Statistica Sinica **21** (2011), 187-206

PSEUDO-LIKELIHOOD ESTIMATION FOR INCOMPLETE DATA

Geert Molenberghs[1,2], Michael G. Kenward[3], Geert Verbeke[2,1] and Teshome Birhanu[1]

[1] *Universiteit Hasselt,* [2] *Katholieke Universiteit Leuven and* [3] *London School of Hygiene and Tropical Medicine*

Abstract: In statistical practice, incomplete measurement sequences are the rule rather than the exception. Fortunately, in a large variety of settings, the stochastic mechanism governing the incompleteness can be ignored without hampering inferences about the measurement process. While ignorability only requires the relatively general missing at random assumption for likelihood and Bayesian inferences, this result cannot be invoked when non-likelihood methods are used. A direct consequence of this is that a popular non-likelihood-based method, such as generalized estimating equations, needs to be adapted towards a weighted version or doubly-robust version when a missing at random process operates. So far, no such modification has been devised for pseudo-likelihood based strategies. We propose a suite of corrections to the standard form of pseudo-likelihood to ensure its validity under missingness at random. Our corrections follow both single and double robustness ideas, and is relatively simple to apply. When missingness is in the form of dropout in longitudinal data or incomplete clusters, such a structure can be exploited toward further corrections. The proposed method is applied to data from a clinical trial in onychomycosis and a developmental toxicity study.

Key words and phrases: Double robustness, frequentist inference, generalized estimating equations, ignorability, inverse probability weighting, likelihood, missing at random, missing completely at random, pseudo-likelihood.

1. Introduction

The applied statistician often encounters correlated outcome data. Common situations include multivariate, clustered, and longitudinal data. Frequently in such settings not all of the planned measurements of subject i's outcome vector $\boldsymbol{y}_i$ are actually observed, turning the statistical analysis into a missing-data problem. For example, in a longitudinal study, a subject's response vector may terminate early for a number of reasons outside the control of the investigator. It is almost always necessary to reflect on the nature of the missingness process and its impact on inferences.

When referring to the missing-value, or non-response, process we use terminology of Little and Rubin (2002, Chap. 6). A non-response process is said to be

missing completely at random (MCAR) if 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 MCAR nor MAR is termed *non-random* (MNAR).

Early work on missing values was largely concerned with the practical consequence of missing-data induced imbalance (Little and Rubin (2002), Molenberghs and Kenward (2007), and Kenward and Molenberghs (2009)). Over the last three decades, a number of developments have taken place, allowing the use of MAR based methods. These include multiple-imputation strategies (Rubin (1987)), and so-called direct-likelihood or direct Bayesian analysis. These rest on *ignorability*, the property ensuring that such analyses are valid under MAR, supplemented with mild regularity conditions, even without explicitly modeling the missing-data mechanism, provided that all incomplete sequences are subjected to analysis (Rubin (1976), Little and Rubin (2002), Molenberghs and Kenward (2007), Fitzmaurice et al. (2009)). The practical implication for likelihood inference is that, as soon as a module is available to handle measurement sequences of unequal length, valid inferences are obtained *without any additional work*. Thanks to the availability of flexible software, linear and generalized linear mixed models can be fitted to incomplete sets of data.

For non-Gaussian outcomes, apart from random-effects models, non-likelihood models have also become popular (Molenberghs and Verbeke (2005)). Because these models specify, in principle, the full likelihood, they can be used to analyze incomplete data as well, under MAR assumptions and making use of the ignorability property. However, marginal models for non-Gaussian data imply complex and hard to manipulate likelihoods. In many practical settings involving outcome sequences of moderate to large length, direct likelihood may be prohibitive. Some authors have voiced concern over these models' vulnerability to mis-specification. In response, a number of alternatives have been formulated, the most popular one undoubtedly being *generalized estimating equations* (GEE; Liang and Zeger (1986), Diggle et al. (2002), Molenberghs and Verbeke (2005)). By transforming the score equations into estimating equations, this method essentially allows one to confine attention to the specification of the first moments of the outcome sequence only (i.e., the mean structure), thereby circumventing the need to address the association structure while still leading to valid inferences. A more detailed review of GEE is provided in Section 3. A number of variations to this theme exist, such as GEE2 (also specifying the second moments; Liang, Zeger, and Qaqish (1992) and alternating logistic regressions (Carey, Zeger, and Diggle (1993)). When data are incomplete, GEE suffers from its frequentist nature and is in its basic form valid only under MCAR. Therefore, Robins, Rotnitzky, and Zhao (1995) have developed so-called *weighted* generalized estimating

equations (WGEE), as well as a number of refinements and extensions in subsequent papers, to allow usage of GEE under not only MAR, but even under MNAR settings. The method rests on Horvitz-Thompson ideas (Cochran (1977)), weighing contributions by the inverse probability of being observed. The method is elegant and enjoys good properties, but requires specification of a model for the weights. More recently, these WGEE have been extended toward so-called doubly robust estimating equations, where the weighting idea is supplemented with the use of a predictive model for the unobserved responses, given the observed ones. Excellent reviews can be found in Scharfstein, Rotnitzky and Robins (1999) and, more recently, Van der Laan and Robins (2003), Bang and Robins (2005), and Rotnitzky (2009).

Next to GEE, pseudo-likelihood methods (PL; le Cessie and van Houwelingen (1991), Geys, Molenberghs and Lipsitz (1998), Geys, Molenberghs and Ryan (1999), Aerts et al. (2002)) have become popular as an alternative to full likelihood, and therefore also to GEE and GEE2. Rather than replacing the score equations with alternative functions, the likelihood itself is replaced by a more tractable function. A detailed discussion is given in Section 3.3. In so-called marginal pseudo-likelihood, the likelihood for an n_i-dimensional response vector is replaced by the product of all pairs, or all triples, or all p-tuples (with p a prespecified number, corresponding to the highest order of association that is still of scientific interest) of outcomes. Computational and statistical performance (e.g., efficiency) have been shown to range from acceptably good to excellent. Evidently, conditional versions of pseudo-likelihood are also possible, where the contributions take the form of conditional densities of a subset of outcomes within a sequence, given another subset of outcomes.

Because pseudo-likelihood is not a full likelihood method, there is no *a priori* guarantee that the method would be valid under MAR. (Note that Rubin (1976) provided sufficient conditions only, without claiming their necessity.) In this paper, we show that a correction is necessary to allow for the use of pseudo-likelihood, and that both singly robust as well as doubly robust versions of PL can be considered, including from a practical standpoint.

Of course, whatever MAR developments are made, one can never exclude the operation of an MNAR mechanism. A number of modeling strategies have been proposed, but at the same time it has been reported that such strategies are very sensitive to unverifiable modeling assumptions. A number of sensitivity analysis tools have been proposed, but because this paper is devoted to ignorability in the PL context, MNAR and the surrounding issues are considered to be outside of its scope. For reviews on the sensitivity issues, see Molenberghs and Kenward (2007) and Fitzmaurice et al. (2009).

General concepts are formally introduced in Section 2. The likelihood, GEE, and pseudo-likelihood inferential paradigms are sketched in Section 3. Pseudo-likelihood approaches to incomplete data are presented in Section 4. The proposed methods are applied to two case studies in Section 5.

2. Concepts

Let the random variable Y_{ij} denote the response for the ith study subject at the jth occasion ($i = 1, \ldots, N$, $j = 1, \ldots, n_i$). Independence across subjects is assumed. We group the outcomes into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ and define a further vector of missingness indicators $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})'$ with $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise. In the specific case of dropout in longitudinal studies, $\boldsymbol{R}_i$ can be replaced by the dropout indicator $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$. The vector $\boldsymbol{Y}_i$ is divided into observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components, respectively.

As stated in the introduction, one needs to consider the density of the full data $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\beta}, \boldsymbol{\psi})$, where the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ describe the measurement and missingness (non-response) processes, respectively. When appropriate, we introduce separate association parameters $\boldsymbol{\alpha}$. Covariates are assumed to be measured and grouped into $\boldsymbol{x}_i$ although, for notational simplicity, this is sometimes dropped. The full density function can be factored as (Rubin (1976), Little and Rubin (2002)):

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\beta}) f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\psi}). \tag{2.1}$$

The first factor is the marginal density of the measurement process, and the second one is the density of the missingness process, conditional on the outcomes.

3. Inferential Frameworks

3.1. Likelihood

When data are incomplete, subject i's likelihood contribution is

$$L_i = \int f(\boldsymbol{y}_i|\boldsymbol{\beta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})\, d\boldsymbol{y}_i^m. \tag{3.1}$$

In general, (3.1) does not simplify but, under MAR, we obtain $L_i = f(\boldsymbol{y}_i^o|\boldsymbol{\beta})$ $f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi})$. Hence, likelihood and Bayesian inferences for the measurement model parameters $\boldsymbol{\beta}$ can be made without explicitly formulating the missing-data mechanism, provided that some mild regularity conditions hold (Rubin (1976)). It is precisely this result that makes so-called direct likelihood analyses, valid under MAR, appealing in a variety of settings (Molenberghs et al. (2004)).

3.2. Generalized estimating equations

A detailed account in given in the Appendix. When inferences focus on population averages, one can directly model all of the marginal expectations $E(Y_{ij}) = \mu_{ij}$ in terms of covariates of interest, through $h(\mu_{ij}) = \boldsymbol{x'_{ij}\beta}$, with $h(\cdot)$ some known link function. The marginal variance depends on the marginal mean according to $\text{Var}(Y_{ij}) = v(\mu_{ij})\phi$, where $v(\cdot)$ is a known variance function and ϕ is a scale parameter. The correlation between Y_{ij} and Y_{ik} is expressed via a correlation matrix $R_i(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is a vector of nuisance parameters; the covariance matrix $V_i = V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \phi A_i^{1/2} R_i A_i^{1/2}$, with A_i the diagonal matrix of marginal variances. Generalized estimating equations take the form (Liang and Zeger (1986))

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta'}} V_i^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu_i}) \;=\; \boldsymbol{0}. \tag{3.2}$$

The variance obtains from the information sandwich

$$\text{Var}\,(\widehat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}, \tag{3.3}$$

where

$$I_0 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta'}}, \qquad I_1 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\beta}} V_i^{-1} \text{Var}(\boldsymbol{y_i}) V_i^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta'}}. \tag{3.4}$$

As stated earlier, GEE is not likelihood based and therefore ignorability (Rubin (1976)) cannot be invoked to establish the method's validity under MAR. Thus, apart from special cases, GEE in its basic form is valid only under MCAR.

In response to this, Robins, Rotnitzky, and Zhao (1995) proposed a class of so-called *weighted* estimating equations. The idea is to weight each subject's contribution to the GEE by the inverse probability, either of being fully observed, or of being observed up to a certain time. Let π_i be the probability for subject i to be completely observed, and π_i' the probability for subject i to drop out on occasion d_i. These can be written as

$$\pi_i = \prod_{\ell=2}^{n_i} (1 - p_{i\ell}), \qquad \pi_i' = \left[\prod_{\ell=2}^{d_i - 1} (1 - p_{i\ell}) \right] \cdot p_{id_i}, \tag{3.5}$$

where $p_{i\ell} = P\left(D_i = \ell | D_i \geq \ell, Y_{i\,\overline{\ell}}, X_{i\,\overline{\ell}}\right)$ are the component probabilities of dropping out at occasion ℓ, given the subject is still in the study, the covariate history $X_{i\,\overline{\ell}}$, and the outcome history $Y_{i\,\overline{\ell}}$. In such a case, one can opt either for WGEE based on the completers only:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta'}} V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \;=\; \boldsymbol{0}, \tag{3.6}$$

with $\widetilde{R}_i = 1$ if a subject is fully observed and 0 otherwise, or, upon using (3.5), for WGEE using all subjects:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{1}{\pi_i'} \frac{\partial \boldsymbol{\mu}_i^o}{\partial \boldsymbol{\beta}'} (V_i^o)^{-1} (\boldsymbol{y}_i^o - \boldsymbol{\mu}_i^o) \;=\; \mathbf{0}. \tag{3.7}$$

Here the superscript 'o' indicates the portion corresponding to the observed data in the corresponding matrix or vector. Of course, with (3.6), the incomplete subjects also contribute through the model for the dropout probabilities π_i.

As stated earlier, (3.6) has been extended to so-called double robustness (Scharfstein, Rotnitzky and Robins (1999), Van der Laan and Robins (2003), Bang and Robins (2005)). We focus on longitudinal data with monotone missingness on the one hand and on incomplete clustered data on the other, each time under MAR. Double robustness is taken up in Section 4.1.

3.3. Pseudo-likelihood

Using Arnold and Strauss (1991), we introduce pseudo-likelihood, the principal idea of which is to replace a numerically challenging joint density by a simpler function assembled from suitable factors.

3.3.1. Definition and properties

Let S be the set of all $2^n - 1$ vectors of length n consisting solely of zeros and ones, with each vector having at least one non-zero entry. Denote by $\boldsymbol{y}_i^{(s)}$ the subvector of $\boldsymbol{y}_i$ corresponding to the components of s that are non-zero. The associated joint density is $f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\beta})$. To define a pseudo-likelihood function, one chooses a set $\delta = \{\delta_s | s \in S\}$ of real numbers with at least one non-zero component. The log of the pseudo-likelihood is then

$$p\ell = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \ln f_s(\boldsymbol{y}_i^{(s)}; \boldsymbol{\beta}). \tag{3.8}$$

Adequate regularity conditions have to be invoked to ensure that (3.8) can be maximized by solving the pseudo-likelihood (score) equations, the latter obtained by differentiating the logarithmic pseudo-likelihood and equating its derivative to zero. These regularity conditions are spelled out in the Appendix. In particular, when the components in (3.8) result from a combination of marginal and conditional distributions of the original distribution, then a valid pseudo-likelihood function results. In particular, the classical log-likelihood function is found by setting $\delta_s = 1$ if s is the vector consisting solely of ones, and 0 otherwise. More details can be found in Varin (2008), Lindsay (1988), and Joe and Lee (2008). Note that Joe and Lee (2008) use weighting for reasons of efficiency in pairwise

likelihood, similar in spirit to Geys, Molenberghs and Lipsitz (1998), but differently from its use here, which focuses on bias correction when data are incomplete. Another important reference is Cox and Reid (2004).

Be θ_0 the true parameter. Under suitable regularity conditions (see also Arnold and Strauss (1991) Geys, Molenberghs and Ryan (1999) Aerts et al. (2002)), it can be shown that maximizing (3.8) produces a consistent and asymptotically normal estimator $\widetilde{\boldsymbol{\beta}}_0$ so that $\sqrt{N}(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0)$ converges in distribution to

$$N_p[\mathbf{0}, I_0(\boldsymbol{\beta}_0)^{-1} I_1(\boldsymbol{\beta}_0) I_0(\boldsymbol{\beta}_0)^{-1}]. \tag{3.9}$$

Precise statements and additional discussion are given in Supplementary Materials B and C.

4. Pseudo-likelihood For Incomplete Data

In the above, it is assumed that data are completely observed according to the study design. Otherwise, PL is valid under the assumption of an MCAR mechanism operating, but this does not generally extend to MAR mechanisms, excepting a limited number of special cases, such as full exchangeability, as is shown in Section 4.3. The reason for this is twofold. First, in line with Kenward and Molenberghs (1998), even likelihood methods commonly have frequentist elements, such as the expected information matrix. Second, because pseudo-likelihood is not a genuine likelihood but rather a modification of it, it no longer enjoys the results derived for the likelihood by Rubin (1976). This second issue is shared with GEE.

Unlike for GEE, little work has been done for PL estimation with incomplete data. A noteworthy exception is Parzen et al. (2006) who apply PL ideas not just to the vector of outcomes, but to the entire vector of outcomes, covariates, and missing-data indicators. In what follows, we follow a different route, using inverse probability weighting and double robustness ideas (Scharfstein, Rotnitzky and Robins (1999), Van der Laan and Robins (2003), Bang and Robins (2005), and Rotnitzky (2009)).

We first present general expressions and establish their validity, and then apply them to pseudo-likelihood. This implies that they hold, beyond pseudo-likelihood, for a large class of estimating equations, in line with the work of Robins, Rotnitzky, and colleagues. Thereafter, we pay particular attention to two special PL families: (1) marginal and (2) full conditional. In the first case, the multivariate normal and models for binary data are considered in more detail. In the second case, an exponential family model for binary clustered data is scrutinized further.

While our results are general, their implementation for general missing-data patterns is more complicated than when missingness is confined to dropout, or arises in a clustered-data setting.

4.1. General forms of estimating equations for incomplete data

Assume that we have a set of estimating equations, whether resulting from full likelihood or pseudo-likelihood, of a conventional generalized estimating equations type, or beyond:

$$\boldsymbol{U} = \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\beta}|\boldsymbol{Y}_i, \boldsymbol{x}_i) \overset{\text{notation}}{=} \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{Y}_i). \tag{4.1}$$

Assume that $E(\boldsymbol{U}) = \boldsymbol{0}$.

First consider two obvious 'naive' estimating equations that originate from (4.1):

$$\boldsymbol{U}_{\text{naive, CC}} = \sum_{i=1}^{N} \widetilde{R}_i \boldsymbol{U}_i(\boldsymbol{Y}_i), \tag{4.2}$$

$$\boldsymbol{U}_{\text{naive, AC}} = \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{Y}_i^o). \tag{4.3}$$

Here, $\widetilde{R}_i = 1$ if subject i is fully observed and 0 otherwise, and $\boldsymbol{U}_i(\boldsymbol{Y}_i^o)$ is the score pertaining the the observed outcomes on subject i. Further, 'naive' refers to the fact that these estimating equations would generally be biased under MAR; 'CC' denotes complete cases, i.e., subjects with all measurements taken, and 'AC' stands for available cases. For the latter, it is necessary to derive the score contribution of the sub-vector of observed components of $\boldsymbol{Y}_i$. Because this involves integration over the incomplete data, it is trivial in the marginal case but less so, for example, for conditionally specified PL functions.

Singly robust versions of (4.2) and (4.3) are:

$$\boldsymbol{U}_{\text{IPWCC}} = \sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \boldsymbol{U}_i(\boldsymbol{Y}_i), \tag{4.4}$$

$$\boldsymbol{U}_{\text{IPWAC}} = \sum_{i=1}^{N} \frac{1}{\pi_i'} \cdot E_{Y^m|y^o} \boldsymbol{U}_i(\boldsymbol{Y}_i), \tag{4.5}$$

$$\boldsymbol{U}_{\text{IPWAC,seq}} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{R_{ij}}{\pi_{ij}} \cdot U_i(Y_{ij}|\boldsymbol{Y}_{i\bar{j}}). \tag{4.6}$$

Here R_{ij} is the indicator for a subject to be observed at occasion j, and π_{ij} is the probability of being observed up to and including occasion j, $\pi_{ij} = \prod_{\ell=1}^{j}(1-p_{i\ell})$. Further, $\boldsymbol{Y}_{i\bar{j}}$ is shorthand for the history $(Y_{i1}, \ldots, Y_{i,j-1})$, and the corresponding function $U_i(Y_{ij}|\boldsymbol{Y}_{i\bar{j}})$ is the score for the outcome at occasion j given the history. Recall that π_i and π_i' have been defined in (3.5). Doubly robust versions are

$$\boldsymbol{U}_{\text{IPWCC,dr}} = \sum_{i=1}^{N}\left[\frac{\widetilde{R}_i}{\pi_i}U_i(\boldsymbol{Y}_i) + \left(1-\frac{\widetilde{R}_i}{\pi_i}\right)E_{Y^m|y^o}\boldsymbol{U}_i(\boldsymbol{Y}_i)\right], \tag{4.7}$$

$$\begin{aligned}\boldsymbol{U}_{\text{IPWAC,dr}} = \sum_{i=1}^{N}\Bigg\{\sum_{j=1}^{n_i}\Bigg[&\frac{R_{ij}}{\pi_{ij}}\cdot\boldsymbol{U}_i(Y_{ij}|\boldsymbol{Y}_{i\bar{j}}) \\ &+\left(1-\frac{R_{ij}}{\pi_{ij}}\right)\cdot E_{Y^m|y^o}\boldsymbol{U}_i(Y_{ij}|\boldsymbol{Y}_{i\bar{j}})\Bigg]\Bigg\}.\end{aligned} \tag{4.8}$$

Here $\boldsymbol{U}_i(Y_{ij}|\boldsymbol{Y}_{i\bar{j}})$ is the score pertaining to outcome Y_{ij} given the history, denoted by $\boldsymbol{Y}_{i\bar{j}}$. It is easy to show single and double robustness for the above definitions. Details can be found in Supplementary Materials D and E.

The predictive model may show varying degrees of complexity, depending on the type of PL function considered. For example, marginal models for continuous data, marginal models for binary data, and conditional models for binary data, may all pose specific challenges. This means that, in some settings, the predictive model might be of higher dimension than the components of the actual PL function and/or contain components of the full likelihood that are not needed for it. While this may seem to defeat the purpose of using PL methodology, there are several practically useful strategies to handle this. To see this, it helps to distinguish between two uses of the likelihood within the framework: estimation and prediction. It is predominantly for estimation that PL leads to important economies by not having to manipulate the full likelihood; for prediction, several alternative strategies are available.

First, even though using the entire joint distribution is often prohibitive for estimation, it may be tractable for prediction purposes if all the necessary parameters are obtained from the likelihood. An example is provided by a full conditional PL, with a counterexample being a purely marginal PL for binary data, consisting of lower-order margins only. Second, a sufficiently rich predictive model could be used, such as logistic regression for example. Evidently, such a predictive model would, strictly speaking, be incompatible with the actual model under consideration but, as Bang and Robins (2005) point out, virtually all parametric models are mis-specified to some extent. In this sense, a reasonable predictive model, coupled with a sensible missingness model for the weights, often considerably increases efficiency and reduces bias. Bang and Robins' simulation results were encouraging in this respect. In a similar vein, with multiple

imputation, Meng (1994) shows that so-called *uncongenial imputation models* can still lead to inferences that are practically acceptable. We return to this in Section 4.3.

4.2. Precision estimation

General expressions for the precision of the estimates obtained using generalized estimating equations and pseudo-likelihood are given in (3.3) and (3.9), respectively.

When the single or doubly robust versions of the previous sections are used, with a parametric model for dropout, then the uncertainty induced by estimation of the $\boldsymbol{\psi}$ parameters needs to be accommodated. As shorthand for any of the forms (4.4), (4.5), (4.7), and (4.8), we write $\boldsymbol{U} = \sum_{i=1}^{N} \boldsymbol{V}_i(\boldsymbol{\beta})$, and the parameters $\boldsymbol{\psi}$ are estimated from score or estimating equations $\boldsymbol{W} = \sum_{i=1}^{N} \boldsymbol{W}_i(\boldsymbol{\psi})$. The entire score for subject i is $\boldsymbol{S}_i = (\boldsymbol{V}_i', \boldsymbol{W}_i')'$. The asymptotic variance-covariance matrix can then be consistently estimated by $\widehat{I}_0^{-1}\widehat{I}_1\widehat{I}_0^{-1}$, with

$$I_0 = \sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} \\ 0 & \frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} \end{pmatrix}, \tag{4.9}$$

$$I_1 = \sum_{i=1}^{N} \boldsymbol{S}_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}) \boldsymbol{S}_i'(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}). \tag{4.10}$$

See also Bang and Robins (2005) and Rotnitzky (2009).

4.3. The case of pseudo-likelihood

In the previous section, we focused on estimating equations in the broadest sense. When we turn to pseudo-likelihood, the generic forms can be made more specific and expanded further:

$$\boldsymbol{U}_{\text{naive, CC}} = \sum_{i=1}^{N} R_i \sum_{s \in S} \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}), \tag{4.11}$$

$$\boldsymbol{U}_{\text{naive, CS}} = \sum_{i=1}^{N} \sum_{s \in S} R_{i,s} \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}), \tag{4.12}$$

$$\boldsymbol{U}_{\text{naive, AC}} = \sum_{i=1}^{N} \sum_{s \in S} \delta_s E_{Y^m|y^o} \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}), \tag{4.13}$$

$$\boldsymbol{U}_{\text{IPWCC}} = \sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \cdot \sum_{s \in S} \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}), \tag{4.14}$$

$$\boldsymbol{U}_{\text{IPWCS}} = \sum_{i=1}^{N} \sum_{s \in S} \frac{R_{i,s}}{\pi_{i,s}} \cdot \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}), \tag{4.15}$$

$$\boldsymbol{U}_{\text{IPWAC}} = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \sum_{j=1}^{n_i} I(j \in s) \cdot \frac{R_{ij}}{\pi_{ij}} \cdot \boldsymbol{U}_s(Y_{ij} | \boldsymbol{Y}_{i\bar{j}}^{(s)}), \tag{4.16}$$

$$\begin{aligned} \boldsymbol{U}_{\text{IPWCC,dr}} = \sum_{i=1}^{N} \Bigg\{ & \frac{\widetilde{R}_i}{\pi_i} \left[\sum_{s \in S} \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}) \right] \\ & + \left(1 - \frac{\widetilde{R}_i}{\pi_i} \right) \cdot E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o} \left[\sum_{s \in S} \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}) \right] \Bigg\}, \end{aligned} \tag{4.17}$$

$$\begin{aligned} \boldsymbol{U}_{\text{IPWCS,dr}} = \sum_{i=1}^{N} \sum_{s \in S} \Bigg\{ & \frac{R_{i,s}}{\pi_{i,s}} \cdot \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)o}) \\ & + \left(1 - \frac{R_{i,s}}{\pi_{i,s}} \right) \cdot \delta_s E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o} \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}) \Bigg\}, \end{aligned} \tag{4.18}$$

$$\begin{aligned} \boldsymbol{U}_{\text{IPWAC,dr}} = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \sum_{j=1}^{n_i} I(j \in s) \Bigg[& \frac{R_{ij}}{\pi_{ij}} \cdot \boldsymbol{U}_s(Y_{ij} | \boldsymbol{Y}_{i\bar{j}}^{(s)}) \\ & + \left(1 - \frac{R_{ij}}{\pi_{ij}} \right) \cdot E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o} \boldsymbol{U}_s(Y_{ij} | \boldsymbol{Y}_{i\bar{j}}^{(s)}) \Bigg], \end{aligned} \tag{4.19}$$

where R_i, π_i, R_{ij}, and π_{ij} retain their former meaning. Similarly, $R_{i,s}$ and $\pi_{i,s}$ are, respectively, the indicator for, and the probability of, observing the corresponding sub-vector $\boldsymbol{Y}_i^{(s)}$ of $\boldsymbol{Y}_i$. Further, 'CS' stands for 'complete sets'.

When the outcome sequence is fully exchangeable, in the sense that the distribution of any sub-vector of $\boldsymbol{Y}_i$ is that of any other sub-vector of equal length or a permutation thereof, then $\boldsymbol{U}_{\text{IPWCS,dr}}$ simplifies considerably:

$$E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o} \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)}) = E_{\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o} \left[\boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)o}) + \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)m} | \boldsymbol{Y}_i^{(s)o}) \right].$$

Now, the expectation over the second term on the right side can be replaced by $E_{\boldsymbol{Y}_i^{(s)m} | \boldsymbol{Y}_i^{(s)o}} \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)m} | \boldsymbol{Y}_i^{(s)o})$, due to full exchangeability and the fact that the score contributions arise from derivatives of sub-vectors of $\boldsymbol{Y}_i$. Upon this replacement, the conditional expectation vanishes. As a consequence, under exchangeability, there is no need to explicitly model the missing data mechanism. Hence, (4.18) reduces to

$$\boldsymbol{U}_{\text{IPW, exch}} = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \boldsymbol{U}_s(\boldsymbol{Y}_i^{(s)o}). \tag{4.20}$$

Thus, in this special but important case, neither the weights nor the conditional expectation are necessary to obtain valid inferences.

We now focus on special cases of pseudo-likelihood. First, we consider the case of pairwise pseudo-likelihood, and then apply this to normally distributed and binary data. Second, we consider a conditional pseudo-likelihood for binary outcomes.

Specific details for pairwise and full conditional pseudo-likelihood are given in Supplementary Material G.

5. Case Studies

With our case studies we aim to illustrate both marginal and conditional pseudo-likelihood, and similarities and differences between the various types of pseudo-likelihood.

5.1. A clinical trial in onychomycosis

The data come from a randomized, double-blind, parallel group study for the comparison of two oral treatments (coded as A and B) for toenail dermatophyte onychomycosis (TDO), described in full detail by De Backer et al. (1996). See also Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005). The aim was to compare the efficacy and safety of 12 weeks of continuous therapy with treatment A, or with treatment B. We consider data from 146 patients in arm A and 148 in arm B, for whom the big toenail was the target nail. Patients were assessed at 0, 1, 2, 3, 6, 9, and 12 months. The response is the unaffected nail length, measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in mm.

The design and data type of this study were sufficiently simple to allow for full likelihood, providing a basis for comparison with the proposed pseudo-likelihood methods. We used several forms of pairwise marginal likelihood, as described in Section G.1, in particular with the multivariate normal versions as in Section G2.

For the unaffected nail length Y_{ij}, measured at time occasion j for patient i, we specified a linear mixed-effects model:

$$\begin{aligned} Y_{ij}|b_i \sim N[&b_i + \beta_0 \cdot I(T_i = 0) + \beta_1 \cdot I(T_i = 1) + \beta_2 t_j \cdot I(T_i = 0) \\ &+\beta_3 t_j \cdot I(T_i = 1), \sigma^2], \\ b_i \sim N(&0, \tau^2), \end{aligned} \tag{5.1}$$

where $T_i = 0$ if patient i received standard treatment and 1 for experimental therapy ($i = 1, \ldots, 298$). Further, t_j is the time at which the jth measurement

Table 1. Toenail Data. (Unaffected nail length outcome). Parameter estimates (purely model-based standard errors; empirically corrected standard errors) for full likelihood, and naive, singly robust, and doubly robust pairwise likelihood.

Effect	Par.	$\boldsymbol{U}_{\text{full.lik.}}$	$\boldsymbol{U}_{\text{naive, CC}}$	$\boldsymbol{U}_{\text{naive, CP}}$	$\boldsymbol{U}_{\text{naive, AC}}$
Int.A	β_0	2.52 (0.247;0.228)	2.77 (0.086;0.272)	2.70 (0.081;0.248)	2.56 (0.075;0.231)
Int.B	β_1	2.77 (0.243;0.249)	2.82 (0.083;0.271)	2.81 (0.078;0.254)	2.77 (0.073;0.250)
Sl.A	β_2	0.56 (0.023;0.045)	0.55 (0.011;0.046)	0.56 (0.011;0.045)	0.57 (0.011;0.045)
Sl.B	β_3	0.61 (0.022;0.043)	0.60 (0.011;0.044)	0.61 (0.011;0.043)	0.61 (0.010;0.043)
R.I.v.	τ^2	6.49 (0.628;0.633)	6.71 (0.226;0.731)	6.67 (0.213;0.680)	6.41 (0.200;0.645)
Res.v.	σ^2	6.94 (0.248;0.466)	7.31 (0.150;0.520)	7.13 (0.140;0.483)	7.05 (0.137;0.472)
Effect	Par.	$\boldsymbol{U}_{\text{wt.lik.}}$	$\boldsymbol{U}_{\text{IPWCC}}$	$\boldsymbol{U}_{\text{IPWCP}}$	$\boldsymbol{U}_{\text{IPWAC}}$
Int.A	β_0	1.85 (0.092;0.303)	2.71 (0.074;0.266)	2.77 (0.079;0.270)	2.59 (0.069;0.237)
Int.B	β_1	2.65 (0.089;0.517)	2.78 (0.073;0.265)	2.82 (0.077;0.269)	2.77 (0.069;0.249)
Sl.A	β_2	0.68 (0.014;0.068)	0.54 (0.010;0.046)	0.53 (0.010;0.044)	0.55 (0.010;0.045)
Sl.B	β_3	0.73 (0.013;0.101)	0.60 (0.010;0.044)	0.59 (0.010;0.044)	0.60 (0.010;0.043)
R.I.v.	τ^2	6.21 (0.235;1.032)	6.66 (0.195;0.717)	6.72 (0.209;0.753)	6.44 (0.187;0.669)
Res.v.	σ^2	5.05 (0.088;0.603)	7.29 (0.130;0.513)	7.59 (0.142;0.562)	7.35 (0.130;0.514)
Effect	Par.		$\boldsymbol{U}_{\text{IPWCC,dr}} = \boldsymbol{U}_{\text{IPWCP,dr}} = \boldsymbol{U}_{\text{IPWAC,dr}}$		
Int.A	β_0		2.52 (0.074;0.226)		
Int.B	β_1		2.77 (0.072;0.247)		
Sl.A	β_2		0.56 (0.011;0.046)		
Sl.B	β_3		0.61 (0.011;0.044)		
R.I.v.	τ^2		6.23 (0.197;0.636)		
Res.v.	σ^2		7.09 (0.139;0.483)		

is taken ($j = 1, \ldots, 7$). Finally, $I(\cdot)$ is an indicator function. Parameter estimates and standard errors, obtained through maximum likelihood and pairwise likelihood, are presented in Table 1.

Observe that all point estimates are relatively close to each other, except for some deviation in the weighted likelihood analysis, a weighted version of conventional likelihood analysis. Note that, with likelihood, there is little rationale for using weights, here leading to a worse fit.

The purely model-based standard errors are meaningful only in the standard likelihood case, where they are reasonable close to the empirically corrected ones. They are not meaningful in the weighted analyses, as they are based on the incorrect assumption that the weights represent replication at the subject (or pair) level. Furthermore, naive standard errors in the pseudo-likelihood case are based on the entirely incorrect assumption that every pair results from independent replication whereas, for example in a completely observed sequence every measurement is used in six different pairs. It is important therefore to use the

empirically corrected standard errors for inferential purposes.

It is clear that using complete cases only resulted in a small loss of efficiency in the naive and IPW cases, whereas the available-case approach makes optimal use of the data. Turning to the doubly robust versions, not only was it confirmed that all three coincide, they were also very close to full likelihood, both in terms of point estimates and precision.

In a relatively large dataset with continuous outcomes, like this one, treating the weights in the weighted analysis as either fixed or random leads to the same standard errors. In the next example, however, this is not the case. The weights were based on a logistic model for dropout in the toenail study:

$$\begin{aligned}&\text{logit}[P(D_i = j|D_i \geq j, T_i, t_j, Y_{i,j-1})]\\&= -3.17(0.24) - 0.28(0.24)T_i + 0.072(0.036)t_j - 0.035(0.036)Y_{i,j-1}. \quad (5.2)\end{aligned}$$

Note that, while the effect of the previous measurement was not significant, the weighted analyses were different from the unweighted ones. In this sense, it is an advantage that the doubly robust versions obviate the need for using weights, as long as the expectations are included. This is not always the case, as it is a consequence of the pairwise marginal nature of the likelihood contributions.

5.2. The national toxicology program data

This developmental toxicity study investigated the dose-response relationship in mice of the potentially hazardous chemical compound di(2-ethylhexyl) phthalate (DEHP), used in vacuum pumps (Windholz (1983)) and as plasticizers for numerous plastic devices made of polyvinyl chloride. The study was conducted in timed-pregnant mice during the period of major organogenesis (Tyl et al. (1988)). The doses selected for the study were 0, 0.025, 0.05, 0.1, and 0.15%, corresponding to a DEHP consumption of 0, 44, 91, 191, and 292 mg/kg/day, respectively. The dams were sacrificed, slightly prior to normal delivery, and the status of uterine implantation sites recorded. A total of 1082 live fetuses were examined for 'malformation,' coded as a binary indicator. Fetuses were clustered within mothers; hence the implied association needs to be accommodated in the analysis. Summary data can be found in Molenberghs and Verbeke (2005).

We fit the models described in Section G.4 to the binary malformation outcome NTP data, with further specification: $\theta_i = \beta_0 + \beta_1 x_i$ and $\delta_i = \beta_d$. Here x_i is rescaled dose, in the sense that the DEHP consumption doses of 0, 44, 91, 191, and 292 mg/kg/day are replaced by unit-interval standardized values 0.0000, 0.1507, 0.3116, 0.6541, and 1.0000, respectively. Given the conditionally specified nature of the model, the conditional version of PL was a more natural choice than the marginal pairwise version.

Table 2. Developmental Toxicity Study (DEHP). Parameter estimates (standard errors) for full likelihood, and naive, singly robust, and doubly robust pseudo-likelihood.

Effect	Par.	$\boldsymbol{U}_{\text{full.lik.}}$	$\boldsymbol{U}_{\text{full.lik., AC}}$	$\boldsymbol{U}_{\text{naive, CC}}$	$\boldsymbol{U}_{\text{naive, AC}}$
Int.	β_0	-1.992 (0.340)	-2.460 (0.535)	-1.772 (2.005)	-1.749 (0.344)
Dose	β_d	2.955 (0.510)	3.207 (0.674)	2.363 (2.644)	2.925 (0.552)
Assoc.	β_a	0.164 (0.027)	0.053 (0.041)	0.163 (0.155)	0.200 (0.029)
Effect	Par.	$\boldsymbol{U}_{\text{IPWCC}}$	$\boldsymbol{U}_{\text{IPWAC}}$	$\boldsymbol{U}_{\text{IPWAC,exch}}$	
Int.	β_0	-2.888 (3.825)	-1.335 (0.831)	-1.470 (0.164)	
Dose	β_d	2.145 (5.969)	4.588 (1.021)	2.225 (0.293)	
Assoc.	β_a	0.130 (0.275)	0.314 (0.055)	0.184 (0.022)	

We considered, apart from full likelihood, naive CC, naive AC, IPWCC, IPWAC, and exchangeable IPWAC. Given the equivalence of the latter to double robustness in the case of exchangeability, there was no need to consider further the other doubly robust versions.

Estimated parameters and standard errors are presented in Table 2. For IPWAC and IPWCC, where explicit models for the weights were needed, we considered (S.62), with parameter estimates (standard errors), $\widehat{\psi}_0 = 1.960(0.110)$, $\widehat{\psi}_1 = 0.018(0.419)$, and $\widehat{\psi}_2 = -2.558(0.391)$.

There are 23 complete litters in the data, where the number of implants equals the number of viable fetuses, out of 108 litters with at least one viable fetus. This dramatic reduction of sample size is reflected in greatly inflated standard errors from $\boldsymbol{U}_{\text{naive, CC}}$ and $\boldsymbol{U}_{\text{IPWCC}}$, up to the point where an otherwise highly significant dose effect is wiped out. Also, the weighted version $\boldsymbol{U}_{\text{IPWAC}}$ shows a decreased efficiency. In contrast, $\boldsymbol{U}_{\text{IPWAC, exch}}$ is efficient and, while doubly robust, does not need an explicit model for the missingness probabilities; hence, it may be preferable.

While in this case it is obviously possible to specify full likelihood, there may be reasons to select one of the singly or doubly robust available-case versions. Indeed, in the case of likelihood, the model parameters are interpreted conditionally on the number of viable fetuses, and this itself is driven by the dose assignment, an experimental variable. The available-case versions take into account the number of implants, m_i. Of course, an available-case likelihood version is in principle possible as well; this was done, labeled $\boldsymbol{U}_{\text{full.lik., AC}}$, and based on the following modification of (S.49):

$$f_i(\boldsymbol{y}_i; \boldsymbol{\Theta}_i, n_i) = \sum_{k=0}^{m_i - n_i} \binom{m_i}{z_i + k} \times \\ \times \exp\left\{\theta_i(z_i + k) - \delta_i(z_i + k)(m_i - z_i + k) - A(\boldsymbol{\Theta}_i)\right\}. \quad (5.3)$$

Again, this expression has the advantage of properly acknowledging the discrepancy between the number of implants and the number of viable fetuses.

6. Concluding Remarks

In this paper, we have laid out a general framework for handling incomplete data predominantly within the pseudo-likelihood setting. Our methodology, applicable under MAR, employs ideas from inverse probability weighting and double robustness. After general development, we focused on the pseudo-likelihood setting, elucidating in detail specific marginal and conditional instances.

Having shown that, under MAR, naive complete-case and available-case estimating equations are biased, we have formulated several alternative versions that overcome this problem, including both singly and doubly robust forms. The second of these requires evaluation of conditional expectations of the unobserved outcomes given the observed ones, which in turn may require joint distributions of a higher order than those used in the singly robust version. While at first sight this seems to undermine the appeal of pseudo-likelihood, the rôle of such joint distributions is solely to construct expectations, and with considerably less computational burden. Sometimes, this might still be impractical, but then the model-based expectation can be replaced by a simpler, but sufficiently rich, model in line with Bang and Robins (2005) and Meng (1994).

While in general doubly robust versions require the specification of both a weight and a predictive model, considerable simplification applies to the important special case of marginal pairwise (or, more generally, n-way) likelihood, also known as composite likelihood. In this case, the doubly robust versions merely require the formulation of a predictive model. In many models these are relatively easy to compute or approximate, as was illustrated for the normal and binary cases. This is a strong asset of the combined use of doubly-robust and composite likelihood ideas. In some cases, though, the formulation of the margins (pairs) may be challenging in its own right. For example, when the conditionally specified model of Section G.4 is used, formulating the full conditional pseudo-likelihood is much easier than the pairs. Thus, there is a tradeoff between simplicity in terms of weights and predictive terms on the one hand, and the pseudo-likelihood contributions themselves on the other.

For the estimation of precision we have indicated how a conventional sandwich-type estimator can be used. Should the derivation of explicit forms be deemed cumbersome, one could resort to such sampling-based methods as stochastic EM, multiple imputation, the bootstrap, and MCMC machinery.

While our work focuses on the MAR setting, in practice one cannot rule out the possibility of an MNAR mechanism. Furthermore, even when MAR is deemed plausible, it is of interest to conduct some form of sensitivity analysis

(Molenberghs and Kenward (2007), Fitzmaurice et al. (2009). Obvious routes include the comparison of PL-based results with those obtained under different paradigms, such as full likelihood and fully Bayesian analyses, the extension of PL to MNAR, advocated by Parzen et al. (2006), and the extension of our results along the lines of Vansteelandt, Rotnitzky and Robins (2007).

We have provided examples of the method, using continuous data from a clinical trial in onychomycosis and binary outcomes from a developmental toxicity study. A simulation study of these methods' operating characteristics is currently ongoing.

The advantage of a variety of proposals is that the user has freedom of selection. Of course, more work is needed to provide further guidance toward such a choice. In particular, small-sample efficiency will be studied in future research. We have indicated, for some specific cases, how standard errors can be derived. These have also been implemented for the data analysis. In the future, standard-error calculations will be undertaken for a variety of other choices. In this respect, it is important to consider methods that do not involve tedious analytical considerations, such as, for example, the jackknife-based method of Heagerty and Lele (1998). While single robustness requires the correct specification of the weights, this requirement is less critical in the doubly robust version, because it is also possible to attain unbiasedness through the predictive term. That said, this result is in need of further qualification. Kang and Schafer (2007) showed empirically that there exist situations where severe biases may occur even when both weight and predictive models are only slightly misspecified. These authors also showed that widely varying weights are a potential risk for bias as well. This underscores that, like any tool in statistics, the user ought to be aware of the relative merits and advantages of the doubly robust method. In this respect, it is highly relevant that, in a number of settings we considered, such as (S.32), the weights cancel from the estimating equations, thereby increasing robustness. A further numerical study of the effect of misspecification of weights and/or predictive terms will be reported elsewhere. Similar issues have been considered by Bang and Robins (2005), Davidian, Tsiatis and Leon (2005), and Rotnitzky (2009).

The case studies were analyzed using a combination of SAS and GAUSS code, which can be obtained from the authors upon request.

Acknowledgements

The authors gratefully acknowledge support from IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. M. (2002). *Topics in Modelling of Clustered Data.* CRC/Chapman & Hall, London.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhyā B* **53**, 233-243.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962-972.

Carey, V. C., Zeger, S. L. and Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517-526.

Cochran, W. G. (1977). *Sampling Techniques.* Wiley, New York.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909-917.

Davidian, M., Tsiatis, A. A. and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with discussion). *Statist. Sci.* **20**, 261-301.

De Backer, M., De Keyser, P., De Vroey, C. and Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day vs. itraconazole 200mg/day-a double-blind comparative trial. *British J. Dermatology* **134**, 16-17.

Diggle, P. J., Heagerty, P., Liang, K-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, New York.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal Data Analysis.* CRC/ Chapman & Hall, Boca Raton.

Geys, H., Molenberghs, G. and Lipsitz, S. R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *J. Statist. Comput. Simulation* **62**, 45-72.

Geys, H., Molenberghs, G. and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *J. Amer. Statist. Assoc.* **94**, 34-745.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.* **93**, 1099-1111.

Joe, H. and Lee, Y. (2008). On weighting of bivariate margins in pairwise likelihood. *J. Multivariate Anal.* **100**, 670-685.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statist. Sci.* **22**, 523-580.

Kenward, M. G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statist. Sci.* **12**, 236-247.

Kenward, M. G. and Molenberghs, G. (2009). Last observation carried forward: A crystal ball? *J. Biopharm. Statist.* **19**, 872-888.

le Cessie, S. and van Houwelingen, J. C. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics* **47**, 1267-1282.

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. *J. Roy. Statist. Soc. Ser. B* **54**, 3-40.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221-239.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9**, 538-558.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Wiley, Chichester.

Molenberghs, G. and Ryan, L. M. (1999). Likelihood inference for clustered multivariate binary data. *Environmetrics* **10**, 279-300.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C. and Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5**, 445-464.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G. and Troxel, A. (2006). Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. *Statist. in Medicine* **25**, 2784-2796.

Robins, J. M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Rotnitzky, A. (2009). Inverse probability weighted methods. In: *Longitudinal Data Analysis* (Edited by G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs), 453-476. CRC/Chapman & Hall, Boca Raton.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *J. Amer. Statist. Assoc.* **94**, 1096-1120 (with Rejoinder, 1135-146).

Tyl, R. W., Price, C. J., Marr, M. C. and Kimmel, C. A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology* **10**, 395-412.

Van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York.

Vansteelandt, S., Rotnitzky, A. and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94**, 841-860.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1-28.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Windholz, M. (1983). *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*. 10th edition. Merck and Co, Rahway, NJ.

I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium.

E-mail: geert.molenberghs@uhasselt.be geert.molenberghs@med.kuleuven.be

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E7HT, UK.

E-mail: mike.kenward@lshtm.ac.uk

I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium.

E-mail: geert.verbeke@med.kuleuven.be

I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium.

E-mail: birhanu.teshomeayele@uhasselt.be

(Received February 2009; accepted September 2009)