Statistical Science
1998, Vol. 13, No. 3, 236–247

Likelihood Based Frequentist Inference When Data Are Missing at Random

M. G. Kenward and G. Molenberghs

Abstract. One of the most often quoted results from the original work of Rubin and Little on the classification of missing value processes is the validity of likelihood based inferences under missing at random (MAR) mechanisms. Although the sense in which this result holds was precisely defined by Rubin, and explored by him in later work, it appears to be now used by some authors in a general and rather imprecise way, particularly with respect to the use of frequentist modes of inference. In this paper an exposition is given of likelihood based frequentist inference under an MAR mechanism that shows in particular which aspects of such inference cannot be separated from consideration of the missing value mechanism. The development is illustrated with three simple setups: a bivariate binary outcome, a bivariate Gaussian outcome and a two-stage sequential procedure with Gaussian outcome and with real longitudinal examples, involving both categorical and continuous outcomes. In particular, it is shown that the classical expected information matrix is biased and the use of the observed information matrix is recommended.

Key words and phrases: Dropout, expected information matrix, likelihood function, likelihood ratio, longitudinal data, observed information matrix, sequential methods.

1. INTRODUCTION

For over two decades, following the pioneering work of Rubin (1976) and Little (1976), there has been a growing literature on the problem of analyzing incomplete data. This is particularly relevant for longitudinal data where partially observed sequences, especially due to dropout (a patient leaves the study at some time after which no more measurements are taken), are very common. Much of this work is based on the classification of missing data mechanisms, described by Little and Rubin (1987). They define *missing completely at random* (MCAR) to be a process in which the probability of dropout is completely independent of the measurement process. A process is termed *missing at random* (MAR) if the probability of dropout is conditionally independent of the unobserved measurements given the observed measurements. Processes that are neither MCAR nor MAR are called *nonignorable* (NI), in which the probability of dropout depends on unobserved measurements. The development of analyses under an NI process presents problems that we do not address here.

Following the original work of Rubin and Little, there has evolved a general view that "likelihood methods" that ignore the missing value mechanism are valid under an MAR process, where likelihood is interpreted in a frequentist sense. This statement needs careful qualification, however, and it is the purpose of this paper to provide an exposition of the precise sense in which frequentist methods of inference are justified under MAR processes.

Rubin (1976) has shown that MAR (and parameter distinctness) is necessary and sufficient to ensure validity of *direct-likelihood* inference when ignoring the process that causes missing data. Here, direct-likelihood inference is defined as an "inference that results solely from ratios of the likelihood function for various values of the parameter," in agreement with the definition in Edwards (1972). In the concluding section of the same paper, Rubin

M. G. Kenward is Reader, Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF, United Kingdom. E-mail: m.g.kenward@ukc.ac.uk. G. Molenberghs is Associate Professor, Biostatistics, Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium.

remarks:

> One might argue, however, that this apparent simplicity of likelihood and Bayesian inference really buries the important issues.... likelihood inferences are at times surrounded with references to the sampling distributions of likelihood statistics. Thus, practically, when there is the possibility of missing data, some interpretations of Bayesian and likelihood inference face the same restrictions as sampling distribution inference. The inescapable conclusion seems to be that when dealing with real data, the practising statistician should explicitly consider the process that causes missing data far more often than he does.

In essence, the problem from a frequentist point of view is that of identifying and using the appropriate sampling distribution. This is obviously relevant for determining distributions of test statistics, expected values of the information matrix and measures of precision.

Little and Rubin (1987) discuss several aspects of this problem and propose using the observed information matrix to circumvent problems associated with the determination of the correct expected information matrix. Laird (1988) makes a similar point in the context of incomplete longitudinal data analysis.

In a variety of settings, several authors have reexpressed this preference for the observed information matrix and derived methods to compute it: Meng and Rubin (1991), the supplemented EM algorithm; Baker (1992), composite link models; Fitzmaurice, Laird and Lipsitz (1994), incomplete longitudinal binary data; and Jennrich and Schluchter (1986). A group of authors has used the observed information matrix, without reference to the problems associated with the expected information: Louis (1982), Meilijson (1989) and Kenward, Lesaffre and Molenberghs (1994).

However, others, while claiming validity of analysis under MAR mechanisms, have used expected information matrices, and other measures of precision that do not account for the missingness mechanism (Murray and Findlay, 1988; Patel, 1991). A number of references is given in Baker (1992). The expected information in these papers is *wrong* because the expectation is taken under MCAR, in which the missing value mechanism is independent of the distribution of the outcome data. It is clear that the problem as identified in the initial work of Rubin (1976) is not fully appreciated in the more recent literature. An exception to this is Heitjan's (1994) clear restatement of the problem.

A recent exchange of correspondence (Diggle, 1992; Heitjan, 1993; and Diggle, 1993) indicates a genuine interest in these issues and suggests a need for clarification. In Section 2 we sketch a general framework of likelihood inference under an MAR process. The difference between the expected information matrix with and without taking the missing data mechanism into account is elucidated and the relevance of this for Wald and score statistics is discussed. In particular, the use of the observed information matrix is recommended. Analytic and numerical illustrations of this difference are provided in Section 3 using as examples, bivariate binary and bivariate Gaussian data, and a simple group sequential setting. In Section 4 three real longitudinal examples are used for practical illustration.

2. INFORMATION AND SAMPLING DISTRIBUTIONS

Let the vector random variable $\mathbf{Y}$ correspond to the complete set of measurements on an individual and let $\mathbf{R}$ be the associated missing value indicator. For a particular realization of this pair $(\mathbf{y}, \mathbf{r})$ the elements of $\mathbf{r}$ take the values 1 and 0 indicating, respectively, whether the corresponding values of $\mathbf{y}$ are observed or not. Let $(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$ denote the partition of $\mathbf{y}$ into the respective sets of observed and missing data. We assume that the joint distribution of $(\mathbf{Y}, \mathbf{R})$ is regular in the sense of Cox and Hinkley (1974, page 281).

We are concerned here with the sampling distributions of certain statistics under MCAR and MAR mechanisms. These mechanisms were described in the Introduction and can be defined more formally as follows (Little and Rubin, 1987). Under an MCAR mechanism $P(\mathbf{R} = \mathbf{r} \mid \mathbf{y}) = P(\mathbf{R} = \mathbf{r})$ and the joint distribution of the *observed* data partitions in an obvious way:

$$f(\mathbf{y}_{\text{obs}}, \mathbf{r}) = f(\mathbf{y}_{\text{obs}}) f(\mathbf{r}).$$

Under an MAR mechanism $\mathrm{P}(\mathbf{R} = \mathbf{r} \mid \mathbf{y}) = \mathrm{P}(\mathbf{R} = \mathbf{r} \mid \mathbf{y}_{\text{obs}})$ and again the joint distribution of the observed data, and hence the likelihood, can be partitioned,

$$f(\mathbf{y}_{\text{obs}}, \mathbf{r}) = f(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}) f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}; \boldsymbol{\beta})$$

for parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. In terms of the log-likelihood function we have

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}_{\text{obs}}, \mathbf{r}) = \ell_1(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}}) + \ell_2(\boldsymbol{\beta}; \mathbf{r}). \tag{1}$$

Unless otherwise stated it is assumed that $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are distinct (the assumption of separability). As described in the Introduction, this partition of the likelihood has, with important exceptions, been taken for granted to mean that, under an MAR mechanism, likelihood methods based on $\ell_1(\cdot)$ alone are valid for inferences about $\boldsymbol{\theta}$ *even when interpreted in the broad frequentist sense*. We now consider more precisely the sense in which the different elements of the frequentist likelihood methodology can be regarded as valid in general under the MAR mechanism. It is now well known that such inferences are valid under an MCAR mechanism (Rubin, 1976, Section 6).

First we note that under the MAR mechanism $\mathbf{r}$ is *not* an ancillary statistic for $\boldsymbol{\theta}$ in the extended sense of Cox and Hinkley (1974, page 35). Hence we are not justified in restricting the sample space from that associated with the pair $(\mathbf{Y}, \mathbf{R})$. In considering the properties of frequentist procedures below we therefore define the appropriate sampling distributions to be that determined by this pair. We call this the *unconditional* sampling framework. By working within this framework we do need to consider the missing value mechanism. We shall be comparing this with the sampling distribution that would apply if $\mathbf{r}$ were fixed by design, that is, if we repeatedly sampled using the distribution $f(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta})$. If this sampling distribution were appropriate, such as in the MCAR case, this would lead directly to the use of $\ell_1(\cdot)$ as a basis for inference. We call this the *naive* sampling framework. Little (1976), in a comment on the paper by Rubin (1976), mentions explicitly the role played by the nonresponse pattern. He argues: "For sampling based inferences, a first crucial question concerns when it is justified to condition on the observed pattern, that is on the event $R = r \ldots$. A natural condition is that R should be ancillary $\ldots$. Otherwise the pattern on its own carries at least some information about θ, which should in principle be used."

Certain elements of the frequentist methodology can be justified immediately from (1). The maximum likelihood estimator obtained from maximizing $\ell_1(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}})$ alone is identical to that obtained from maximizing the complete log-likelihood function. Similarly the maximum likelihood estimator of $\boldsymbol{\beta}$ is functionally independent of $\boldsymbol{\theta}$ and so any maximum likelihood ratio concerning $\boldsymbol{\theta}$, with common $\boldsymbol{\beta}$, will involve $\ell_1(\cdot)$ only. Because these statistics are identical whether derived from $\ell_1(\cdot)$ or the complete log-likelihood it follows at once that they have the required properties under the naive sampling framework. See, for example, Rubin (1976), Little (1976) and Little and Rubin (1987, Section 5.2).

An important element of likelihood-based frequentist inference is the derivation measures of precision of the maximum likelihood estimators from the information. For this either the observed information i_O can be used, where

$$i_O(\theta_j, \theta_k) = -\frac{\partial^2 \ell(\cdot)}{\partial \theta_j \partial \theta_k},$$

or the expected information i_E, where

$$i_E(\theta_j, \theta_k) = E\{i_O(\theta_j, \theta_k)\}. \tag{2}$$

The argument above justifying the use of the maximum likelihood estimators from $\ell_1(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}})$ applies equally well to the use of the inverse of the *observed* information derived from $\ell_1(\cdot)$ as an estimate of the asymptotic variance–covariance matrix of these estimators. This has been pointed out by Little and Rubin (1987, Section 8.2.2) and Laird (1988, page 307). In addition there are other reasons for preferring the observed information matrix (see, e.g., Efron and Hinkley, 1978). Given the relative ease with which the observed information matrix can be calculated, using numerical differentiation if necessary, its use in missing data problems should be the rule rather than the exception.

The use of the expected information matrix is more problematical. The expectation in (2) needs to be taken over the *unconditional* sampling distribution (the *unconditional information* i_U) and consequently the use of the naive sampling framework (producing the *naive information* i_N) can lead to inconsistent estimates of precision. In the next section we give three examples of the bias resulting from the use of the naive framework. It is possible, however, as we show below, to calculate the unconditional information by taking expectations over the appropriate distribution and so correct this bias. Although this added complication is generally unnecessary in practice, given the availability of the observed information, it does allow a direct examination of the effect of ignoring the missing value mechanism on the expected information.

More formally, it can be shown that, under the usual regularity conditions, the maximum likelihood estimator is asymptotically normally distributed around the true parameter and with variance–covariance matrix the inverse of the unconditional expected information. This is immediate from Welsh (1996, page 197, Corollary 4.6), with the Fisher information properly calculated over the unconditional space.

This leaves us with two options: the observed information matrix and the unconditional expected information matrix. While the first one is standard

and routinely implemented in statistical packages, the second one is not and has the further disadvantage that it requires correct specification of the MAR missingness mechanism. Thus, practically, the observed information matrix is the only choice.

As part of the process of frequentist inference we also need to consider the sampling distribution of the test statistics. Provided that use is made of the likelihood ratio, or Wald score statistics based on the observed information, then reference to a null asymptotic χ^2 distribution will be appropriate because this is derived from the *implicit* use of the unconditional sampling framework. Only in those situations in which the sampling distribution is explicitly constructed must care be taken to ensure that the unconditional framework is used; that is, account must be taken of the missing data mechanism.

3. ILLUSTRATION

3.1 Bivariate Binary Data

Suppose that each member of the pair of observations (Y_{i1}, Y_{i2}), from unit i, $i = 1, \ldots, n$, is a binary random variable, with associated probabilities $P(Y_{i1} = 1) = \lambda$ and $P(Y_{i2} = 1) = \theta$. It is assumed that an MAR mechanism is operating with respect to the second observation; that is, the probability of Y_{i2} being missing depends on Y_{i1} alone. It follows that Y_{i1} is always observed. We want to compare the naive information i_N with the unconditional information i_U for this setup. We begin by assuming that Y_{i1} and Y_{i2} are independent. The joint distribution of Y_{i1}, Y_{i2} and R_i can then be partitioned as follows: $f(y_{i1}, y_{i2}, r_i) = f(y_{i1})f(y_{i2})f(r_i \mid y_{i1})$. It follows at once that the observed information for θ can be expressed

$$i_O(\theta, \theta) = \frac{1}{\theta^2}\sum_{i=1}^{m} y_{i2} + \frac{1}{(1-\theta)^2}\left(m - \sum_{i=1}^{m} y_{i2}\right), \tag{3}$$

where m denotes the number of observations observed on the second occasion. The other elements of the information matrix are not relevant to the development.

The naive information is obtained from (3) by taking expectations over the joint distribution of m independent binary random variables with parameter θ; that is, we take expectations over the observed pattern of observations but not conditional on r, the realization of the random variable R associated with the occurrence of that particular pattern. Hence from (3) we get $i_N(\theta, \theta) = m\theta^{-1}(1-\theta)^{-1}$.

The unconditional information is derived in two steps. First we obtain the conditional expectation of (3) with respect to $Y \mid R$. For this we need

$$\begin{aligned} E_{Y|r_i=1}(Y_{i2}) &= P(Y_{i2} = 1 \mid r_i = 1) \\ &= P(Y_{i2} = 1) = \theta, \end{aligned}$$

because of the independence of (Y_{i1}, R_i) and Y_{i2} under the MAR mechanism. It follows that $E_{Y|R}\{i_O(\theta, \theta)\} = m\theta^{-1}(1-\theta)^{-1}$ for m the number of observations on the second occasion. We are now treating m as the realization of a random variable M, over which we take expectations to obtain the unconditional information. Setting $\pi = P(R_i = 1)$ we have

$$i_U(\theta, \theta) = E_R\left(m\frac{1}{\theta(1-\theta)}\right) = \frac{E_R(m)}{\theta(1-\theta)} = \frac{n\pi}{\theta(1-\theta)}.$$

Replacing π by the estimate m/n it can be seen that in practice the naive and unconditional information are equivalent and sampling based inferences that use the naive information are valid. Under independence, the data are *observed at random* (OAR) and the result above is just a manifestation of the general validity of sampling based methods under the combination of MAR and OAR, or equivalently MCAR, as pointed out by Heitjan (1994, page 706).

We now introduce dependence between Y_{i1} and Y_{i2}. This can be expressed through the conditional success probabilities of Y_{i2}: $\theta_1 = P(Y_{i2} = 1 \mid y_{i1} = 1)$ and $\theta_0 = P(Y_{i2} = 1 \mid y_{i1} = 0)$.

The off-diagonal elements of the observed information matrix are zero, so we need consider only the information for one of θ_0 and θ_1 to contrast the naive and unconditional forms of the expected information. For θ_1 the observed information reads

$$\begin{aligned} i_O(\theta_1, \theta_1) = &\frac{1}{\theta_1^2}\sum_{i=1}^{m} y_{i1}y_{i2} \\ &+ \frac{1}{(1-\theta_1)^2}\sum_{i=1}^{m} y_{i1}(1-y_{i2}). \end{aligned} \tag{4}$$

For the naive information it follows at once, taking expectations in (4),

$$i_N(\theta_1, \theta_1) = \frac{m\lambda}{\theta_1(1-\theta_1)}. \tag{5}$$

For the unconditional information we need to consider first the conditional expectations over $Y \mid R$. Define $\eta_0 = P(R_i = 1 \mid y_{i1} = 0)$ and $\eta_1 = P(R_i = 1 \mid y_{i1} = 1)$. It follows that $P(R_i = 1) = \eta =$

$\lambda\eta_1 + (1-\lambda)\eta_0$. We then have

$$E_{Y|R}(Y_{i2}) = P(Y_{i2} = 1 \mid r_i = 1) = \frac{\lambda\theta_1\eta_1 + (1-\lambda)\theta_0\eta_0}{\lambda\eta_1 + (1-\lambda)\eta_0},$$

$$E_{Y|R}(Y_{i1}Y_{i2}) = P(Y_{i1} = 1, Y_{i2} = 1 \mid r_i = 1) = \frac{\lambda\theta_1\eta_1}{\lambda\eta_1 + (1-\lambda)\eta_0} = \frac{\lambda\theta_1\eta_1}{\eta},$$

and similarly for $E_{Y|R}(Y_{i1}(1 - Y_{i2}))$. Combining these with (4) we get

$$E_{Y|R}\{i_O(\theta_1, \theta_1)\} = \frac{m\lambda\eta_1}{\eta\theta_1(1-\theta_1)}. \tag{6}$$

We now take expectations over R. Noting that $E_R(m) = n\eta$ we have

$$i_U(\theta_1, \theta_1) = \frac{n\lambda\eta_1}{\theta_1(1-\theta_1)}, \tag{7}$$

with a similar expression for $i_U(\theta_0, \theta_0)$.

We are now in a position to consider the conditions under which the naive and unconditional expectations are equivalent. From (5) and (7), it can be seen that conditions for $E_R(i_N(\theta_1, \theta_1)) = i_U(\theta_1, \theta_1)$ and $E_R(i_N(\theta_0, \theta_0)) = i_U(\theta_0, \theta_0)$ are $E_R(m/n) = \eta_1 = \eta_0$ and hence $\eta = \eta_1 = \eta_0$, the requirement for an MCAR mechanism to operate. It follows, as expected, that the MCAR mechanism is both a necessary and sufficient condition for the equivalence of the two forms of information.

These findings are illustrated with some numerical results. It is necessary to consider only the diagonal elements of i_N and i_U because the off-diagonal elements are all zero. We take a sample of size $n = 1{,}000$ and consider various settings for the parameters (as shown in Table 1). We performed a simulation run of 500 replicates for each setting. Results are presented in Table 2. The simulations agree very closely with the unconditional information. Although not reported here, simulation runs with larger sample sizes produced similar results with improved agreement between theoretical and simulated values.

TABLE 1
Bivariate binary data: parameter settings

Model	λ	θ_1	θ_0	η_1	η_0
1	0.50	0.25	0.75	0.75	0.25
2	0.50	0.25	0.75	0.25	0.75
3	0.25	0.40	0.60	0.40	0.60
4	0.25	0.40	0.60	0.60	0.40

To illustrate the point that basing the computation of test statistics on either the observed information matrix or the unconditional expectation is sufficient to obtain valid inference, we consider three Wald test statistics. The null hypotheses H_{01}–H_{03} are that each of the three parameters λ, θ_1 and θ_0 are equal to the true value. The four parameter settings displayed in Table 1 are revisited (Table 3) under both the unconditional sampling framework (i.e., with m, the number of complete cases, varying at random), as well as under a few naive frameworks, where the value of m is considered fixed at three possible values: (1) at its expected value, (2) at about two standard deviations below its expected value, (3) at a value well below the minimal m observed under the unconditional sampling scheme. Should the correct reference distribution be χ^2 with one degree of freedom, then the coverage probability, for 500 replicates, has probability interval (93.05; 96.95). Clearly, the values obtained under the unconditional framework are in agreement. Combining all 12 coverages leads to 94.88, well within the interval (94.44; 95.56). In fact, the first naive framework, where m equals its expected value, shows an only slightly increased dispersion. However, suspicion is raised for the second naive framework, while the third one is dramatically different. As expected, the behavior of hypothesis tests concerning θ_1 and θ_0 is much less affected by the choice of sampling framework. These conclusions are supported by QQ plots for the Wald test statistics against the quantiles of a χ^2 reference distribution.

TABLE 2
Bivariate binary data: diagonal of the information matrix (naive, unconditional, and simulated). Sample size is $n = 1{,}000$ *(500 replications)*

	Naive $i_N(\cdot,\cdot)$			Uncond. $i_U(\cdot,\cdot)$			Simulated $\widehat{i_O}(\cdot,\cdot)$		
Model	λ	θ_1	θ_0	λ	θ_1	θ_0	λ	θ_1	θ_0
1	4000	1333	1333	4000	2000	667	4004	2001	678
2	4000	1333	1333	4000	667	2000	4004	672	2014
3	5333	573	1719	5333	417	1875	5349	420	1879
4	5333	469	1406	5333	625	1250	5348	631	1259

TABLE 3

Bivariate binary data: coverage probabilities (×1,000) for Wald test statistics. Sample size is n = 1,000 (500 replications). The null hypotheses are H_{01}: $\hat{\lambda} = \lambda$, H_{02}: $\hat{\theta}_1 = \theta_1$, H_{03}: $\hat{\theta}_0 = \theta_0$. For the naive sampling frameworks, m denotes the fixed number of complete cases

	Uncond.			Naive(1)				Naive(2)				Naive(3)			
Model	H_{01}	H_{02}	H_{03}	m	H_{01}	H_{02}	H_{03}	m	H_{01}	H_{02}	H_{03}	m	H_{01}	H_{02}	H_{03}
1	946	944	948	500	968	964	928	470	884	958	954	400	68	952	942
2	946	948	954	500	972	932	956	470	894	942	962	400	54	962	950
3	954	936	956	550	938	942	940	520	952	938	954	450	822	940	960
4	954	942	958	540	960	966	936	420	944	936	958	350	778	954	944

3.2 Bivariate Gaussian Data

Little and Rubin (1987) state: "If the data are MCAR, the expected information matrix of $\theta = (\mu, \Sigma)$ represented as a vector" is block diagonal. "The observed information matrix, which is calculated and inverted at each iteration of the Newton–Raphson algorithm, is not block diagonal with respect to μ and Σ, so this simplification does not occur if standard errors are based on this matrix. On the other hand, the standard errors based on the observed information matrix are more conditional and thus valid when the data are MAR but not MCAR, and hence should be preferable to those based on [the expected information] in applications."

Suppose now that we have n independent pairs of observations (Y_{i1}, Y_{i2}) each with a bivariate Gaussian distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and variance–covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

It is assumed that m complete pairs, and only the first member (Y_{i1}) of the remaining pairs, are observed. The log-likelihood can be expressed as the sum of the log-likelihoods for the complete and incomplete pairs:

$$\ell = \sum_{i=1}^{m} \ln f(y_{i1}, y_{i2} \mid \mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}) + \sum_{i=m+1}^{n} \ln f(y_{i1} \mid \mu_1, \sigma_{11}),$$

which, in the Gaussian setting, has kernel

$$\ell = -\frac{n-m}{2} \ln \sigma_{11} - \frac{m}{2} \ln |\Sigma| - \frac{1}{2\sigma_{11}} \sum_{i=m+1}^{n} (y_{i1} - \mu_1)^2 - \frac{1}{2} \sum_{i=1}^{m} \begin{pmatrix} y_{i1} - \mu_1 \\ y_{i2} - \mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_{i1} - \mu_1 \\ y_{i2} - \mu_2 \end{pmatrix}.$$

Straightforward differentiation produces the elements of the observed information matrix that relate to $\boldsymbol{\mu}$:

$$i_O(\boldsymbol{\mu}, \boldsymbol{\mu}) = (n-m) \begin{pmatrix} \sigma_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + m \boldsymbol{\Sigma}^{-1}$$

and

$$i_O(\mu_j, \sigma_{kl}) = \begin{cases} \displaystyle\sum_{i=m+1}^{n} \frac{y_{i1} - \mu_1}{\sigma_{11}^2} + \sum_{i=1}^{m} \mathbf{e}_1^T \boldsymbol{\Sigma}^{-1} \mathbf{E}_{11} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} y_{i1} - \mu_1 \\ y_{i2} - \mu_2 \end{pmatrix}, & j = k = l = 1, \\ \displaystyle\sum_{i=1}^{m} \mathbf{e}_j^T \boldsymbol{\Sigma}^{-1} \mathbf{E}_{kl} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} y_{i1} - \mu_1 \\ y_{i2} - \mu_2 \end{pmatrix}, & \text{otherwise,} \end{cases} \tag{8}$$

for

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

$$\mathbf{E}_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{E}_{12} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and

$$\mathbf{E}_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

For the naive information we just take expectations of these quantities over $(Y_{i1}, Y_{i2})^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, m$ and $Y_{i1} \sim \mathrm{N}(\mu_1, \sigma_{11})$ for $i = m+1, \ldots, n$. It follows at once that the cross-terms linking the mean and variance–covariance parameters vanish, establishing the familiar orthogonality property of these sets of parameters in the Gaussian setting. We now examine the behavior of the expected information under the actual sampling process implied by the MAR mechanism.

We need to consider first the conditional expectation of these quantities given the occurrence of R, the dropout pattern. Because (Y, R) enters the expression for $i_U(\boldsymbol{\mu}, \boldsymbol{\mu})$ only through m, the naive and

unconditional information matrices for $\boldsymbol{\mu}$ are effectively equivalent. However, we show now that this is not true for the cross-term elements of the information matrices. Define $\alpha_j = E(Y_{i1} \mid r_i = j) - \mu_1$. For the conditional expectation of Y_{i2} we have

$$\begin{aligned} E_{Y|R}(Y_{i2}) &= E(Y_{i2} \mid r_i = 1) \\ &= \int \left\{ y_{i2} \int f(y_{i2} \mid y_{i1})\, dy_{i2} \right\} f(y_{i1} \mid r_i = 1)\, dy_{i1} \\ &= \mu_2 - \sigma_{12}\sigma_{11}^{-1}\mu_1 \\ &\quad + \frac{\sigma_{12}}{\sigma_{11} P(r_i = 1)} \int y_{i1} f(y_{i1},\ r_i = 1)\, dy_{i1} \\ &= \mu_2 + \sigma_{12}\sigma_{11}^{-1}\{E(Y_{i1} \mid r_i = 1) - \mu_1\} \end{aligned}$$

or

$$E_{Y|R}(Y_{i2} - \mu_2) = \beta\alpha_1$$

for $\beta = \sigma_{12}\sigma_{11}^{-1}$. Hence

$$E_{Y|R}\left\{ \begin{pmatrix} Y_{i1} - \mu_1 \\ Y_{i2} - \mu_2 \end{pmatrix} \right\} = \alpha_1 \begin{pmatrix} 1 \\ \beta \end{pmatrix}.$$

Noting that

$$\boldsymbol{\Sigma}^{-1} \begin{pmatrix} 1 \\ \beta \end{pmatrix} = \begin{pmatrix} \sigma_{11}^{-1} \\ 0 \end{pmatrix} = \sigma_{11}^{-1} \mathbf{e_1},$$

we then have, from (8),

$$E_{Y|R}\{i_O(\mu_j, \sigma_{kl})\} = \begin{cases} (n-m)\dfrac{\alpha_0}{\sigma_{11}^2} + m\dfrac{\alpha_1}{\sigma_{11}} \mathbf{e_1}^T \boldsymbol{\Sigma}^{-1} \mathbf{E}_{11} \mathbf{e_1}, & \\ \qquad\qquad\qquad\qquad\qquad\qquad j = k = l = 1, & \\ m\dfrac{\alpha_1}{\sigma_{11}} \mathbf{e_j}^T \boldsymbol{\Sigma}^{-1} \mathbf{E}_{kl} \mathbf{e_1}, & \text{otherwise.} \end{cases}$$

Finally, taking expectations over R, we get the following for the cross-terms of the unconditional information matrix:

$$i_U(\boldsymbol{\mu}, \sigma_{11}) = \frac{n}{\sigma_{11}} \left\{ \frac{(1-\pi)\alpha_0}{\sigma_{11}} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{\pi\alpha_1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{pmatrix} \sigma_{22} \\ -\sigma_{12} \end{pmatrix} \right\}, \tag{9}$$

$$i_U(\boldsymbol{\mu}, \sigma_{12}) = \frac{n\pi\alpha_1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{pmatrix} -\beta \\ 1 \end{pmatrix}, \tag{10}$$

$$i_U(\boldsymbol{\mu}, \sigma_{22}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{11}$$

for $\pi = P(r_i = 1)$. In contrast to the naive information these cross-terms do not all vanish, and the orthogonality of mean and variance–covariance parameters is lost under the MAR mechanism. One implication of this is that, although the information relating to the linear model parameters alone is not affected by the move from an MCAR to an MAR mechanism, the asymptotic variance–covariance matrix *is* affected due to the induced nonorthogonality and therefore the dropout mechanism cannot be regarded as ignorable as far as the estimation of precision of the linear model parameters is concerned. It can also be shown that the expected information for the variance–covariance parameters is not equivalent under the MCAR and MAR dropout mechanisms, but the expressions are rather more involved. Assuming that π is nonzero, it can be seen that the necessary and sufficient condition for the terms in (9) and (10) to be equal to zero is that $\alpha_0 = \alpha_1 = 0$, the condition defining, as expected, an MCAR mechanism.

We now illustrate these findings with a few numerical results. The off-diagonal unconditional information elements (9)–(11) are computed for sample size $n = 1{,}000$, mean vector $(0, 0)^T$ and two covariance matrices: (1) $\sigma_{11} = \sigma_{22} = 1$ and correlation $\rho = \sigma_{12} = 0.5$, and (2) $\sigma_{11} = 2$, $\sigma_{33} = 3$ and $\rho = 0.5$ leading to $\sigma_{12} = \sqrt{6}/2$. Further, two MAR dropout mechanisms are considered. They are both of the logistic form

$$P(r_1 = 1 | y_{i1}) = \frac{\exp(\gamma_0 + \gamma_1 y_{i1})}{1 + \exp(\gamma_0 + \gamma_1 y_{i1})}.$$

We choose $\gamma_0 = 0$ and (a) $\gamma_1 = 1$ or (b) $\gamma_1 = -\infty$. The latter mechanism implies $r_i = 1$ if $y_{i1} \geq 0$ and $r_i = 0$ otherwise. Both dropout mechanisms yield $\pi = 0.5$. In all cases $\alpha_1 = -\alpha_0$, with α_1 in the four possible combinations of covariance and dropout parameters: (1a) 0.4132, (1b) 0.7263, (2a) $\sqrt{2/\pi}$, (2b) $2/\sqrt{\pi}$. Numerical values for (9)–(11) are presented in Table 4, as well as the average from the observed information matrices in a simulation with 500 replicates.

Obviously, these elements are far from zero, as would be found with the naive estimator. They are of the same order of magnitude as the upper left block of the information matrix (pertaining to the mean parameters), which are

$$\begin{pmatrix} 1166.67 & -333.33 \\ -333.33 & 666.67 \end{pmatrix}.$$

We performed a limited simulation study to verify the coverage probability for the Wald tests under the unconditional and a selection of conditional frameworks. (See Table 5.) The hypotheses considered are

TABLE 4

Bivariate normal data: computed and simulated values for the off-diagonal block of the unconditional information matrix. Sample size is $n = 1{,}000$ (500 replications). The true model has zero mean vector. Two true covariances Σ and two dropout parameters γ_1 are considered

Parameters			Unconditional $i_U(\mu, \cdot)$			Simulated $\widehat{i_O}(\mu, \cdot)$		
Σ		γ_1	σ_{11}	σ_{12}	σ_{22}	σ_{11}	σ_{12}	σ_{22}
1	0.5	1	−68.87	137.75	0.00	−69.36	137.95	−0.04
0.5	1		137.75	−275.49	0.00	137.88	−276.83	−0.04
2	$\sqrt{6}/2$	1	−30.26	49.42	0.00	−30.21	49.54	0.04
$\sqrt{6}/2$	3		49.42	−80.70	0.00	49.52	−81.31	0.06
1	0.5	$-\infty$	132.98	−265.96	0.00	135.67	−267.66	0.16
0.5	1		−265.96	531.92	0.00	−267.73	537.58	−0.02
2	$\sqrt{6}/2$	$-\infty$	47.02	−76.78	0.00	49.52	−78.73	−0.02
$\sqrt{6}/2$	3		−476.78	125.38	0.00	−78.58	126.91	0.02

TABLE 5

Bivariate normal data: true values are as in the third model of Table 4. Coverage probabilities (×1,000) for Wald test statistics. Sample size is $n = 1{,}000$ (500 replications). The null hypotheses are H_{04}: $\mu_1 = 0$, H_{05}: $\mu_2 = 0$, H_{06}: $\mu_1 = \mu_2 = 0$. For the naive sampling frameworks, m denotes the fixed number of complete cases

Hypothesis	Uncond.	$m = 500$	$m = 450$	$m = 400$
H_{04}	933	996	187	0
H_{05}	953	952	913	830
H_{06}	952	992	338	0

H_{04}: $\mu_1 = 0$, H_{05}: $\mu_2 = 0$ and H_{06}: $\mu_1 = \mu_2 = 0$. The simulations have been restriced to the first covariance matrix used in Table 4 and to the second dropout mechanism ($\gamma_1 = -\infty$). The coverages for the unconditional framework are in good agreement with a χ^2 reference distribution; the first naive framework (500 complete cases) leads to a conservative procedure, whereas the second and the third lead to extreme liberal behavior that is most marked for hypotheses H_{04} and H_{06}. This is to be expected, because by fixing $m = 500$, the proportion of positive first outcomes is constrained to be equal to its predicted value. This has the effect of reducing the variability of $\hat{\mu}_1$. The second and the third frameworks also supress the variability, but introduce bias at the same time. The comparative insensitivity of the behavior of test for H_{05} to the sampling framework is because μ_1 has only an indirect influence through the correlation between the outcomes on both occasions. It should be noted that due to numerical problems, not all simulations led to 500 succesful estimations. On average, 489 convergences were observed, the lowest value being 460 for H_{05} in the first naive sampling frame.

3.3 Sampling with a Stopping Rule

Suppose that n i.i.d. $N(\mu, 1)$ observations $y_1, \ldots, y_n$ are collected and, if the sample fails to satisfy a given stopping rule, a further n observations $y_{n+1}, \ldots, y_{2n}$ are collected, with the same distribution. This represents a very simple form of a group sequential trial (Armitage, 1975). As in the previous examples the final sample size N is a random variable, but in this case taking only one of two values: n or $2n$. The aim is to estimate μ, and the naive approach leads to the estimator $\hat{\mu} = N^{-1}\sum_{i=1}^{N} y_i$ with corresponding information $i_N(\mu) = N$. It is well known that this naive estimator is biased: denoting the probability of stopping by $\pi = P(N = n)$, it can be shown that $E(\hat{\mu}) = \pi(\mu_0 - \mu)/2 + \mu$ for $\mu_0 = E(n^{-1}\sum_{i=1}^{n} Y_i \mid N = n)$.

An alternative estimator that uses all the information is based on the joint distribution of the $\{Y_i\}$ and N. Using the *pattern-mixture* decomposition of the joint distribution $f(y_1, \ldots, y_N \mid N)f(N)$ we get the kernel of the log-likelihood function

$$\ell(\mu) = -\frac{1}{2}\sum_{i=1}^{N}(y_i - \mu)^2 + \left(2 - \frac{N}{n}\right)\ln \pi + \left(\frac{N}{n} - 1\right)\ln(1 - \pi).$$

For a given stopping rule this can be evaluated. Suppose that the rule is to stop if $\sum_{i=1}^{n} Y_i < 0$. Then $\pi = \Phi(-\sqrt{n}\mu)$ and

$$\ell(\mu) = -\frac{1}{2}\sum_{i=1}^{N}(y_i - \mu)^2 + \left(2 - \frac{N}{n}\right)\ln\{\Phi(-\sqrt{n}\mu)\} + \left(\frac{N}{n} - 1\right)\ln\{\Phi(\sqrt{n}\mu)\},$$

with associated score function

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N}(y_i - \mu) + \frac{\phi(\sqrt{n}\mu)}{\sqrt{n}}\left\{\frac{N - n}{\Phi(\sqrt{n}\mu)} - \frac{2n - N}{\Phi(-\sqrt{n}\mu)}\right\}.$$

The observed information is then

$$i_O(\mu) = -\frac{\partial^2 l}{\partial \mu^2}$$
$$= N - \phi'(\sqrt{n}\mu)\left\{\frac{N-n}{\Phi(\sqrt{n}\mu)} - \frac{2n-N}{\Phi(-\sqrt{n}\mu)}\right\}$$
$$+ \phi^2(\sqrt{n}\mu)\left\{\frac{N-n}{\Phi^2(\sqrt{n}\mu)} + \frac{2n-N}{\Phi^2(-\sqrt{n}\mu)}\right\},$$

where $\phi' = \partial\phi/\partial\mu$. The data enter this expresssion only through the sample size N, so to obtain the expected information it is necessary only to take expectations with respect to N. Given $P(N = 2n) = \Phi(\sqrt{n}\mu)$, it follows that $E(N) = n\{1 + \Phi(\sqrt{n}\mu)\}$. Hence

$$i_U(\mu) = n\{1 + \Phi(\sqrt{n}\mu)\} + \frac{n\phi^2(\sqrt{n}\mu)}{\Phi(\sqrt{n}\mu)\Phi(-\sqrt{n}\mu)}.$$

The stopping rule setup explored here is an example of an MAR process in which the parameters of the response and missing value mechanism are *not* separable. As expected we have seen that the naive and unconditional information do not coincide. Although the naive ML estimator remains consistent even when separability does not hold, as pointed out by Diggle (1992), the asymptotic framework required in this setting for this consistency ($n \to \infty$) is of little value from a practical viewpoint.

To illustrate these findings numerically, we considered an experiment with $n = 100$ and $\mu = 0$. This leads to $i_U = 100(1.5 + 2/\pi) = 213.66$. The naive information on the other hand is merely the expectation of the sample size: $i_N = 150$. After 5,000 runs, we found for the naive estimator 149.96, and 212.51 for the unconditional one. The unweighted average of the sample means was -0.0203, while a weighted average (weighted by the sample size to correct for the smaller information contents when early stopping occurred) is -0.0002. A small negative bias is to be expected, because in the experiments where the stopping rule was applied, small outcomes are favored.

4. EXAMPLES

In this section, the points made above are illustrated using three real longitudinal examples, two with a categorical response and one with a continuous response.

The first example is a multicenter study involving 315 patients that were treated by fluvoxamine for psychiatric symptoms described as possibly resulting from a dysregulation of serotonine in the brain. The data are discussed in Molenberghs and Lesaffre (1994). After recruitment to the study, the patient was assessed at four visits. The therapeutic effect and the extent of side effects were scored at each visit on an ordinal scale. The side-effect response is coded as (1) none, (2) not interfering with functionality, (3) interfering significantly with functionality, (4) side-effects surpasse the therapeutic effect. Similarly, the effect of therapy is recorded on a four-point ordinal scale: (1) no improvement or worsening; (2) minimal improvement; (3) moderate improvement, (4) important improvement. Thus, a side effect occurs if new symptoms occur, while there is therapeutic effect if old symptoms disappear. We have 299 patients who have at least one measurement, including 242 completers. An important covariate is previous *duration* of the disease. In a previous analysis of these data evidence was found for an MAR process operating on the side-effects outcome in the sense that there was clear dependence of dropout on previous side-effect measurement, while for therapeutic effect there was little evidence for a dependence of dropout on the previous measurement, even though there was some dependence on the current, possibly unobserved, measurement (Molenberghs, Kenward and Lesaffre, 1997). In conclusion, the mechanism for side effects is at least MAR, whereas an MCAR mechanism can be formulated that is consistent with the therapeutic outcome.

We will first study two dichotomized versions (category 1 versus higher categories 2, 3 and 4; and categories 1 and 2 versus 3 and 4) of side effects and therapeutic effects at the first and the last measurement occasions. The data are shown in Table 6; the analysis is shown in Table 7. The model of Section 3.1 is fitted to all four tables, which is particularly illustrative because naive and unconditional standard error estimates for λ (the success probability at the first occasion) coincide, concentrating potential differences between both estimators in the parameters θ_0 and θ_1 (the conditional success probabilities at the last occasion, given failure or success at the first occasion, respectively). For the first analysis of side effects, there are only small differences and inference at a common significance level is unaffected. This is different in setting 2. Indeed, the naive significance probability for H_0: $\theta_0 = 0.5$ is 0.0319, while the unconditional version is 0.1306. Note that θ_1 is substantially different from θ_0 and, more important, that the missingness probabilities η_1 and η_0 are very different. For therapeutic effect, neither of the two settings leads to differences in standard errors of any importance.

TABLE 6
Psychiatric study: dichotomized outcome at first and last measurement occasions

Setting	Outcome	Dichot.	(0, 0)	(0, 1)	(1, 0)	(1, 1)	(0, ∗)	(1, ∗)
1	Side	1/234	89	13	57	65	26	49
2	Side	12/34	203	5	14	2	48	27
3	Ther.	1/234	11	1	124	88	7	68
4	Ther.	12/34	77	9	119	19	28	47

The analysis considered above is based on a simple Markov-type model. It concentrates the discrepancy between the naive and robust frameworks in the conditional probabilities θ_j ($j = 0, 1$). Marginal models do not have this feature. As an illustration, we analyze side effects at the first, the second and the fourth measurement occasion, on a three-category scale (with original categories 3 and 4 combined). A trivariate odds ratio model (Molenberghs and Lesaffre, 1994) is adopted. Briefly, marginal cumulative logits for each outcome are combined with global marginal log odds ratios for the pairwise and third-order interactions in order to specify the joint distribution. Note that this model falls outside the regular exponential family. Generally, one might expect larger differences between observed and naive expected information for nonexponential family models. The marginal logits are assumed to depend on *duration*, whereas the log odds ratios are assumed constant. Molenberghs, Kenward and Lesaffre (1997) observed that dropout in the side effects outcome depends both on the previous measurement as well as on the value of *duration*. We analyzed the set of 222 complete cases as well as all available data. Table 8 reports on the value of the (naive and unconditional) Wald statistic for a number of hypotheses. Although not spectacular, the differences between naive expected and observed information based tests is larger for the MAR analysis than for the complete case analysis. In particular, the *P*-value for the hypothesis of no *duration* effect (MAR) changes from 0.0049 with the naive expected information to 0.0110 with the observed information. In this example it was seen consistently that in MAR analyses the observed information yielded smaller test statistics than the naive expected information. For completers only analyses, this was not always the case.

In the previous study of moderate size, for which there existed some preliminary evidence for an MAR mechanism in the side effects, differences appeared between inferences based on observed and naive expected information. Woolson and Clarke (1984) analyzed data from the Muscatine Coronary Risk Factor Study, a longitudinal study of coronary risk factors in 4,856 school children (1971–1981). These authors analyzed classifications of the children as obese versus not obese made in 1977, 1979 and 1981. Apart from the outcome, the sex of the child and the age stratum (8, 10, 12 or 14) were recorded. All possible missing value patterns occur. There is no evidence that the missing data mechanism would be more complex than MCAR. We have fitted an odds ratio model, with the logit of each measurement depending on *sex* and *age* (linear trend). Categorization of *age* gave very similar results. Table 9 presents the Wald test statistics: the differences between statistics in the completers' analysis are very small. Although differences are slightly larger in the MAR analyses, there is no qualititative difference in the inference based on these tests.

We will now consider a relatively small example with a continuous response, analyzed in Crépeau, Koziol, Reid and Yuh (1985). Fifty-four rats were divided into five treatment groups corresponding to exposure to increasing doses of halothane (0%, 0.25%, 0.5%, 1% and 2%). The groups were of sizes 11, 10, 11, 11 and 11 rats, respectively. Following an induced heart attack in each rat the blood pressure was recorded on nine unequally spaced occasions. A number of rats died during the course of the experiment, including all rats from group 5 (2%

TABLE 7
Psychiatric study: analysis of the data in Table 6. Parameter estimates (naive standard errors; unconditional standard errors) are shown

Par.	Side 1/234	Side 12/34	Ther. 1/234	Ther. 12/34
λ	0.572(0.029; 0.029)	0.144(0.020; 0.020)	0.937(0.014; 0.014)	0.619(0.028; 0.028)
θ_1	0.533(0.044; 0.045)	0.125(0.058; 0.083)	0.415(0.034; 0.034)	0.138(0.029; 0.029)
θ_0	0.128(0.034; 0.033)	0.024(0.011; 0.011)	0.083(0.073; 0.080)	0.105(0.033; 0.033)
η_1	0.714	0.372	0.757	0.746
η_0	0.797	0.813	0.632	0.754

TABLE 8
Psychiatric study: *side effects at times* 1, 2 *and* 4. *Wald test statistics for the completers only and for an MAR analysis*

		Compl. cases		MAR	
Hypothesis	**df**	**Expect.**	**Obs.**	**Expect.**	**Obs.**
Common *duration* effect	2	1.36	1.19	2.54	2.44
No *duration* effect	3	2.98	2.54	12.90	11.13
Common two-way association	2	10.70	9.99	11.48	9.13
Intercepts equal across times	4	28.73	28.83	34.96	33.44
Common difference between intercepts	2	0.16	0.16	2.07	1.48
Linear trend in first intercept	1	0.0099	0.0099	0.16	0.18
Linear trend in second intercept	1	0.020	0.018	1.15	0.85
Linear trend in both intercepts	2	0.034	0.033	1.18	0.91

halothane). Following the original authors we omit this group from the analysis, leaving 43 rats, of which 23 survived the experiment. Examination of the data from these four groups does not provide any evidence of an MAR dropout process, although this observation must be considered in the light of the small sample size. A Gaussian multivariate linear model with an unconstrained covariance matrix was fitted to the data. There was very little evidence of a treament-by-time interaction and the following results are based on the use of a model with additive effects for treatment and time. The Wald statistics for the treatment main effect on three degrees of freedom are equal to 46.95 and 30.82, respectively using the naive expected and observed information matrices. Although leading to the same qualitative conclusions the figures are notably discrepant. A first reaction may be to attribute this difference to the incompleteness of the data. However, the lack of evidence for an MAR process together with the relatively small sample size points to another cause. The equivalent analysis of the 24 completers produces Wald statistics of 45.34 and 26.35, respectively. That is, the effect can be attributed to a combination of small sample variation and possible model misspecification. A theoretical reason for this difference might be that the expected value of the off-diagonal block of the information matrix of the maximum likelihood estimates (describing covariance between mean and covariance parameters) has expectation zero but is likely to depart from this in small samples. As a consequence, the variances of the estimated treatment effects will be higher when derived from the observed information, thereby reducing the Wald statistic. To summarize, this example provides an illustration of an alternative source of discrepancy between the naive expected and observed information matrices, which is likely to be associated with the use, in smaller samples, of covariance matrices with many parameters.

5. CONCLUDING REMARKS

The literature overview in the Introduction indicates an early awareness of problems with conventional likelihood based frequentist inference in the MAR setting. Specifically, several authors point to the use of the observed information matrix as a way to circumvent issues with the expected information matrix. In spite of this, it seems that a broad awareness of this problem has diminished while the number of methods formulated to deal with the MAR situation has risen dramatically in recent years. We therefore feel that a restatement and exposition of this important problem is timely. Three easily accessible and simply formulated settings have been used to illuminate the issues while a number of real examples have been used to explore the implications in practice. The different status of the observed information and the conventional expected informa-

TABLE 9
Muscatine Coronary Risk Factor Study: *Wald test statistics for the completers only and for an MAR analysis*

		Compl. cases		MAR	
Hypothesis	**df**	**Expect.**	**Obs.**	**Expect.**	**Obs.**
Common *sex* effect	2	5.55	5.54	1.50	1.49
No *sex* effect	3	5.56	5.55	6.84	6.82
Common *age* effect	2	22.20	21.37	40.03	38.59
No *age* effect	3	22.40	21.66	46.17	45.39
Common two-way association	2	15.99	16.08	17.26	16.63
Common intercept across time	2	22.27	21.72	45.55	45.56
Linear trend in intercepts	1	0.10	0.10	2.16	2.09

tion (called the naive information in this work) has been clearly shown by contrasting both with the expected information, where the expectation takes the missingness pattern into account (referred to as the unconditional information).

We can conclude from this that, provided the observed information matrix is used, conventional likelihood based frequentist inference is applicable in the MAR setting.

REFERENCES

ARMITAGE, P. (1975). *Sequential Medical Trials*. Blackwell, Oxford.

BAKER, S. G. (1992). A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *J. Comput. Graph. Statist.* **1** 63–76.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

CRÉPEAU, H., KOZIOL, J., REID, N. and YUH, Y. S. (1985). Analysis of incomplete multivariate data from repeated measurements experiments. *Biometrics* **41** 505–514.

DIGGLE, P. J. (1992). On informative and random dropouts in longitudinal studies. Letter to the Editor. *Biometrics* **48** 947.

DIGGLE, P. J. (1993). Estimation with missing data. Reply to a Letter to the Editor. *Biometrics* **49** 580.

EDWARDS, A. W. F. (1972). *Likelihood.* Cambridge Univ. Press.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65** 457–487.

FITZMAURICE, G. M., LAIRD, N. M. and LIPSITZ, S. R. (1994). Analysing incomplete longitudinal binary responses: A likelihood-based approach. *Biometrics* **50** 601–612.

HEITJAN, D. F. (1993). Estimation with missing data. Letter to the Editor. *Biometrics* **49** 580.

HEITJAN, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81** 701–708.

JENNRICH, R. I. and SCHLUCHTER, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42** 805–820.

KENWARD, M. G., LESAFFRE, E. and MOLENBERGHS, G. (1994). An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50** 945–953.

LAIRD, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine* **7** 305–315.

LITTLE, R. J. A. (1976). Inference about means for incomplete multivariate data. *Biometrika* **63** 593–604.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233.

MEILIJSON, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. Roy. Statist. Soc. Ser. B* **51** 127–138.

MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: the SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.

MOLENBERGHS, G., KENWARD, M. G. and LESAFFRE, E. (1997). The analysis of longitudinal ordinal data with informative dropout. *Biometrika* **84** 33–44.

MOLENBERGHS G. and LESAFFRE, E. (1994). Marginal modeling of correlated ordinal data using an n-way Placket distribution. *J. Amer. Statist. Assoc.* **89** 633–644.

MURRAY, G. D. and FINDLAY, J. G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine* **7** 941–946.

PATEL, H. I. (1991). Analysis of incomplete data from clinical trials with repeated measurements. *Biometrika* **78** 609–619.

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

WELSH, A. H. (1996). *Aspects of Statistical Inference.* Wiley, New York.

WOOLSON, R. F. and CLARKE, W. R. (1984). Analysis of categorical incomplete longitudinal data. *J. Roy. Statist. Soc. Ser. A* **147** 87–99.