Missing data mechanisms and pattern-mixture models

Geert Molenberghs
Universiteit Hasselt and Universi…
894 PUBLICATIONS **20,765** CITATIONS

SEE PROFILE

Bart Michiels
Johnson & Johnson
30 PUBLICATIONS **671** CITATIONS

SEE PROFILE

Statistica Neerlandica (1998) Vol. 52, nr. 2, pp. 153–161

Monotone missing data and pattern-mixture models

G. Molenberghs* and B. Michiels

Biostatistics, Limburgs Universitair Centrum, Universitaire Campus,
B3590 Diepenbeek, Belgium

M. G. Kenward

Institute of Mathematics and Statistics, The University, Canterbury, Kent
CT2 7NF, UK

P. J. Diggle

Department of Mathematics and Statistics, Lancaster University,
Lancaster LA1 4YF, UK

It is shown that the classical taxonomy of missing data models, namely missing completely at random, missing at random and informative missingness, which has been developed almost exclusively within a selection modelling framework, can also be applied to pattern-mixture models. In particular, intuitively appealing identifying restrictions are proposed for a pattern-mixture MAR mechanism.

Key Words & Phrases: missing at random, selection model.

1 Introduction

Modern missing data terminology is largely due to Rubin (1976) and Little and Rubin (1987). Their taxonomy of missing data mechanisms, which distinguishes between missing completely at random, missing at random, and informative missingness, is widely used. It is usually presented in the selection modelling framework (S), where the joint distribution of measurement and missingness processes is factorized into the marginal measurement distribution and the conditional distribution of the missingness indicators, given the outcomes. Recently, Little (1993) has suggested pattern-mixture models (PM) as a valuable alternative to selection models. An early reference is Glynn, Laird and Rubin (1986). PM models are expressed in terms of the opposite factorisation of the joint distribution.

Although S and PM models are interchangeable from a probabilistic point of view, in the sense that they represent different factorisations 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

* gmolenb@luc.ac.be

the probability of missingness depends on the unobserved outcomes. One attraction of S models is that they fit naturally into Little and Rubin's taxonomy, whereas PM models appear not to do so. The aim of this paper is to show, on the contrary, that PM models can be classified similarly, and further that the intermediate category of "missing at random" is connected to particular kinds of restrictions on the parameters of a PM model in the case of monotone missingness. This suggests to us that a purely philosophical debate about the relative merits of the S and PM paradigms is unhelpful. Instead, the focus of debate should shift to a consideration of 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 an S model seems to be obvious. On the other hand, if one is interested in the treatment effect, for various dropout patterns separately, then a PM model is a natural choice.

2 Missing data setting

In this article, we will restrict attention to a longitudinal data setting, where missingness is due to dropout. It will be shown in Section 5 that the results obtained for this case cannot be generalized to non-monotone patterns.

For each subject in the study, an outcome is intended to be observed at T time points, yielding the outcome vector Y of length T. Some subjects drop out during the study, so that for these we only observe the early components of Y. The missingness indicator will be denoted by R, indicating the time of last measurement. A person completed the study if $R = T$.

We will refer to Y as the complete data. The term full data will be used for the pair (Y, R). Our objective is to describe the joint density $f(y, r)$ of the full data.

3 Selection models

In an S model, the joint density $f(y, r)$ is factorised as

$$f(y, r) = f(y)f(r \mid y) \tag{1}$$

The classical taxonomy considers the structure of $f(r \mid y)$: missing data are said to be missing completely at random (MCAR) if the probability of missingness is independent of the data, i.e. $f(r \mid y) = f(r)$, and missing at random (MAR) if a subject's missingness mechanism depends on its observed outcomes only, $f(r = t \mid y_1, \ldots, y_T) = f(r = t \mid y_1, \ldots, y_t)$, for $t = 1, \ldots, T$. In either case, the joint density of the observed data factorises as

$$f(y_1, \ldots, y_t, r = t) = f(y_1, \ldots, y_t)f(r = t \mid y_1, \ldots, y_t)$$

with the conditional density further reducing to $f(r = t)$ in the MCAR case. If, in addition, the parameter vectors associated with the two components, say θ and ϕ, are disjoint, it follows that the log-likelihood can be expressed as the sum of two log-likelihoods, one for the measurement process parameters θ and one for the missingness process parameters ϕ. For this reason both MCAR and MAR processes are called ignorable in the context of likelihood inference, provided that the disjoint parameter condition is satisfied (Rubin, 1976). In contrast, non-ignorable missingness corresponds to a process for which $f(r \mid y)$ depends on at least one of the unobserved components of y, and possibly on the observed components $y_1, \ldots, y_t$, in which case (1) does not simplify. Maximizing the likelihood is then a somewhat more complex task. Model (1) was considered by Diggle and Kenward (1994) for informative dropout in longitudinal continuous data and by Molenberghs, Kenward and Lesaffre (1997) for categorical outcomes.

A valid criticism of S models for informative dropout, as pointed out by several discussions of Diggle and Kenward (1994), is that they often lead to very flat likelihoods or even, in extreme cases, to non-identifiability of one or more model parameters. Furthermore, particular S models rest on assumptions which are fundamentally untestable, in the sense that one can devise multiple models which would be distinguishable on the basis of complete data, but which are identical with respect to the observed data.

4 Pattern-mixture models

Little (1993, 1995) advocates PM models as a valuable alternative to S models. In a PM model, the joint density of $f(y, r)$ is factorized as

$$f(y, r) = f(r)f(y \mid r)$$

Because different densities (possibly with different parameters) are considered for each of the observed values of r, PM models are chronically under-identified. At first sight, this leaves them open to the same criticism as S models. However, Little (1993) claims that the PM approach is more honest, because parameters for which the data provide information are clearly distinguished from parameters for which there is no information at all. For example, if Y has two components Y_1 and Y_2, the second of which is incomplete, the observed data clearly give no information about the conditional density of Y_2 given that it is missing. Little therefore proceeds by imposing identifying restrictions, i.e. the inestimable parameters are identified by linking them to their estimable counterparts.

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

Clearly, S models and PM models coincide under MCAR, since in either case the joint density simplifies to $f(y)f(r)$. Next, we show that MAR can be expressed in a PM 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.

In our setting of longitudinal data with dropouts, CCMV can be defined formally as the condition that

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

whereas ACMV is the condition that

$$\forall t \geq 2, \quad \forall j < t : f(y_t \,|\, y_1, \ldots, y_{t-1}, r = j) = f(y_t \,|\, y_1, \ldots, y_{t-1}, r \geq t) \qquad (2)$$

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

Theorem 1. *For longitudinal data with dropouts, MAR* $\Longleftrightarrow$ *ACMV.*

The proof of Theorem 1 is given in the Appendix.

An interesting by-product of this theorem is that, since MAR corresponds to a set of (untestable) restrictions (ACMV) in the PM framework, MAR itself is also untestable. This fact is often overlooked in the S 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 $r = t$ and some other incomplete pattern $r = s$, than between those for $r = t$ and the completers, $r = T$.

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 $r = 2, \ldots, T$ collapse into a single pattern $r \geq 2$. Applying ACMV restrictions to $r = 1$ and $r \geq 2$ leads to the construction of the density $f(y_2 \,|\, y_1, r = 1) = f(y_2 \,|\, y_1, r \geq 2)$, as in (2). Multiplying by $f(y_1 \,|\, r = 1)$ leads to $f(y_1, y_2 \,|\, r = 1)$, thus determining the joint densities of $f(y_1, y_2 \,|\, r)$ for all $r = 1, \ldots, T$. Next, $f(y_3 \,|\, y_1, y_2, r)(r = 1, 2)$ can be calculated from $f(y_3 \,|\, y_1, y_2, r \geq 3)$. We then proceed by induction to construct all joint densities.

5 Non-monotone patterns: A counter example

It has to be noted that the result of Theorem 1 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:

$$f(r \,|\, y) = \Pr(r_1, r_2 \,|\, y_1, y_2) = \begin{cases} p & \text{if } (r_1, r_2) = (0, 0) \\ q_{y_1} & \text{if } (r_1, r_2) = (1, 0) \\ s_{y_2} & \text{if } (r_1, r_2) = (0, 1) \\ 1 - p - q_{y_1} - s_{y_2} & \text{if } (r_1, r_2) = (1, 1) \end{cases} \qquad (3)$$

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 $r = (0, 1)$:

$$r = (1, 0) : f(y_1, y_2 \mid r = (1, 0)) = f(y_1 \mid r = (1, 0)) \cdot f(y_2 \mid y_1, r = (1, 1)) \quad (4)$$

$$r = (0, 1) : f(y_1, y_2 \mid r = (0, 1)) = f(y_2 \mid r = (0, 1)) \cdot f(y_1 \mid y_2, r = (1, 1)) \quad (5)$$

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 (4) we derive

$$\frac{f(r = (1, 0) \mid y_1, y_2) f(y_1, y_2)}{f(r = (1, 0))} = \frac{f(r = (1, 0) \mid y_1) f(y_1)}{f(r = (1, 0))} \, \frac{f(r = (1, 1) \mid y_1, y_2) f(y_1, y_2)}{f(r = (1, 1) \mid y_1) f(y_1)}$$

$$\Updownarrow$$

$$f(r = (1, 1) \mid y_1, y_2) = f(r = (1, 1) \mid y_1)$$

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

Similarly, condition (5) implies that q_{y_1} is constant.

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

6 Conclusion

In a missing data context, the choice of modelling framework needs careful consideration. The simplicity of the classical MCAR, MAR, and informative 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. For the latter, the interpretation is equally instructive as MAR. The intermediate case corresponds to an explicit and reasonably natural set of restrictions on the unidentifiable components of the full data distribution.

Acknowledgement

The authors thank the referee for suggestions and improvements.

References

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

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. (1994), A class of pattern-mixture models for normal incomplete data, *Biometrika* **81**, 471–483.

Little, R. J. A. (1995), Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**, 1112–1121.

Little, R. J. A. and D. B. Rubin (1987), *Statistical analysis with missing data*, John Wiley and Sons, New York.

Molenberghs, G., M. G. Kenward and E. Lesaffre (1997), The analysis of longitudinal ordinal data with nonrandom dropout, *Biometrika* **84**, 33–44.

Glynn, R. J., N. M. Laird and D. B. Rubin (1986), Selection modeling versus mixture modeling with nonignorable nonresponse, in: H. Wainer (ed.), *Drawing inferences from self-selected samples*, Springer-Verlag, New York, 115–142.

Rubin, D. B. (1976), Inference and missing data, *Biometrika* **63**, 581–592.

Received: August 1995. Revised: February 1997.

Appendix

The MAR assumption states that

$$f(r = t \,|\, y_1, \dots, y_T) = f(r = t \,|\, y_1, \dots, y_t) \tag{6}$$

and the ACMV assumption that

$$\forall t \geq 2, \quad \forall j < t : f(y_t \,|\, y_1, \dots, y_{t-1}, r = j) = f(y_t \,|\, y_1, \dots, y_{t-1}, r \geq t) \tag{7}$$

First, a lemma will be established.

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

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

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

$$\begin{aligned} f(y_t \mid y_1, \ldots, y_{t-1}, r \geq t) f(r \geq t) &= f(y_t \mid y_1, \ldots, y_{t-1}) \\ &\quad - \sum_{i=1}^{t-1} f(y_t \mid y_1, \ldots, y_{t-1}, r = i) f(r = i) \\ &= f(y_t \mid y_1, \ldots, y_{t-1}) \\ &\quad - \sum_{i=1}^{t-1} f(y_t \mid y_1, \ldots, y_{t-1}) f(r = i) \\ &= f(y_t \mid y_1, \ldots, y_{t-1}) \left[1 - \sum_{i=1}^{t-1} f(r = i) \right] \\ &= f(y_t \mid y_1, \ldots, y_{t-1}, r = j) \left[1 - \sum_{i=1}^{t-1} f(r = i) \right] \\ &= f(y_t \mid y_1, \ldots, y_{t-1}, r = j) f(r \geq t) \end{aligned}$$

This completes the proof.
We are now able to prove Theorem 1.

MAR $\Rightarrow$ ACMV

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

$$Q = \frac{f(y_1, \ldots, y_T) f(r = i \mid y_1, \ldots, y_i)}{f(y_1, \ldots, y_i) f(r = i \mid y_1, \ldots, y_i)} = f(y_{i+1}, \ldots, y_T \mid y_1, \ldots, y_i) \tag{8}$$

Further, one can always write:

$$\begin{aligned} Q &= \frac{f(y_{i+1}, \ldots, y_T \mid y_1, \ldots, y_i, r = i) f(y_1, \ldots, y_i \mid r = i) f(r = i)}{f(y_1, \ldots, y_i \mid r = i) f(r = i)} \\ &= f(y_{i+1}, \ldots, y_T \mid y_1, \ldots, y_i, r = i) \end{aligned} \tag{9}$$

Equating expressions (8) and (9) for Q we see that

$$f(y_{i+1}, \ldots, y_T \mid y_1, \ldots, y_i, r = i) = f(y_{i+1}, \ldots, y_T \mid y_1, \ldots, y_i). \tag{10}$$

To show that (10) implies the ACMV conditions (7), we will use the induction principle on t. First, consider the case $t = 2$.

Using (10) for $i = 1$, and integrating over $y_3, \ldots, y_T$, we obtain

$$f(y_2 \mid y_1, r = 1) = f(y_2 \mid y_1)$$

leading to, using Lemma 1,

$$f(y_2 \mid y_1, r = 1) = f(y_2 \mid y_1, r \geq 2)$$

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

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

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

$$f(y_{j+1}, \ldots, y_i \,|\, y_1, \ldots, y_j, r = j) = f(y_{j+1}, \ldots, y_i \,|\, y_1, \ldots, y_j) \qquad (11)$$

After integration over $y_{i+2}, \ldots, y_T$, equation (10) leads to

$$f(y_{j+1}, \ldots, y_{i+1} \,|\, y_1, \ldots, y_j, r = j) = f(y_{j+1}, \ldots, y_{i+1} \,|\, y_1, \ldots, y_j) \qquad (12)$$

Dividing (12) by (11) and equating the left and right hand sides, we find that

$$f(y_{i+1} \,|\, y_1, \ldots, y_i, r = j) = f(y_{i+1} \,|\, y_1, \ldots, y_i)$$

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

$$\forall t \geq 2, \quad \forall j < t : f(y_t \,|\, y_1, \ldots, y_{t-1}, r = j) = f(y_t \,|\, y_1, \ldots, y_{t-1}) \qquad (13)$$

We now factorise the full data density as

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

Using (13), it follows that

$$\begin{aligned}f(y_1, \ldots, y_T, r = i) &= f(y_1, \ldots, y_i \,|\, r = i) f(r = i) \prod_{t=i+1}^{T} f(y_t \,|\, y_1, \ldots, y_{t-1})\\ &= f(y_1, \ldots, y_i \,|\, r = i) f(r = i) f(y_{i+1}, \ldots, y_T \,|\, y_1, \ldots, y_i)\\ &= \frac{f(y_1, \ldots, y_i \,|\, r = i) f(r = i)}{f(y_1, \ldots, y_i)}\\ &\quad \times f(y_1, \ldots, y_i) f(y_{i+1}, \ldots, y_T \,|\, y_1, \ldots, y_i)\\ &= \frac{f(y_1, \ldots, y_i \,|\, r = i) f(r = i)}{f(y_1, \ldots, y_i)} f(y_1, \ldots, y_T)\\ &= f(r = i \,|\, y_1, \ldots, y_i) f(y_1, \ldots, y_T) \qquad (14)\end{aligned}$$

An alternative factorisation of $f(y, r)$ gives

$$f(y_1, \ldots, y_T, r = i) = f(r = i \mid y_1, \ldots, y_T) f(y_1, \ldots, y_T) \tag{15}$$

It follows from (14) and (15) that

$$f(r = i \mid y_1, \ldots, y_T) = f(r = i \mid y_1, \ldots, y_i)$$

completing the proof of Theorem 1.