PSEUDO-LIKELIHOOD ESTIMATION FOR INCOMPLETE DATA

SUPPLEMENTARY MATERIALS

Geert Molenberghs[1,2] **Michael G. Kenward**[3]

Geert Verbeke[2,1] **Teshome Birhanu**[1]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[3] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E7HT, UK*

A Generalized Estimating Equations

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. This is typically done via $h(\mu_{ij}) = \boldsymbol{x'_{ij}}\boldsymbol{\beta}$, with $h(\cdot)$ some known link function, such as the logit link for binary responses. 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 (overdispersion) 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 of $\boldsymbol{Y}_i$ can then be written as $V_i = V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \phi A_i^{1/2} R_i A_i^{1/2}$, with A_i the matrix with the marginal variances on the main diagonal and zeros elsewhere.

Generalized estimating equations take the form

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

The nuisance parameter $\boldsymbol{\alpha}$ needs to be replaced by a consistent estimate; Liang and Zeger (1986) proposed a moment-based estimator for this.

Assuming that the marginal mean $\boldsymbol{\mu_i}$ has been correctly modeled, it can be shown that, under mild regularity conditions, the estimator $\widehat{\boldsymbol{\beta}}$ obtained from solving (S.1) is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and with covariance matrix

$$\text{var}(\widehat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}, \tag{S.2}$$

where

$$I_0 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}'}, \qquad I_1 = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}'}{\partial \boldsymbol{\beta}} V_i^{-1} \text{Var}(\boldsymbol{y_i}) V_i^{-1} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}'}. \tag{S.3}$$

In practice, $\text{Var}(\boldsymbol{y_i})$ in (S.3) is replaced by $(\boldsymbol{y_i} - \boldsymbol{\mu_i})(\boldsymbol{y_i} - \boldsymbol{\mu_i})'$, which is unbiased on the sole condition, again, that the mean was correctly specified.

As stated earlier, GEE is not likelihood based and therefore ignorability (Rubin 1976) cannot be invoked to establish the method's validity under MAR. Therefore, apart from special cases, GEE in its basic form will be 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 then to weigh each subject's contribution to the GEEs 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 at occasion d_i. These can be written as

$$\pi_i = \prod_{\ell=2}^{n_i} (1 - p_{i\ell}), \tag{S.4}$$

$$\pi_i' = \left[\prod_{\ell=2}^{d_i - 1} (1 - p_{i\ell}) \right] \cdot p_{id_i}, \tag{S.5}$$

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 choose either for WGEE based on the completers only:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \;=\; \mathbf{0}, \tag{S.6}$$

with $\widetilde{R}_i = 1$ if a subject is fully observed and 0 otherwise, or, upon using (6), for WGEE using all subjects:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{1}{\pi_i'} \frac{\partial \boldsymbol{\mu}_i^o}{\partial \boldsymbol{\beta}'} (V_i^o)^{-1} (\boldsymbol{y}_i^o - \boldsymbol{\mu}_i^o) \;=\; \mathbf{0}. \tag{S.7}$$

Here, the superscript 'o' indicates the portion corresponding to the observed data in the corresponding matrix or vector. In (S.6), the incomplete subjects contribute through the model for the dropout probabilities π_i. The above development only focuses on dropout but can be generalized to encompass non-monotone missingness as well (Vansteelandt, Rotnitzky, and Robins 2007).

Estimators from WGEE enjoy robustness properties similar to the ones from regular GEE, i.e., the correlation structure does not need to be correctly specified. Applying WGEE is technically feasible and can be conducted using the SAS procedure GENMOD. Of course, some extra programming is needed to construct the weights.

As stated earlier, (S.6) has been extended towards so-called double robustness (Scharfstein, Rotnitzky, and Robins 1999, Van der Laan and Robins 2003, Bang and Robins 2005). We will 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.

B Consistency and Asymptotic Normality of the Pseudo-likelihood Estimator

We first list the required regularity conditions on the density functions $f_s(\boldsymbol{y}^{(s)};\boldsymbol{\beta})$.

A0 The densities $f_s(\boldsymbol{y}^{(s)};\boldsymbol{\beta})$ are distinct for different values of the parameter $\boldsymbol{\beta}$.

A1 The densities $f_s(\boldsymbol{y}^{(s)};\boldsymbol{\beta})$ have common support, which does not depend on $\boldsymbol{\beta}$.

A2 The parameter space Ω contains an open region ω of which the true parameter value $\boldsymbol{\beta}_0$ is an interior point.

A3 ω is such that for all s, and almost all $\boldsymbol{y}^{(s)}$ in the support of $\boldsymbol{Y}^{(s)}$, the densities admit all third derivatives

$$\frac{\partial^3 f_s(\boldsymbol{y}^{(s)};\boldsymbol{\beta})}{\partial\theta_j\partial\theta_k\partial\theta_\ell}.$$

A4 The first and second logarithmic derivatives of f_s satisfy

$$E_{\boldsymbol{\beta}}\left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)};\boldsymbol{\beta})}{\partial\theta_k}\right) = 0, \qquad k = 1,\ldots,q,$$

and

$$0 < E_{\boldsymbol{\beta}} \left(\frac{-\partial^2 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\beta})}{\partial\theta_k \partial\theta_\ell} \right) < \infty, \qquad k, \ell = 1, \ldots, q.$$

A5 The matrix I_0, defined in (S.8), is positive definite.

A6 There exist functions M_{klr} such that

$$\sum_{s \in S} \delta_s E_{\boldsymbol{\beta}} \left| \frac{\partial^3 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\beta})}{\partial\theta_k \partial\theta_\ell \partial\theta_r} \right| < M_{k\ell r}(\boldsymbol{y})$$

for all $\boldsymbol{y}$ in the support of f and for all $\boldsymbol{\theta} \in \omega$ and $m_{k\ell r} = E_{\boldsymbol{\beta}_0}(M_{k\ell r}(Y)) < \infty$.

Theorem 1, proven by Arnold and Strauss (1991), guarantees the existence of at least one solution to the pseudo-likelihood equations, which is a consistent and asymptotically normal estimator. Without loss of generality, we can assume $\boldsymbol{\beta}$ is constant. Replacing it by $\boldsymbol{\beta}_i$, and modeling it as a function of covariates is straightforward.

Theorem 1 (Consistency and Asymptotic Normality) *Assume that* $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ *are i.i.d. with common density that depends on* $\boldsymbol{\beta}_0$. *Then under regularity conditions (A1)–(A6):*

1. *the pseudo-likelihood estimator* $\tilde{\boldsymbol{\beta}}_N$, *defined as the maximizer of (9), converges in probability to* $\boldsymbol{\beta}_0$.

2. $\sqrt{N}(\tilde{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0)$ *converges in distribution to* $N_p(\boldsymbol{0}, I_0(\boldsymbol{\beta}_0)^{-1} I_1(\boldsymbol{\beta}_0) I_0(\boldsymbol{\beta}_0)^{-1})$ *with* $I_0(\boldsymbol{\beta})$ *defined by*

$$I_{0,k\ell}(\boldsymbol{\beta}) = -\sum_{s \in S} \delta_s E_{\boldsymbol{\beta}} \left(\frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\beta})}{\partial\theta_k \partial\theta_\ell} \right) \tag{S.8}$$

and $I_1(\boldsymbol{\beta})$ *by*

$$I_{1,k\ell}(\boldsymbol{\beta}) = \sum_{s,t \in S} \delta_s \delta_t E_{\boldsymbol{\beta}} \left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)}; \boldsymbol{\beta})}{\partial\theta_k} \frac{\partial \ln f_t(\boldsymbol{y}^{(t)}; \boldsymbol{\beta})}{\partial\theta_\ell} \right). \tag{S.9}$$

C Pairwise and Higher-order Marginal Pseudo-likelihood

C.1 Pairwise Pseudo-likelihood

As stated earlier, marginal models for non-Gaussian data can become prohibitive when subjected to full maximum likelihood inference, especially with large within-unit replication. le Cessie and van Houwelingen (1991) and Geys, Molenberghs, and Lipsitz (1998) replace the true contribution of a vector of correlated binary data to the full likelihood, written as $f(y_{i1}, \ldots, y_{in_i})$, by the product of all pairwise contributions $f(y_{ij}, y_{ik})$ $(1 \leq j < k \leq n_i)$, to obtain a pseudo-likelihood function. Also the term *composite likelihood* is encountered in this context. Renard, Molenberghs, and Geys (2004) refer to this particular instance of pseudo-likelihood as *pairwise likelihood*. Grouping the outcomes for subject i into a vector $\boldsymbol{Y}_i$, the contribution of the ith cluster to the log pseudo-likelihood then specializes to

$$p\ell_i = \sum_{j<k} \ln f(y_{ij}, y_{ik}), \tag{S.10}$$

if it contains more than one observation. Otherwise $p\ell_i = f(y_{i1})$. Extension to three-way and higher-order pseudo-likelihood is straightforward. All of these are special cases of (9).

C.2 Full Conditional Pseudo-likelihood

Some models lend themselves more easily to conditioning than to marginalization, such as log-linear models (Molenberghs and Verbeke 2005, Ch. 12). Upon noting that

$$f(y_{ij}|y_{ik}, k \neq j) = \frac{f(y_{i1}, \ldots, y_{in_i})}{f(y_{i1}, \ldots, y_{i,j-1}, y_{i,j+1}, \ldots, y_{in_i})} = \frac{f_{1}(\boldsymbol{y}_i^{(1)})}{f_{s_j}(\boldsymbol{y}_i^{(s_j)})},$$

a full conditional likelihood contribution becomes:

$$p\ell_i = n_i \cdot \ln f_{1}(\boldsymbol{y}_i^{(1)}) \; - \; \sum_{j=1}^{n_i} \ln f_{s_j}(\boldsymbol{y}_i^{(s_j)}).$$

Here, $\mathbf{1}$ is a vector of ones and s_j is a vector of ones, with a single 0 in the jth entry. Evidently, alternative versions of conditional pseudo-likelihood are possible. For example, one could consider all pairs, conditioning upon the remaining $n_i - 2$ outcomes. This setting has been considered by Geys, Molenberghs, and Ryan (1999) for the analysis of the NTP data (Section 5.2). This particular setting, focusing on the missing-data aspect, is taken up in Section G.4.

D Single-robustness Theorem 2

The following theorem establishes single robustness.

Theorem 2 (Single robustness of $U_{\mathbf{IPWCC}}$, $U_{\mathbf{IPWAC}}$, and $U_{\mathbf{IPWAC,seq}}$.) *Under MAR, and if $p_{i\ell}$ in (6)–(6) is non-parametrically or correctly parametrically specified as $p_{i\ell}(\boldsymbol{\psi})$, then U_{IPWCC}, U_{IPWAC}, and $U_{IPWAC,seq}$ produce consistent estimators.*

In the above, and also in what follows, the same regularity conditions apply as in Rotnitzky (2009). In particular, it is important that the probability of being observed for a measurement be bounded away from zero.

Proof. This follows from their expectation being 0, as follows:

$$\begin{aligned} E(\boldsymbol{U}_{\text{IPWCC}}) &= E_Y\left\{\sum_{i=1}^N E_{R_i|Y_i}\left[\frac{\widetilde{R}_i}{\pi_i}\boldsymbol{U}_i(\boldsymbol{Y}_i)\right]\right\} \\ &= E_Y\left\{\sum_{i=1}^N \left[\frac{E_{R_i|Y_i}(\widetilde{R}_i)}{\pi_i}\boldsymbol{U}_i(\boldsymbol{Y}_i)\right]\right\} \\ &= E_Y\left[\sum_{i=1}^N U_i(\boldsymbol{Y}_i)\right] = \mathbf{0}. \end{aligned} \tag{S.11}$$

$$\begin{aligned} E(\boldsymbol{U}_{\text{IPWAC}}) &= E_Y\left\{\sum_{i=1}^N E_{R_i|Y_i}\left[\frac{R_i'}{\pi_i'}E_{Y^m|y^o}\boldsymbol{U}_i(\boldsymbol{Y}_i)\right]\right\} \\ &= E_Y\left\{\sum_{i=1}^N \left[\frac{E_{R_i|Y_i}(R_i')}{\pi_i'}E_{Y^m|y^o}\boldsymbol{U}_i(\boldsymbol{Y}_i)\right]\right\} \\ &= \sum_{i=1}^N E_Y E_{Y^m|y^o}\boldsymbol{U}_i(\boldsymbol{Y}_i) = E_Y\left[\sum_{i=1}^N \boldsymbol{U}_i(\boldsymbol{Y}_i)\right] = \mathbf{0}. \end{aligned} \tag{S.12}$$

$$E(\boldsymbol{U}_{\text{IPWAC,seq}}) = E_Y\left\{\sum_{i=1}^N E_{R_i|Y_i}\left[\sum_{j=1}^{n_i}\frac{R_{ij}}{\pi_{ij}}E_{Y^m|y^o}\boldsymbol{U}_i(\boldsymbol{Y}_{ij}|\boldsymbol{Y}_{i\overline{j}})\right]\right\}$$

$$
\begin{aligned}
&= E_Y \left\{ \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} E_{R_j|R_{\overline{j}}Y} \frac{R_{ij}}{\pi_{ij}} E_{Y^m|y^o} \boldsymbol{U}_i(\boldsymbol{Y}_{ij}|\boldsymbol{Y}_{i\overline{j}}) \right] \right\} \\
&= E_Y \left[\sum_{i=1}^{N} \sum_{j=1}^{n_i} \boldsymbol{U}_i(\boldsymbol{Y}_i) \right] = \mathbf{0}. \qquad \text{(S.13)}
\end{aligned}
$$

Here, $E_{R_j|R_{\overline{j}}Y}$ is the expectation relative to R_j, given the missingness history up to occasion j and given the outcomes $\boldsymbol{Y}$. Note that, in the CC case, we used $E_{R_i|Y_i}(R_i) = E_{R|Y^o}(R_i) = \pi_i$, owing to MAR. A similar statement holds in the AC case. This completes the proof.

E Double-robustness Theorem 3

We now establish double robustness.

Theorem 3 (Double robustness of $U_{\mathbf{IPWCC,dr}}$ and $U_{\mathbf{IPWAC,dr}}$.) *Under MAR, and (a) if $p_{i\ell}$ in (6)–(6) is non-parametrically or correctly parametrically specified as $p_{i\ell}(\boldsymbol{\psi})$ and/or (b) if the predictive models in (17) and (18) are correctly specified, then $\boldsymbol{U}_{IPWCC,dr}$ and $U_{IPWAC,dr}$ are consistent.*

Proof. If condition (a) holds, then the result trivially follows from Theorem 2 and the observation that the expectation of the first factors of the second terms on the right hand sides equal zero. Under condition (b), write $E_{R_i|Y_i}(R_i) = E_{R|Y^o}(R_i) = \lambda_i$. Then,

$$
\begin{aligned}
E(\boldsymbol{U}_{\text{IPWCC,dr}}) &= E_Y \left\{ \sum_{i=1}^{N} \left[\frac{\lambda_i}{\pi_i} \boldsymbol{U}_i(\boldsymbol{Y}_i) + \left(1 - \frac{\lambda_i}{\pi_i}\right) E_{Y_i^m|y_i^o} \boldsymbol{U}_i(\boldsymbol{Y}_i) \right] \right\} \\
&= \sum_{i=1}^{N} \left\{ \frac{\lambda_i}{\pi_i} E_{Y_i^m} E_{Y_i^m|Y_i^o} \left[\boldsymbol{U}_i(\boldsymbol{Y}_i) \right] \right. \\
&\quad \left. + \left(1 - \frac{\lambda_i}{\pi_i}\right) E_{Y_i^m} E_{Y_i^m|Y_i^o} \left[E_{Y_i^m|y_i^o} \boldsymbol{U}_i(\boldsymbol{Y}_i) \right] \right\} \\
&= \sum_{i=1}^{N} E_{Y_i^m} E_{Y_i^m|Y_i^o} \left[\boldsymbol{U}_i(\boldsymbol{Y}_i) \right] = \sum_{i=1}^{N} E_Y \left[\boldsymbol{U}_i(\boldsymbol{Y}_i) \right] = \mathbf{0}. \qquad \text{(S.14)}
\end{aligned}
$$

The AC case starts with similar logic for the case condition (a) holds. When (b) holds, but not necessarily (a):

$$\begin{aligned}
E(\boldsymbol{U}_{\text{IPWAC,dr}}) &= E_Y \left\{ \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \frac{\lambda_{ij}}{\pi_{ij}} \boldsymbol{U}_i(Y_i|\boldsymbol{Y}_{i\overline{j}}) \right.\right. \\
&\quad \left.\left. + \left(1 - \frac{\lambda_{ij}}{\pi_{ij}}\right) E_{Y_i^m|y_i^o} \boldsymbol{U}_i(Y_i|\boldsymbol{Y}_{i\overline{j}}) \right] \right\} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ \frac{\lambda_{ij}}{\pi_{ij}} E_{Y_i^m} E_{Y_i^m|Y_i^o} \left[\boldsymbol{U}_i(Y_i|\boldsymbol{Y}_{i\overline{j}}) \right] \right. \\
&\quad \left. + \left(1 - \frac{\lambda_{ij}}{\pi_{ij}}\right) E_{Y_i^m} E_{Y_i^m|Y_i^o} \left[E_{Y_i^m|y_i^o} U_i(Y_i|\boldsymbol{Y}_{i\overline{j}}) \right] \right\} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} E_{Y_i^m} E_{Y_i^m|Y_i^o} \left[U_i(Y_i|\boldsymbol{Y}_{i\overline{j}}) \right] \\
&= \sum_{i=1}^{N} E_Y \left[\boldsymbol{U}_i(\boldsymbol{Y}_i) \right] = \boldsymbol{0}. \qquad \text{(S.15)}
\end{aligned}$$

This completes the proof.

F Sandwich Estimator for U_{IPWCC} and $U_{IPWCC,dr}$ with Normal Data

Write a subject's contribution to (S.26) as

$$\boldsymbol{V}_i = \frac{\widetilde{R}_i}{\pi_i} \sum_{j<k} \boldsymbol{U}_i(y_{ij}, y_{ik}) = \frac{\widetilde{R}_i}{\pi_i} \sum_{j<k} \frac{\partial \ell_{ijk}}{\partial \boldsymbol{\beta}} = \frac{\widetilde{R}_i}{\pi_i} \boldsymbol{U}_i. \qquad \text{(S.16)}$$

The model for missingness can be written in logistic form as:

$$\pi_i = \prod_{j=2}^{n_i} \left(1 + e^{\boldsymbol{z}'_{ij}\boldsymbol{\psi}}\right)^{-1},$$

where $\boldsymbol{z}_{ij}$ is a vector containing relevant covariates and outcomes from the history prior to occasion j. Then,

$$\begin{aligned}
\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\beta}} &= \frac{\widetilde{R}_i}{\pi_i} \cdot K' \frac{\partial^2 \ell_{ijk}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma}) \partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'} K, \qquad \text{(S.17)} \\
\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} &= \frac{\widetilde{R}_i}{\pi_i} \cdot \boldsymbol{U}_i \sum_{k=2}^{n_i} \boldsymbol{z}_{ik} p_{ik}, \qquad \text{(S.18)}
\end{aligned}$$

with

$$K = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\alpha}} \end{pmatrix}, \qquad p_{ik} = \frac{e^{\boldsymbol{z}'_{ik}\boldsymbol{\psi}}}{1 + e^{\boldsymbol{z}'_{ik}\boldsymbol{\psi}}}.$$

Next, the estimating equation W_i for the $\boldsymbol{\psi}$ parameters follows from its logistic structure, with data of the form $(R_{ij}, \boldsymbol{z}_{ij})$, for $i = 1, \ldots, N$ and $j = 1, \ldots, d_i$, and $R_{ij} = 0$ if $j < d_i$, and 1 otherwise. Following standard generalized linear models theory, we have that

$$\boldsymbol{W}_i = \sum_{j=2}^{d_i} \boldsymbol{z}'_{ij}(R_{ij} - p_{ij}). \tag{S.19}$$

Hence,

$$\frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} = -\sum_{j=2}^{d_i} (\boldsymbol{z}_{ij} \cdot \boldsymbol{z}'_{ij}) p_{ij}(1 - p_{ij}). \tag{S.20}$$

The sandwich estimator then follows from plugging the expressions (S.16) and (S.19) for the scores, and (S.17), (S.18), and (S.20) for the second derivatives, into (19) and (20). We still need an expression for

$$\frac{\partial^2 \ell_{ijk}}{\partial(\beta, \alpha)\partial(\beta, \alpha)'}.$$

Define

$$H^{(2)} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\sigma}}, \qquad Q^{(2)} = \frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\sigma}},$$

with $\boldsymbol{h} = (h_{jj}, h_{jk}, h_{kk})'$ and $\boldsymbol{Q} = (Q_{jj}, Q_{jk}, Q_{kk})'$. Then,

$$H^{(2)} = \frac{1}{\varphi^2} \begin{pmatrix} -\frac{1}{2}\sigma_{kk}^2 & \sigma_{jj}\sigma_{kk} & \frac{1}{2}\sigma_{jk}^2 \\ -\sigma_{kk}\sigma_{jk} & \sigma_{jj}\sigma_{kk} + \sigma_{jk}^2 & -\sigma_{jj}\sigma_{jk} \\ \frac{1}{2}\sigma_{jk}^2 & \sigma_{jj}\sigma_{kk} & -\frac{1}{2}\sigma_{jj}^2 \end{pmatrix}.$$

The generic element of $Q^{(2)}$ is

$$Q_{\sigma,\tau} = -\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\Sigma^{-1}\left(S_\sigma \Sigma^{-1} S_\tau + S_\tau \Sigma^{-1} S_\sigma\right)\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i).$$

Finally,

$$\frac{\partial^2 \ell_{ijk}}{\partial(\beta, \alpha)\partial(\beta, \alpha)'} = \left(\begin{array}{c|c} -\Sigma^{-1} & T^{(2)} \\ \hline T^{(2)'} & H^{(2)} + Q^{(2)} \end{array}\right),$$

where $T^{(2)}$ is a 2×3 matrix with columns $-\Sigma^{-1} S_\sigma \Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$.

We now consider the doubly robust version (S.29). Evidently, $\boldsymbol{W}_i$ and $\partial \boldsymbol{W}_i/\partial \boldsymbol{\psi}$ remain as before, with the same holding true for the form of $\boldsymbol{S}_i$ and A_i. However, the contribution $\boldsymbol{V}_i$ of subject i changes and can also be written as

$$\begin{aligned}
\boldsymbol{V}_i &= \boldsymbol{V}_i^{(1)} + \left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) \boldsymbol{V}_i^{(2)}, \\
\boldsymbol{V}_i^{(1)} &= \sum_{j<k<d_i} \boldsymbol{U}_i(y_{ij}, y_{ik}), \\
\boldsymbol{V}_i^{(2)} &= \sum_{j=1}^{d_i-1} (n_i - d_i + 1)\boldsymbol{U}_i(y_{ij}) + \sum_{j<d_i\leq k} E\left[\boldsymbol{U}_i(y_{ik}|y_{ij})\right] + \sum_{d_i\leq j<k} E\left[\boldsymbol{U}_i(y_{ij}, y_{ik})\right].
\end{aligned}$$

We need only the derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. Regarding the latter, we obtain:

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = -\frac{\widetilde{R}_i}{\pi_i} \boldsymbol{V}_i^{(2)} \sum_{k=2}^{n_i} \boldsymbol{z}_{ik} p_{ik},$$

while for the former, the general form is

$$\frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{V}_i^{(1)}}{\partial \boldsymbol{\beta}} + \left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) \frac{\partial \boldsymbol{V}_i(2)}{\partial \boldsymbol{\beta}}.$$

Now, denote by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{n_i})'$, the entire mean vector and by $\boldsymbol{\sigma} = \text{vech}(\Sigma)$, the vector of unique variance-covariance matrix elements. It then easily follows that

$$\frac{\partial \boldsymbol{V}_i^{(1)}}{\partial \boldsymbol{\beta}} = K' \left(\sum_{j<k<d_i} \frac{\partial^2 \ell_{ijk}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'} \right) K, \tag{S.21}$$

$$\begin{aligned}
\frac{\partial \boldsymbol{V}_i^{(2)}}{\partial \boldsymbol{\beta}} = K' \Bigg[& \sum_{j<d_i} (n_i - d_i + 1) \frac{\partial^2 \ell_{ij}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'} + \sum_{j<d_i\leq k} \frac{\partial}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})} E\left(\frac{\partial \ell_{ik|j}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'}\right) \\
& + \sum_{d_i\leq j<k} \frac{\partial}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})} E\left(\frac{\partial \ell_{ijk}}{\partial(\boldsymbol{\mu}, \boldsymbol{\sigma})'}\right) \Bigg] K.
\end{aligned} \tag{S.22}$$

The derivatives in (S.21)–(S.22) follow in the same fashion as in the single robust case, starting from explicit expressions (S.36)–(S.39).

G Details for Pairwise and Full Conditional Pseudo-likelihood

G.1 Pairwise Likelihood

While in principle general missingness could be considered, we focus on the important special case of dropout, to streamline mathematical development. The forms (21)–(29) take the following form for the specific case of pairwise likelihood:

$$\boldsymbol{U}_{\text{naive, CC}} = \sum_{i=1}^{N} R_i \sum_{j<k} \boldsymbol{U}_i(Y_{ij}, Y_{ik}), \tag{S.23}$$

$$\boldsymbol{U}_{\text{naive, CP}} = \sum_{i=1}^{N} \sum_{j<k<d_i} \boldsymbol{U}_i(Y_{ij}, Y_{ik}), \tag{S.24}$$

$$\boldsymbol{U}_{\text{naive, AC}} = \sum_{i=1}^{N} \left[\sum_{j<k<d_i} \boldsymbol{U}_i(Y_{ij}, Y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \boldsymbol{U}_i(Y_{ij}) \right], \tag{S.25}$$

$$\boldsymbol{U}_{\text{IPWCC}} = \sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \left[\sum_{j<k} \boldsymbol{U}_i(Y_{ij}, Y_{ik}) \right], \tag{S.26}$$

$$\boldsymbol{U}_{\text{IPWCP}} = \sum_{i=1}^{N} \sum_{j<k<d_i} \frac{R_{ijk}}{\pi_{ijk}} \cdot \boldsymbol{U}_i(Y_{ij}, Y_{ik}), \tag{S.27}$$

$$\boldsymbol{U}_{\text{IPWAC}} = \sum_{i=1}^{N} \sum_{j<k} \left[\frac{R_{ij}}{\pi_{ij}} \cdot \boldsymbol{U}_i(Y_{ij}) + \frac{R_{ik}}{\pi_{ik}} \cdot \boldsymbol{U}_i(Y_{ik}|Y_{ij}) \right], \tag{S.28}$$

$$\begin{aligned} \boldsymbol{U}_{\text{IPWCC,dr}} = \sum_{i=1}^{N} \Bigg\{ & \frac{\widetilde{R}_i}{\pi_i} \left[\sum_{j<k} \boldsymbol{U}_i(Y_{ij}, Y_{ik}) \right] \\ & + \left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) E_{\boldsymbol{Y}_i^m|\boldsymbol{Y}_i^o} \left[\sum_{j<k} \boldsymbol{U}_i(Y_{ij}, Y_{ik}) \right] \Bigg\}, \end{aligned} \tag{S.29}$$

$$\begin{aligned} \boldsymbol{U}_{\text{IPWCP,dr}} = \sum_{i=1}^{N} \sum_{j<k<n_i} \Bigg[& \frac{R'_{ijk}}{\pi'_{ijk}} \cdot \boldsymbol{U}_i(Y_{ij}, Y_{ik}) \\ & + \left(1 - \frac{R'_{ijk}}{\pi'_{ijk}}\right) \cdot E_{\boldsymbol{Y}_i^m|\boldsymbol{Y}_i^o} \boldsymbol{U}_i(Y_{ij}, Y_{ik}) \Bigg], \end{aligned} \tag{S.30}$$

$$\boldsymbol{U}_{\text{IPWAC,dr}} = \sum_{i=1}^{N} \sum_{j<k} \left[\frac{R_{ij}}{\pi_{ij}} \cdot \boldsymbol{U}_i(Y_{ij}) + \frac{R_{ik}}{\pi_{ik}} \cdot \boldsymbol{U}_i(Y_{ik}|Y_{ij}) \right.$$

$$+\left(1-\frac{R_{ij}}{\pi_{ij}}\right)\cdot E_{\boldsymbol{Y}_i^m|\boldsymbol{Y}_i^o}\boldsymbol{U}_i(Y_{ij})$$
$$+\left(1-\frac{R'_{ik}}{\pi'_{ik}}\right)\cdot E_{\boldsymbol{Y}_i^m|\boldsymbol{Y}_i^o}\boldsymbol{U}_i(Y_{ik}|Y_{ij})\Bigg]. \tag{S.31}$$

Here, $R'_i = d_i$ if subject i drops out at occasion d_i. We can now write $\pi_i = \prod_{\ell=2}^{n_i}(1-p_{i\ell})$, where still $p_{i\ell} = P\left(D_i = \ell | D_i \geq \ell, Y_{i\overline{\ell}}, X_{i\overline{\ell}}\right)$. The second term in (S.25) results from all pairs with the first component observed and the second one unobserved.

It is interesting, and easy to show, that all three of the doubly robust versions coincide in this case, which adds to their attraction:

$$\begin{aligned}\boldsymbol{U}_{\text{IPWCC,dr}} &= \boldsymbol{U}_{\text{IPWCP,dr}} = \boldsymbol{U}_{\text{IPWAC,dr}} \\ &= \sum_{i=1}^{N}\Bigg\{\sum_{j<k<d_i}\boldsymbol{U}_i(Y_{ij},Y_{ik}) + \sum_{j=1}^{d_i-1}(n_i-d_i+1)\cdot\boldsymbol{U}_i(Y_{ij}) \\ &\quad + \sum_{j<d_i\leq k}E[\boldsymbol{U}_i(Y_{ik}|Y_{ij})] + \sum_{d_i\leq j<k}E[\boldsymbol{U}_i(Y_{ij},Y_{ik})]\Bigg\}.\end{aligned} \tag{S.32}$$

A key feature in (S.32) is that the need to model the missing-data mechanism is avoided. Note that this expression is related to (S.25) in the sense that both terms of the latter expression occur here as well, with in addition the predictive terms. There are two types of predictive terms, corresponding to: (a) a pair with the first component observed and the second one missing; (b) a pair with both components missing.

All predictive models involve two types of contributions: for $E[\boldsymbol{U}_i(Y_{ik}|Y_{ij})]$ where Y_{ij} is observed but Y_{ik} is not, and for $E[\boldsymbol{U}_i(Y_{ij}, Y_{ik})]$ with both unobserved. These will be considered for the special but important cases that follow next.

It is very easy to derive an exchangeable form, starting from (S.31), because then, in this expression, the expectations vanish. Hence, clearly, the exchangeable form is equal to (S.25), making the naive available case version not only valid, but actually doubly robust. Of course, this is the case only under exchangeability.

An important observation is that in the doubly robust versions (S.32), the need to specify the missing-data model is avoided, even though the predictive model for the unobserved outcomes is needed.

G.2 Multivariate Normal

Assume $\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}, \Sigma)$. Then first, suppressing the index i from notation, and writing down the expressions for observed values, we find:

$$\begin{aligned}
\boldsymbol{U}(y_k|y_j) &= \frac{\partial(\mu_{k|j}, \sigma_{kk|j})}{\partial(\mu_j, \mu_k, \sigma_{jj}, \sigma_{jk}, \sigma_{kk})} \cdot \frac{\partial \ln \phi(y_k|y_j; \mu_{k|j}, \sigma_{kk|j})}{\partial(\mu_{k|j}, \sigma_{kk|j})} \\
&= \begin{pmatrix} -\frac{\sigma_{jk}}{\sigma_{jj}} & 0 \\ 1 & 0 \\ -\frac{\sigma_{jk}}{\sigma_{jj}^2}(y_j - \mu_j) & \frac{\sigma_{jk}^2}{\sigma_{jj}^2} \\ \frac{y_j - \mu_j}{\sigma_{jj}} & -\frac{2\sigma_{jk}}{\sigma_{jj}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{y_k - \mu_{k|j}}{\sigma_{kk|j}} \\ -\frac{1}{2\sigma_{kk|j}} + \frac{1}{2}\frac{(y_k - \mu_{k|j})^2}{\sigma_{kk|j}^2} \end{pmatrix},
\end{aligned} \tag{S.33}$$

where $\phi(\cdot)$ is the normal density with mean and variance given by:

$$\mu_{k|j} = \mu_k + \frac{\sigma_{jk}}{\sigma_{jj}}(y_j - \mu_j) \qquad \text{and} \qquad \sigma_{kk|j} = \frac{\sigma_{jj}\sigma_{kk} - \sigma_{jk}^2}{\sigma_{jj}}.$$

The only stochastic elements in (S.33) are the conditional residual and its square. We need to take their expectation conditional upon the observed outcomes, producing for the second factor in (S.33):

$$\begin{pmatrix} \frac{\sigma_{jj}\Sigma_{k\overline{d}}\Sigma^{-1}_{\overline{d}\,\overline{d}}(\boldsymbol{Y}_{\overline{d}} - \boldsymbol{\mu}_{\overline{d}}) - \sigma_{jk}(y_j - \mu_j)}{\sigma_{jj}\sigma_{kk} - \sigma_{jk}^2} \\ \frac{\sigma_{jj}\left(\sigma_{jk}^2 - \sigma_{jj}\Sigma_{k\overline{d}}\Sigma^{-1}_{\overline{d}\,\overline{d}}\Sigma_{\overline{d}k}\right) + \left[\sigma_{jj}\Sigma k\overline{d}\,\Sigma^{-1}_{\overline{d}\,\overline{d}}(\boldsymbol{Y}_{\overline{d}} - \boldsymbol{\mu}_{\overline{d}}) - \sigma_{jk}(y_j - \mu_j)\right]^2}{2(\sigma_{jj}\sigma_{kk} - \sigma_{jk}^2)^2} \end{pmatrix}. \tag{S.34}$$

Here, $\overline{d}$ refers to the set of indices $(1, 2, \ldots, d-1)$, corresponding to the observed portion of $\boldsymbol{Y}$.

Turning to the other expectation, we find:

$$\begin{aligned}
\boldsymbol{U}(y_j, y_k) &= \frac{\partial \ln \phi(y_j, y_k; \mu_j, \mu_k, \sigma_{jj}, \sigma_{jk}, \sigma_{kk})}{\partial(\mu_j, \mu_k, \sigma_{jj}, \sigma_{jk}, \sigma_{kk})} \\
&= \begin{pmatrix} \Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \\ h_{jj} + Q_{jj} \\ h_{jk} + Q_{jk} \\ h_{kk} + Q_{kk} \end{pmatrix},
\end{aligned} \tag{S.35}$$

where

$$h_{jj} = -\frac{1}{2}\frac{\sigma_{kk}}{\varphi}, \qquad h_{jk} = \frac{\sigma_{jk}}{\varphi}, \qquad h_{kk} = -\frac{1}{2}\frac{\sigma_{jj}}{\varphi},$$

$$\varphi = \sigma_{jj}\sigma_{kk} - \sigma_{jk}^2,$$

$$Q_\sigma = \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\Sigma^{-1}S_\sigma\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),$$

$$S_{jj} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad S_{jk} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad S_{kk} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Here, S_σ is generic notation for either one of the three pairs (j,j), (j,k), and (k,k).

To calculate the expectation of (S.35), we need:

$$\begin{aligned} E(\boldsymbol{Y}|\boldsymbol{y}_{\overline{d}}) &= \boldsymbol{\mu}_{jk}^c = \boldsymbol{\mu} + \Sigma_{jk,\overline{d}}\Sigma_{\overline{d},\overline{d}}^{-1}(\boldsymbol{y}_{\overline{d}} - \boldsymbol{\mu}_{\overline{d}}), && \text{(S.36)} \\ \text{var}(\boldsymbol{Y}|\boldsymbol{y}_{i\overline{d}}) &= \Sigma_{jk,jk} - \Sigma_{jk,\overline{d}}\Sigma_{\overline{d},\overline{d}}^{-1}\Sigma_{\overline{d},jk}. && \text{(S.37)} \end{aligned}$$

It now follows that

$$E\left[\boldsymbol{U}(y_j,y_k)|\boldsymbol{y}_{\overline{d}}\right] = \begin{pmatrix} \Sigma_{jk,jk}^{-1}\Sigma_{jk,\overline{d}}\Sigma_{\overline{d},\overline{d}}^{-1}(\boldsymbol{y}_{\overline{d}} - \boldsymbol{\mu}_{\overline{d}}) \\ h_{jj} + E[Q_{jj}|\boldsymbol{y}_{\overline{d}}] \\ h_{jk} + E[Q_{jk}|\boldsymbol{y}_{\overline{d}}] \\ h_{kk} + E[Q_{kk}|\boldsymbol{y}_{\overline{d}}] \end{pmatrix}, \qquad \text{(S.38)}$$

where some straightforward algebra produces:

$$\begin{aligned} E[Q_\sigma|\boldsymbol{y}_{\overline{d}}] &= \frac{1}{2}\text{tr}\left\{\Sigma_{jk,jk}^{-1}S_\sigma\Sigma_{jk,jk}^{-1}\left[\Sigma_{jk,jk} + \Sigma_{jk,\overline{d}}\Sigma_{\overline{d},\overline{d}}^{-1}\times\right.\right. \\ &\quad \left.\left.\times\left((\boldsymbol{y}_{\overline{d}} - \boldsymbol{\mu}_{\overline{d}})(\boldsymbol{y}_{\overline{d}} - \boldsymbol{\mu}_{\overline{d}})' - \Sigma_{\overline{d},\overline{d}}\right)\Sigma_{\overline{d},\overline{d}}^{-1}\Sigma_{\overline{d},jk}\right]\right\}. \qquad \text{(S.39)} \end{aligned}$$

In the special case of two measurements, the first of which always observed, $\overline{d} = 1$ in (S.34), i.e., it refers to the first measurement. Hence, both expectations in (S.34) reduce to 0, implying in turn that then $E_{y^m|y^o}\boldsymbol{U}(y_2|y_1) = E_{y_2|y_1}\boldsymbol{U}(y_2|y_1) = \boldsymbol{0}$, as it should because in this simple case pseudo-likelihood coincides with full likelihood.

For each of the estimators, the sandwich estimator can be computed. For the case of IPWCC and its doubly robust version, Appendix F provides generic expressions.

G.3 Marginal Pseudo-likelihood for Binary Data

Let us assume that we have a model for multivariate and hence also for bivariate binary data. For example, using the notation $\nu_{ij} = P(Y_{ij} = 1)$, $\nu_{ijk} = P(Y_{ij} = 1, Y_{ik} = 1)$, and $\nu_{ik|j}(\ell) = P(Y_{ik} = 1|y_{ij} = \ell)$ $(\ell = 0, 1)$, pairwise Plackett (1965) probabilities take the form

$$\nu_{ijk} = \begin{cases} \frac{1+(\nu_{ij}+\nu_{ik})(\psi_{ijk}-1)-S(\nu_{ik},\nu_{ij},\psi_{ijk})}{2(\psi_{ijk}-1)} & \text{if } \psi_{ijk} \neq 1, \\ \nu_{ij}\nu_{ik} & \text{if } \psi_{ijk} = 1, \end{cases} \tag{S.40}$$

with

$$S(\nu_{ij}, \nu_{ik}, \psi_{ijk}) = \sqrt{[1 + (\nu_{ij} + \nu_{ik})(\psi_{ijk} - 1)]^2 + 4\psi_{ijk}(1 - \psi_{ijk})\nu_{ij}\nu_{ik}}$$

and the pairwise odds ratio, also termed global cross ratio (Dale 1986):

$$\psi_{ijk} = \frac{P(Y_{ij} = 1, Y_{ik} = 1)P(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0)P(Y_{ij} = 0, Y_{ik} = 1)}.$$

When the Bahadur (1961) model is used instead, (S.40) is replaced by

$$\nu_{ijk} = \nu_{ij}\nu_{ik} \left[1 + \rho_{ijk} \frac{1 - \nu_{ij}}{\sqrt{\nu_{ij}(1 - \nu_{ij})}} \cdot \frac{1 - \nu_{ik}}{\sqrt{\nu_{ik}(1 - \nu_{ik})}}\right]. \tag{S.41}$$

In both cases, expressions for the multivariate probabilities exist as well. In the odds ratio case, this leads to the so-called multivariate Dale model (Molenberghs and Lesaffre 1994, Molenberghs and Verbeke 2005). The expressions are implicit and fitting the model is computationally very demanding. The multivariate Bahadur model can be written as $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i) \cdot c(\boldsymbol{y}_i)$, where

$$\begin{aligned} f_1(\boldsymbol{y}_i) &= \prod_{j=1}^{n_i} \nu_{ij}^{y_{ij}}(1 - \nu_{ij})^{1-y_{ij}}, \\ c(\boldsymbol{y}_i) &= 1 + \sum_{j_1<j_2} \rho_{ij_1j_2}e_{ij_1}e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3}e_{ij_1}e_{ij_2}e_{ij_3} + \dots \\ &\quad +\rho_{i12\dots n_i}e_{i1}e_{i2}\dots e_{in_i}, \\ e_{ij} &= \frac{y_{ij} - \nu_{ij}}{\sqrt{\nu_{ij}(1 - \nu_{ij})}}. \end{aligned}$$

Here, the ρ parameters are pairwise and higher-order correlations. Even though the model admits a convenient and concise closed form, its fitting is less than trivial, owing to strong and intractable constraints on the parameter space, be it in fully general or second-order form (where the third- and higher-order correlations are set equal to zero). This makes pseudo-likelihood attractive.

A generic contribution to the pairwise log-likelihood takes the form:

$$\begin{aligned} p\ell_{ijk} &= y_{ij}y_{ik}\ln\nu_{ijk} + y_{ij}(1-y_{ik})\ln(\nu_{ij}-\nu_{ijk}) + (1-y_{ij})y_{ik}\ln(\nu_{ik}-\nu_{ijk}) \\ &\quad +(1-y_{ij})(1-y_{ik})\ln(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}). \end{aligned}$$

As before, let $\boldsymbol{\beta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$, where $\nu_{ij} = \nu_{ij}(\boldsymbol{\beta})$ and the association parameters are functions of $\boldsymbol{\alpha}$. Hence, $\nu_{ijk} = \nu_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Pairwise and conditional contributions to the score take the form:

$$\begin{aligned} \boldsymbol{U}_{ijk} &= \frac{y_{ij}y_{ik}}{\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}\nu_{ijk} + \frac{y_{ij}(1-y_{ik})}{\nu_{ij}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}(\nu_{ij}-\nu_{ijk}) \\ &\quad +\frac{(1-y_{ij})y_{ik}}{\nu_{ik}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}(\nu_{ik}-\nu_{ijk}) \\ &\quad +\frac{(1-y_{ij})(1-y_{ik})}{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}), \end{aligned} \tag{S.42}$$

$$\begin{aligned} \boldsymbol{U}_{ik|j} &= \frac{y_{ij}y_{ik}\nu_{ij}}{\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\nu_{ijk}}{\nu_{ij}}\right) + \frac{y_{ij}(1-y_{ik})\nu_{ij}}{\nu_{ij}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}}\right) \\ &\quad +\frac{(1-y_{ij})y_{ik}(1-\nu_{ij})}{\nu_{ik}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}}\right) \\ &\quad +\frac{(1-y_{ij})(1-y_{ik})(1-\nu_{ij})}{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}}\right). \end{aligned} \tag{S.43}$$

In addition, we need expectations of these over the conditional distribution of the unobserved outcomes given the observed ones. Evidently, because (S.42)–(S.43) are linear in the triplet y_{ij}, y_{ik}, and $y_{ij}y_{ik}$, it suffices to calculate the expectations over these. Their corresponding probabilities are

$$\nu_{ij|\overline{d}} = \frac{\nu_{i\overline{d}j}}{\nu_{i\overline{d}}}, \qquad \nu_{ijk|\overline{d}} = \frac{\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}}. \tag{S.44}$$

Combining (S.42)–(S.43) with (S.44) leads to:

$$\begin{aligned} E\left(\boldsymbol{U}_{ijk}\right) &= \frac{\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}\nu_{ijk} + \frac{\nu_{i\overline{d}j}-\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}(\nu_{ij}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\beta}}(\nu_{ij}-\nu_{ijk}) \\ &\quad +\frac{\nu_{i\overline{d}k}-\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}(\nu_{ik}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\beta}}(\nu_{ik}-\nu_{ijk}) \end{aligned}$$

$$+\frac{\nu_{i\overline{d}}-\nu_{i\overline{d}j}-\nu_{i\overline{d}k}+\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\beta}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}), \qquad \text{(S.45)}$$

$$\begin{aligned}E(\boldsymbol{U}_{ik|j}) &= \frac{y_{ij}\nu_{i\overline{d}k}\nu_{ij}}{\nu_{i\overline{d}}\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\nu_{ijk}}{\nu_{ij}}\right)+\frac{y_{ij}(\nu_{i\overline{d}}-\nu_{i\overline{d}k})\nu_{ij}}{\nu_{i\overline{d}}(\nu_{ij}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}}\right)\\
&+\frac{(1-y_{ij})\nu_{i\overline{d}k}(1-\nu_{ij})}{\nu_{i\overline{d}}(\nu_{ik}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}}\right)\\
&+\frac{(1-y_{ij})(\nu_{i\overline{d}}-\nu_{i\overline{d}k})(1-\nu_{ij})}{\nu_{i\overline{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\beta}}\times\\
&\times\left(\frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}}\right). \qquad \text{(S.46)}\end{aligned}$$

As already mentioned at the end of Section 4.1, all probabilities involving $\overline{d}$ are potentially high-dimensional; they would follow from the multivariate Dale model, the multivariate Bahadur model, etc. We have seen, however, that several alternative routes are open. For example, here, one could simply resort to the singly robust version. Alternatively, the expectations could be replaced by simple, e.g., logistic, models: $E_{\boldsymbol{Y}_i^m|\boldsymbol{y}_i^o}(y_{ij})$ could be written as a standard logistic model where the existing covariates are supplemented with $\boldsymbol{y}_{i\overline{d}}$, whereas for $E_{\boldsymbol{Y}_i^m|\boldsymbol{y}_i^o}(y_{ij}y_{ik})$ the pairwise model under consideration can be used, again supplementing the covariate information with $\boldsymbol{y}_{i\overline{d}}$.

Further, (S.42)–(S.46) require derivatives with respect to the univariate and pairwise probabilities. For most pairwise models, such as the Bahadur and Dale models, they are reasonably straightforward and have been derived by various authors. See Molenberghs and Verbeke (2005) for details.

The derivation of the sandwich estimator follows from logic similar to that laid out in Section G.2.

G.4 Conditional Pseudo-likelihood for Binary Data

Consider a single clustered outcome, such as in the National Toxicology Program Data (Section 5.2) and assume the model (Molenberghs and Ryan 1999, Aerts *et al* 2002, Molenberghs and Verbeke 2005):

$$f_i(\boldsymbol{y}_i;\boldsymbol{\Theta}_i) = \qquad \text{(S.47)}$$

$$\exp\left\{\sum_{j=1}^{n_i}\theta_{ij}y_{ij}+\sum_{j<j'}\delta^*_{ijj'}y_{ij}y_{ij'}+\ldots+\omega_{i1\ldots n_i}y_{i1}\ldots y_{in_i}-A(\mathbf{\Theta}^*_i)\right\}.$$

or its quadratic simplification (Zhao and Prentice 1990, Thélot 1985, Molenberghs and Ryan 1999):

$$f_i(\boldsymbol{y}_i;\mathbf{\Theta}_i{}^*,n_i)=\exp\left\{\sum_{j=1}^{n_i}\theta^*_i y_{ij}+\sum_{j<j'}\delta^*_i y_{ij}y_{ij'}-A(\mathbf{\Theta}^*_i)\right\}, \tag{S.48}$$

with δ^*_i describing the association between pairs of measurements within the ith unit. It is useful to code the outcomes as 1 and -1, rather than 1 and 0, whenever the number of measurements per unit is variable, to ensure coding invariance. Focusing on an exchangeable situation, define the number of measurements from unit i with a positive response to be Z_i. Model (S.48) then becomes, upon absorbing constant terms into the normalizing constant and using the re-parameterization $\theta_i=2\theta^*_i$ and $\xi_i=2\delta^*_i$:

$$f_i(\boldsymbol{y}_i;\mathbf{\Theta}_i,n_i)=\exp\left\{\theta_i z_i^{(1)}+\xi_i z_i^{(2)}-A(\mathbf{\Theta}_i)\right\}, \tag{S.49}$$

with $z_i^{(1)}=z_i$ and $z_i^{(2)}=-z_i(n_i-z_i)$. The normalizing constant takes the form:

$$A(\mathbf{\Theta}_i)=\ln\left[\sum_{k=0}^{n_i}\begin{pmatrix}n_i\\k\end{pmatrix}\exp\left\{\theta_i k^{(1)}+\xi_i k^{(2)}\right\}\right],$$

where $k^{(1)}=k$ and $k^{(2)}=-k(n_i-k)$. For model (S.49), independence corresponds to $\xi_i=0$. A positive δ_i corresponds to classical clustering or overdispersion, whereas a negative parameter value occurs in the under-dispersed case. As such, estimation of the association parameter can be of interest.

Fitting the model is awkward for long sequences, owing to the presence of the normalizing constant. Therefore, it is convenient to replace the corresponding likelihood function by a pseudo-likelihood alternative, found by replacing the joint density $f_i(\boldsymbol{y}_i;\mathbf{\Theta}_i)$ by the product of univariate full conditional densities $f(y_{ij}|\{y_{ij'}\},j'\neq j;\mathbf{\Theta}_i)$ for $j=1,\ldots,n_i$. This idea can be put into the framework (9) by choosing $\delta_{1_{n_i}}=n_i$ and $\delta_{s_j}=-1$ for $j=1,\ldots,n_i$ where $\mathbf{1}_{n_i}$ is a vector of ones and $\boldsymbol{s}_j$ consists of ones everywhere, except for the jth entry. For all other vectors $\boldsymbol{s}$, δ_s equals zero. This pseudo-likelihood has the effect of replacing a joint mass function with a complicated normalizing constant by n_i univariate functions of logistic type.

If we can assume that outcomes within a unit are exchangeable, then there are merely two types of contribution: (1) the conditional probability of an additional success, given there are $z_i - 1$ successes and $n_i - z_i$ failures (this contribution occurs with multiplicity z_i):

$$p_{is} = \frac{\exp\left[\theta_i - \delta_i(n_i - 2z_i + 1)\right]}{1 + \exp\left[\theta_i - \delta_i(n_i - 2z_i + 1)\right]},$$

and (2) the conditional probability of an additional failure, given there are z_i successes and $n_i - z_i - 1$ failures (with multiplicity $n_i - z_i$):

$$p_{if} = \frac{\exp\left[-\theta_i + \delta_i(n_i - 2z_i - 1)\right]}{1 + \exp\left[-\theta_i + \delta_i(n_i - 2z_i - 1)\right]}.$$

The log PL contribution for unit i can then be expressed as

$$p\ell_i = z_i \ln p_{is} + (n_i - z_i) \ln p_{if}. \tag{S.50}$$

The contribution of unit i to the pseudo-likelihood score vector takes the form

$$\left[\begin{array}{c} z_i(1 - p_{is}) - (n_i - z_i)(1 - p_{if}) \\ -z_i(n_i - 2z_i + 1)(1 - p_{is}) + (n_i - z_i)(n_i - 2z_i - 1)(1 - p_{if}) \end{array}\right].$$

Note that, if $\delta_i \equiv 0$, then $p_{is} \equiv 1 - p_{if}$ and the first component of the score vector is a sum of terms $z_i - n_i p_{is}$, i.e., standard logistic regression follows.

Data can be incomplete, for example, because some litter mates die or get resorbed into the uterus line. Let there be m_i litter mates, n_i of which are viable and assessed for success/failure. This then means that (S.50) would pertain to the observed data only, whereas there are an additional $m_i - n_i$ missing outcomes.

The general expressions (21)–(29) now take the form:

$$\begin{aligned}
\boldsymbol{U}_{\text{naive, CC}} &= \sum_{i=1}^{N} R_i \boldsymbol{U}_i(z_i, n_i - z_i) = \sum_{i=1}^{N} R_i \boldsymbol{U}_i(z_i, m_i - z_i), && \text{(S.51)} \\
\boldsymbol{U}_{\text{naive, AC}} &= \sum_{i=1}^{N} \boldsymbol{U}_i(z_i, n_i - z_i), && \text{(S.52)} \\
\boldsymbol{U}_{\text{IPWCC}} &= \sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i(m_i|m_i)} \boldsymbol{U}_i(z_i, n_i - z_i), && \text{(S.53)}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{U}_{\mathrm{IPWAC}} &= \sum_{i=1}^{N} \frac{I(n_i|m_i)}{\pi_i(n_i|m_i)} \boldsymbol{U}_i^o(z_i, n_i, m_i), && \text{(S.54)}\\
\boldsymbol{U}_{\mathrm{IPWCC,dr}} &= \sum_{i=1}^{N} \left\{ \frac{\widetilde{R}_i}{\pi_i(m_i|m_i)} \boldsymbol{U}_i(z_i, n_i - z_i) \right.\\
&\quad \left. + \left[1 - \frac{\widetilde{R}_i}{\pi_i(m_i|m_i)}\right] E_{k|z_i,n_i}\left[\boldsymbol{U}_i(z_i + k, m_i - z_i - k)\right] \right\} && \text{(S.55)}\\
\boldsymbol{U}_{\mathrm{IPWAC,dr}} &= \sum_{i=1}^{N} \left\{ \frac{I(n_i|m_i)}{\pi_i(n_i|m_i)} \boldsymbol{U}_i^o(z_i, n_i, m_i) \right.\\
&\quad \left. + \left[1 - \frac{I(n_i|m_i)}{\pi_i(m_i|m_i)}\right] E_{k|z_i,n_i}\left[\boldsymbol{U}_i(z_i + k, m_i - z_i - k)\right] \right\} && \text{(S.56)}
\end{aligned}$$

Here, R_i is the usual indicator for a complete cluster, and $I(n_i|m_i)$ is an indicator for observing n_i out of m_i litter mates. Furthermore, $\pi_i(n_i|m_i)$ is the probability of observing n_i out of m_i litter mates. Evidently, $\pi(m_i|m_i)$ is the special case of observing a complete cluster. Result (S.56) follows from observing that the observed version of the score and the expectation over the incomplete data follow, in this case, in exactly the same way.

The quantity $\boldsymbol{U}_i^o(z_i, n_i, m_i)$ in (S.54) and (S.56) follows from

$$p\ell_i^o = \ln \left\{ \sum_{k=0}^{m_i - n_i} \begin{pmatrix} m_i - n_i \\ k \end{pmatrix} p_{is}(z_i, k)^{z_i + k} [1 - p_{if}(z_i, k)]^{m_i - z_i - k} \right\}, \qquad \text{(S.57)}$$

and then constructing

$$\boldsymbol{U}_i^o = \frac{\partial p\ell_i^o}{\partial(\theta_i, \delta_i)}, \qquad \text{(S.58)}$$

where

$$\begin{aligned}
\mathrm{logit}[p_{is}(z_i, k)] &= \theta_i - \delta_i[m_i - 2(z_i + k) + 1],\\
\mathrm{logit}[p_{if}(z_i, k)] &= -\theta_i + \delta_i[m_i - 2(z_i + k) - 1].
\end{aligned}$$

Note the difference between (S.50) and (S.57). In the former only the observed data are included, while in the latter there is summation over the missing outcomes.

In the NTP data, especially for the higher dose groups, complete clusters may be rare, thence the AC versions become not only attractive, but actually necessary to make progress.

Overall, the AC forms are slightly more cumbersome, owing to somewhat less tractable expressions, such as (S.57). Consider full exchangeability, whence form (30) can be used, we obtain:

$$\boldsymbol{U}_{\text{IPWAC,exch}} = \sum_{i=1}^{N} \boldsymbol{U}_i^o(z_i, n_i, m_i). \tag{S.59}$$

Even though the missing-data mechanism is removed, as follows from (30) in general, construction (S.57)–(S.58) needs to be used. This is different from the pairwise likelihood case, thanks to the marginal specification of the latter. Of course, (S.59) can be used with a numerical optimizer or equation solver, thanks to the explicit expression (S.57).

Now, using (S.49), the expectations can be written as:

$$E_{k|z_i,n_i}\left[\boldsymbol{U}_i(z_i+k, m_i-z_i-k)\right] = \frac{\sum_{k=0}^{m_i-n_i} e^{\theta_i k - \delta_i k(m_i-2z_i-k)} \boldsymbol{U}_i(z_i+k, m_i-z_i-k)}{\sum_{k=0}^{m_i-n_i} e^{\theta_i k - \delta_i k(m_i-2z_i-k)}}.$$

To formulate a sensible missingness model in this case, write the individual responses as $(y_{i1}, \ldots, y_{in_i}, y_{i,n_i+1}, \ldots, y_{im_i})$, with the first n_i observed and the later $m_i - n_i$ missing. Likewise, the missingness indicators are $(r_{i1}, \ldots, r_{in_i}, r_{i,n_i+1}, \ldots, r_{im_i})$, the first set being 1 and the second part 0. Let x_i indicate the dose administered to litter i. Now, the joint distribution of $\boldsymbol{Y}_i$ and $\boldsymbol{R}_i$ factors as

$$\begin{aligned} & f(y_{i1}, \ldots, y_{in_i}, y_{i,n_i+1}, \ldots, y_{im_i}|x_i)\times \\ & \qquad \times f(r_{i1}, \ldots, r_{in_i}, r_{i,n_i+1}, \ldots, r_{im_i}|y_{i1}, \ldots, y_{in_i}, y_{i,n_i+1}, \ldots, y_{im_i}, x_i). \end{aligned}$$

Here, the first factor is the one for which pseudo-likelihood is considered, whereas the second one can be written in summary-statistics form, thanks to exchangeability: $f(n_i, m_i - n_i|z_i, n_i - z_i, x_i)$. To explicitly acknowledge within-cluster correlation, a beta-binomial model (Skellam 1948, Kleinman 1973, Molenberghs and Verbeke 2005), for example, would be a reasonable choice:

$$p_i = \frac{B[n_i + \nu_i(\rho^{-1}-1), m_i - n_i + (1-\nu_i)(\rho^{-1}-1)]}{B[\nu_i(\rho^{-1}-1), (1-\nu_i)(\rho^{-1}-1)]}, \tag{S.60}$$

in terms of the mean parameter ν_i and correlation ρ, and then

$$f_i(n_i, m_i - n_i|\nu_i, \rho) = \begin{pmatrix} m_i \\ n_i \end{pmatrix} p_i^{m_i - n_i}(1 - p_i)^{n_i}. \tag{S.61}$$

Here, $B(\cdot, \cdot)$ is the beta function. One might write, for example:

$$\text{logit}(\nu_i) = \psi_0 + \psi_1 n_i + \psi_2(z_i/n_i). \tag{S.62}$$

Fitting the model and other manipulation is straightforward (Molenberghs and Verbeke 2005), even though it is not commonly implemented in standard statistical software. Alternatively, one might choose to simplify matters and simply replace (S.60) by a logistic regression, in which case (S.61) and (S.62) would be retained.

For the sandwich estimator, take for example $IPWCC$, which can be written in shorthand as

$$\boldsymbol{U}_{\text{IPWCC}} = \sum_{i=1}^{N} V_i = \sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \boldsymbol{U}_i.$$

Then,

$$\frac{\partial \boldsymbol{V}_i}{\partial(\theta, \delta)} = \frac{\widetilde{R}_i}{\pi_i} Q_i, \qquad \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} = -\frac{\widetilde{R}_i}{\pi_i^2} \frac{\partial \pi_i}{\partial \boldsymbol{\psi}} \boldsymbol{U}_i.$$

Here, Q_i has elements:

$$\begin{aligned} q_{i,11} &= -z_i p_{is}(1 - p_{is}) - (n_i - z_i) p_{if}(1 - p_{if}), \\ q_{i,12} = q_{i,21} &= z_i(n_i - 2z_i + 1) p_{is}(1 - p_{is}) + (n_i - z_i)(n_i - 2x_i - 1) p_{if}(1 - p_{if}), \\ q_{i,22} &= -z_i(n_i - 2z_i + 1)^2 p_{is}(1 - p_{is}) \\ &\quad -(n_i - z_i)(n_i - 2z_i - 1)^2 p_{if}(1 - p_{if}). \end{aligned}$$

The derivative w.r.t. ψ evidently depends on whether the beta-binomial model, or rather simpler logistic regression is chosen. Finally, let $\boldsymbol{W}_i$ be the beta-binomial score equation contribution of litter i. From this, the derivative $\partial \boldsymbol{W}_i / \partial \boldsymbol{\psi}$ follows immediately. For the other forms, similar calculations apply.

References

Ashford, J.R. and Sowden, R.R. (1970) Multivariate probit analysis. *Biometrics*, **26**, 535–546.

Bahadur, R.R. (1961) A representation of the joint distribution of responses of n dichotomous items. In: *Studies in Item Analysis and Prediction*, H. Solomon (Ed.), Stanford Mathematical Studies in the Social Sciences VI. Stanford, California, Stanford University Press.

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

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

Fitzmaurice, G.M., Molenberghs, G., and Lipsitz, S.R. (1995) Regression models for longitudinal binary responses with informative dropouts. *Journal of the Royal Statistical Society, Series B*, **57**, 691–704.

Hartley, H.O. and Hocking, R. (1971) The analysis of incomplete data. *Biometrics*, **27**, 7783–808.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003) A Local Influence Approach applied to Binary Data from a Psychiatric Study. *Biometrics*, **59**, 410–419.

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.

Kenward, M.G, Goetghebeur, E.J.T., and Molenberghs, G. (2001) Sensitivity analysis of incomplete categorical data. *Statistical Modelling*, **1**, 31–48.

Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association*, **68**, 46–54.

Lang, J.B. and Agresti, A. (1994) Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**, 625–632.

Lipsitz, S.R., Laird, N.M., and Harrington, D.P. (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153–160.

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

Molenberghs, G. and Lesaffre, E. (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998) Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001) Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93–113.

Plackett, R.L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Renard, D., Molenberghs, G., and Geys, H. (2004) A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, **44**, 649–667.

Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. (1998) Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, **93**, 1321–1339.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

Thélot, C. (1985). Lois logistiques à deux dimensions. *Annales de l'Insée*, **58**, 123–149.

Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.