen heel groot is.

Brown heeft de protective estimator enkel gedefinieerd voor Normaal verdeelde gegevens. In Hoofdstuk 4 breiden we deze schatter verder uit naar categorische

gegevens. Het schatten van de kansen vraagt het oplossen van een stelsel, maar een vereenvoudigde methode is mogelijk, waarbij een aantal kleine stelsels moet opgelost worden. Een extra voordeel van deze methode is dat een contradictie in de voorwaarden tot uiting komt als negatieve kansen. Er worden ook drie methoden behandeld om de precisie van de schattingen te bepalen.

De assumpties van Little en Rubin zijn gebaseerd op selectiemodellen. Een andere factorisatie van de volledige dichtheid geeft aanleiding tot *pattern-mixture modellen*. Hier bekijkt men voor elk patroon van ontbrekende waarden een afzonderlijk meetmodel, en wordt het model van de ontbrekende waarden apart behandeld. Aangezien er gegevens ontbreken, is het duidelijk dat niet voor alle patronen het meetmodel identificeerbaar zal zijn. Daarom heeft men restricties nodig. Little (1993, 1995) geeft een overzicht van een aantal restricties. De MCAR-assumptie, die onafhankelijkheid veronderstelt tussen het meetproces en het proces van de ontbrekende waarden, is natuurlijk hetzelfde in een selectiemodel en een pattern-mixture model. Maar de andere assumpties kunnen niet eenvoudig vertaald worden naar restricties voor pattern-mixture modellen.

In Hoofdstuk 5 wordt de ACMV-restrictie gedefinieerd. Volgens deze restrictie moet men, om een dichtheid op een bepaald tijdstip, gegeven de vorige metingen, te berekenen voor een patroon waar niet alle informatie beschikbaar is, deze informatie gaan lenen bij alle patronen waar deze dichtheid wel kan berekend worden. In het geval van dropout hebben we in dit hoofdstuk bewezen dat ACMV en MAR equivalent zijn. Dit geeft de mogelijkheid om voor pattern-mixture modellen eenzelfde soort opsplitsing te maken als voor selectiemodellen. Daardoor kan men een selectiemodel en een pattern-mixture model fitten aan data, waarbij voor beide modellen dezelfde veronderstelling gemaakt wordt over het ontbreken van metingen. Zo kan men een sensitiviteitsanalyse uitvoeren. Als de twee modellen vergelijkbare resultaten geven, bij voorbeeld wat betreft het effect van de behandeling, kan men veel meer vertrouwen hebben in deze resultaten. Een ander voordeel is dat de resultaten van het selectiemodel soms van klinisch belang zijn, maar soms is men juist geïnteresseerd in de resultaten van het pattern-mixture model: is het effect van de behandeling verschillend voor patiënten die op een ander moment de studie ver-

laten? Door de equivalentie van beide modellen, kan dàt model gekozen worden dat het beste aansluit bij de klinische vraag.

In Hoofdstuk 6 hebben we zo het therapeutisch effect en de nevenwerkingen bestudeerd van het toedienen van fluvoxamine aan psychiatrische patiënten. Gebruik makend van een bivariaat Dale model hebben we de kans op nevenwerkingen en de kans op therapeutisch effect geschat, met als covariaten de leeftijd, het geslacht en de psychiatrische voorgeschiedenis van de patiënten. Beide modellen gaven vergelijkbare resultaten: voor de kans op nevenwerkingen was er stijging met de leeftijd, en deze kans was hoger voor patiënten die uitvielen na de eerste meting, dan voor patiënten die ook op de tweede meting aanwezig waren. Er was ook een sterke associatie aanwezig tussen de verschillende metingen. De psychiatrische voorgeschiedenis had helemaal geen invloed, en het geslacht had een klein effect op de kans op nevenwerkingen. Voor het therapeutisch effect was er helemaal geen invloed van leeftijd, geslacht, of psychiatrische voorgeschiedenis. Ook hier was er wel een sterke associatie aanwezig. Al deze conclusies kunnen zowel uit het selectiemodel als uit het pattern-mixture model getrokken worden.

Een andere dataset, waar patiënten met borstkanker behandeld worden met Vorozole of met Megestrol Acetate, wordt behandeld in Hoofdstuk 8. Voor deze patiënten werd gekeken naar de FLIC-score, een maat voor de levenskwaliteit tijdens de behandeling. Gebruik makend van mixed models, gaven het selectiemodel en het pattern-mixture model weer vergelijkbare resultaten: er is geen invloed van de behandeling op de FLIC-score.

Een combinatie van selectiemodellen en pattern-mixture modellen is mogelijk door gebruik te maken van de theorie van de pseudo-likelihood (zie Hoofdstuk 7). Voor de interessante stukken van zowel het selectiemodel als het pattern-mixture model worden aparte, relevante modellen gedefinieerd, gebruik makend van dezelfde assumptie voor het ontbreken van gegevens. Natuurlijk zijn er milde voorwaarden nodig om de compatibiliteit te verzekeren. Die modellen worden dan gecombineerd in een pseudo-likelihood, dewelke dan gefit wordt.

In de meeste studies wordt gebruik gemaakt van selectiemodellen. Dit wordt mee veroorzaakt door de handige opsplitsing van de veronderstellingen omtrent het

ontbreken van gegevens. Maar vermits dezelfde opsplitsing nu ook mogelijk is voor pattern-mixture modellen, kan zo'n model ook gemakkelijk gebruikt worden. Dit geeft de mogelijkheid om een model te gebruiken dat beter aansluit bij de klinische vraag, of om meer vertrouwen te hebben in de gevonden resultaten, als zowel het selectiemodel als het pattern-mixture model tot hetzelfde besluit leiden.

Toch is er nog meer onderzoek nodig naar pattern-mixture modellen, bij voorbeeld in het geval van niet-monotone ontbrekende waarden, of om de sensitiviteitsanalyse verder uit te werken.

References

Agresti, A. (1990) *Categorical Data Analysis*. New York: Wiley.

Arnold, B.C., Castillo, E. and Sarabia, J.M. (1992) *Conditionally Specified Distributions, Lecture Notes in Statistics* **73**, New York: Springer Verlag.

Arnold, B.C. and Strauss, D. (1991) Pseudolikelihood estimation: some examples. *Sankhya: the Indian Journal of Statistics - Series B*, **53**, 233–243.

Baker, S.G. (1992) A simple method for computing the observed information matrix wxhen using the EM algorithm with categorical data. *Journal of Computational and Graphical Statistics*, **1**, 63–76.

Baker, S.G. and Laird, N.M. (1988) Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, **83**, 62–69.

Bartlett, M.S. (1937) Some examples of statistical methods of research in agriculture and applied botany. *Journal of the Royal Statistical Society - Series B*, **4**, 137–170.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society - Series B*, **36**, 192–225.

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

Brown, C.H. (1990) Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, **46**, 143–155.

Clayton, D. and Hills, M. (1993) *Statistical Models in Epidemiology.* Oxford: Oxford University Press.

Conaway, M.R., Waternaux, C., Allred, E., Bellinger, D. and Leviton, A. (1992) Pre-natal blood lead levels and learning difficulties in children: an analysis of non-randomly missing categorical data. *Statistics in Medicine*, **11**, 799–811.

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

Cressie, N. (1991) *Statistics for Spatial Data.* New York: Wiley.

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.

Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns.* London: Mathematics in Biology, Academic Press.

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

Diggle, P.J. (1990) Time Series: A Biostatistical Introduction. Oxford: Oxford University Press.

Diggle, P.J. and Kenward, M.G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

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

Ekholm, A. and Skinner, C. (1998) The muscatine children's obesity data reanalysed using pattern mixture models. *Applied Statistics*, **47**, 251–263.

Fay, R.E. (1986) Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, **81**, 354–365.

Fuchs, C. (1982) Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, **77**, 270–278.

Gelman, A. and Speed, T.P. (1993) Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society - Series B*, **55**, 185–188.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geys, H., Molenberghs, G. and Ryan, L.M. (1997) Pseudo-likelihood Inference for Clustered Binary Data. *Communications in Statistics: Theory and Methods*, **26**, 2743–2767.

Geys, H., Molenberghs, G. and Ryan, L.M. (1999) Pseudo-likelihood Modelling of Multivariate Outcomes in Developmental Toxicology. *Journal of the American Statistical Association*, **94**, 000–000.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1986) Selection Modelling versus mixture modelling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples*, Ed. H. Wainer, pp. 115–142. New York: Springer Verlag.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1993) Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups. *Journal of the American Statistical Association*, **88**, 984–993.

Goss, P.E., Winer, E.P., Tannock, I.F., Schwartz, L.H. and Kremer, A.B. (1998) A randomized phase III trial comparing the new potent and selective third-generation aromatase inhibitor vorozole with megestrol acetate in postmenopausal advanced breast cancer patients. *Submitted*.

Greenlees, W.S., Reece, J.S. and Zieschang, K.D. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, **77**, 251–261.

Hand, D.J. and Crowder, M.J. (1995) *Practical Longitudinal Data Analysis*. London: Chapman and Hall.

Heitjan, D.F. and Rubin, D.B. (1991) Ignorability and coarse data. *Annals of Statistics*, **19**, 2244–2253.

Heyting, A., Tolboom, J.T.B.M. and Essers, J.G.A. (1992) Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine*, **11**, 2043–2061.

Hogan, J.W. and Laird, N.M. (1997) Mixture Models for the Joint Distribution of Repeated Measures and Event Times. *Statistics in Medicine*, **16**, 239–258.

Holland, P.W. (1986) A Comment on Remarks by Rubin and Hartigan. In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples* (pp. 149–151). New York: Springer.

Kenward, M.G., Lesaffre, E. and Molenberghs, G. (1994) An Application of Maximum Likelihood and Generalized Estimating Equations to the Analysis of Ordinal Data from a Longitudinal Study with Cases Missing at Random, *Biometrics*, **50**, 945–953.

Koch, G., Singer, J., Stokes, M., Carr, G., Cohen, S. and Forthofer, R. (1991) Some aspects of weighted least-squares analysis for longitudinal categorical data. In: J.H. Dwyer, Feinleib, M., Lipper, P. and Hoffmeister, H. (Eds.), *Statistical Models for Longitudinal Studies of Health* (pp. 215–260). Oxford: Oxford University Press.

Laird, N.M. (1988) Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305–315.

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., Lange, N. and Stram, D. (1987) Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, **82**, 97–105.

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 modelling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**, 625–632.

Le Cessie, S. and Van Houwelingen, J.C. (1994) Logistic regression for correlated binary data. *Applied Statistics*, **43**, 95-108.

Lehnen, R.G. and Koch, G.G. (1974) Analyzing panel data with uncontrolled attrition. *Public Opinion Quarterly*, **38**, 40–56.

Lesaffre, E., Molenberghs, G. and Dewulf, L. (1996) Effect of dropouts in a longitudinal study: an application of a repeated ordinal model. *Statistics in Medicine*, **15**, 1123–1141.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (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.

Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Lindsey, J.K. (1993) *Models for Repeated Measurements*. Oxford: Oxford University Press.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153–160.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996) *SAS System for Mixed Models*. SAS Institute Inc, Cary, NC, USA.

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*, **43**, 78.

Little, R.J.A. (1995) Modelling 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.

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society - Series B*, **44**, 226-233.

Longford, N.T. (1993) *Random Coefficient Models*. Oxford: Oxford University Press.

Mardia, K.V. (1970). *Families of Bivariate Distributions*, London: Griffin.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models. Second Edition*. London: Chapman and Hall.

Meilijson, I. (1989) A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society - Series B*, **51**, 127–138.

Meng, X.-L. and Rubin, D.B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, **86**, 899–909.

Michiels, B. and Molenberghs, G. (1995) Protective estimation of longitudinal categorical data with nonrandom dropout. *Lecture Notes in Statistics*, **104**, 177–184.

Michiels, B. and Molenberghs, G. (1997) Protective estimation of longitudinal categorical data with nonrandom dropout. *Communications in Statistics: Theory and Methods*, **26**, 65–94.

Michiels, B., Molenberghs, G., Bijnens, L. and Vangeneugden, T. (1998) Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Submitted*.

Michiels, B., Molenberghs, G. and Lipsitz, S.R. (1998) A Pattern-Mixture Odds Ratio Model for Incomplete Categorical Data. *Submitted*.

Molenberghs, G. and Goetghebeur, E. (1997) Simple fitting algorithms for incomplete categorical data. *Journal of the Royal Statistical Society, Series B*, **59**, 401–414.

Molenberghs, G., Goetghebeur, E., Lipsitz, S.R. and Kenward, M.G. (1999) Non-Random Missingness in Categorical Data: Strengths and Limitations. *American Statistician*, **00**, 000–000.

Molenberghs, G., Goetghebeur, E., Lipsitz, S.R., Kenward, M.G., Lesaffre, E. and Michiels, B. (1999) Missing data perspectives of the fluvoxamine data set: a review. *Statistics in Medicine*, **00**, 000–000.

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*, **00**, 000-000.

Molenberghs, G., Michiels, B., Kenward, M.G. and Diggle, P.J. (1998) Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.

Molenberghs, G., Michiels, B. and Kenward, M.G. (1998) Pseudo-likelihood for combined selection and pattern-mixture models for incomplete data. *Biometrical Journal*, **40**, 557–572.

Molenberghs, G., Michiels, B. and Lipsitz, S.R. (1999) Selection Models and Pattern-Mixture Models for Incomplete Data with Covariates. *Biometrics*, **00**, 000-000.

Nelder, J.A. and Mead, R. (1965) A simplex method for function minimisation. *The Computer Journal*, **7**, 303–313.

Park, T. and Brown, M.B. (1994) Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, **89**, 44–52.

Plackett, R.L. (1965) A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Ripley, B.D. (1981) *Spatial Statistics*. New York: John Wiley & Sons.

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 semi-parametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D.B. (1977) Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, **72**, 538–543.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

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. (1996) Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, **91**, 473–489.

Rubin, D.B. and Schenker, N. (1986) Multiple imputation for interval estimation from random samples. *Journal of the American Statistical Association*, **81**, 366–374.

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 and Hall.

Schipper, H., Clinch, J. and McMurray, A. (1984) Measuring the quality of life of cancer patients: the Functional-Living Index-Cancer: development and validation. *Journal of Clinical Oncology*, **2**, 472–483.

Schluchter, M. (1988) Analysis of incomplete multivariate data using linear models with structured covariance matrices. *Statistics in Medicine*, **7**, 317–324.

Sheiner, L.B., Beal, S.L. and Dunne, A. (1997) Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials. *Journal of the American Statistical Association*, **92**, 1235–1244.

Smith, D.M., Robertson, B. and Diggle, P.J. (1996) *Object-oriented Software for the Analysis of Longitudinal Data in S*. Technical Report MA 96/192. Department of Mathematics and Statistics, University of Lancaster, LA1 4YF, United Kingdom.

Stasny, E.A. (1986) Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, **81**, 42–47.

Verbeke, G., Lesaffre, E., Molenberghs, G., Thijs, H. and Kenward, M.G. (1998) Sensitivity analysis for non-random dropout: a local influence approach. *Submitted*.

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models In Practice: A SAS Oriented Approach*, Lecture Notes in Statistics 126. New York: Springer-Verlag.

Vonesh, E.F. and Chinchilli, V.M. (1997) *Linear and non-linear models for the analysis of repeated measurements*. Basel: Marcel Dekker.

Wainer, H. (1989) Eelworms, Bullet Holes, and Geraldine Ferraro: Some Problems with Statistical Adjustment and Some Solutions. *Journal of Educational Statistics*, **14**, 121–140.

Welsh, A.H. (1996) *Aspects of Statistical Inference*. New York: Wiley.

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) Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, **7**, 337–346.

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.

Yates, F. (1933) The analysis of replicated experiments when the field results are incomplete. *Empirical Journal of Experimental Agriculture*, **1**, 129–142.

Zhao, L.P., Lipsitz, S. and Lew, D. (1996) Regression analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165–1182.