First-order marginalised transition random effects models with probit link function

Özgür Asar [a,*] , Ozlem Ilk [b]

[a] CHICAS, Medical School, Lancaster University, UK
[b] Department of Statistics, Middle East Technical University, Turkey

July 10, 2015

Abstract

Marginalised models, also known as marginally specified models, have recently become a popular tool for analysis of discrete longitudinal data. Despite being a novel statistical methodology, these models introduce complex constraint equations and model fitting algorithms. On the other hand, there is a lack of publicly available software to fit these models. In this paper, we propose a three-level marginalised model for analysis of multivariate longitudinal binary outcome. The implicit function theorem is introduced to approximately solve the marginal constraint equations explicitly. *probit* link enables direct solutions to the convolution equations. Parameters are estimated by maximum likelihood via a Fisher-Scoring algorithm. A simulation study is conducted to examine the finite-sample properties of the estimator. We illustrate the model with an application to the data set from the Iowa Youth and Families Project. The R package `pnmtrem` is prepared to fit the model.

Keywords: correlated data, implicit differentiation, link functions, maximum likelihood estimation, subject-specific inference, statistical software.

1 Introduction

Longitudinal data comprise repeated measurements on the same subjects across time. Whilst data from the same subjects are typically dependent on each other, data from different subjects are typically independent. Often, multiple responses, e.g. multiple health outcomes or distress variables, from each subject are collected. These responses introduce two types of dependencies: 1) within-response (serial) dependence, and 2) multivariate response dependence at a given time point. To draw valid statistical inferences, both of these dependencies should be taken into account.

Conventional models for analysis of longitudinal data are marginal, transition and random effects models (Diggle et al., 2002). A recently popular method for discrete longitudinal data analysis is the framework of marginalised models, also known as marginally specified models. The framework typically combines the underlying features of the conventional models, and enables likelihood-based

*Corresponding author: Özgür Asar, E-mail address: o.asar@lancaster.ac.uk, Tel.: +44 (0) 1524 593519.

inference for marginal mean parameters. Heagerty and Zeger (2000) define marginally specified models as a re-parameterised version of transition and/or random effects models in terms of the marginal mean and additional dependence parameters. Heagerty (1999, 2002), in his seminal papers, develops marginalised random effects and marginalised transition models, respectively. Both of these models are two-level logistic regression models. Whilst covariate effects are captured in the first levels, serial dependence is captured in the second levels via random effects and response history, respectively. Heagerty and Kurland (2001) show that marginal regression parameter estimates based on marginalised random effects models are less sensitive to dependence structure misspecification compared to those based on conventional random effects models. Heagerty (2002) and Lee and Mercante (2010) prove that parameters of the first and second levels of marginalised transition models are orthogonal. The marginalised modeling paradigm was primarily developed for binary data (Schildcrout and Heagerty, 2007; Ilk and Daniels, 2007; Lee et al., 2009; along with the aforementioned works of Heagerty). Later, it has been extended to ordinal (Caffo and Griswold, 2006; Lee and Daniels, 2007; Lee et al., 2013), count (Lee et al., 2011; Iddi and Molenberghs, 2012) and nominal data (Lee and Mercante, 2010). Amongst these works, Ilk and Daniels (2007) propose a three-level marginalised model for *multivariate longitudinal binary data*, called marginalised transition random effects model. With this model, whilst covariate effects are captured in the first level, serial and multivariate response dependencies are captured in the second and third levels via response history and random effects, respectively. In this paper, we extend marginalised transition random effects model in terms of link function, from *logit* to *probit*, and the parameter estimation methodology, from *Bayesian methods* (BM) to *maximum likelihood* (ML) estimation.

probit and *logit* are popular link functions for modelling categorical data. These link functions are defined as the inverses of the distribution functions of the standard normal and the standard logistic distribution, respectively. They have similar behaviours in terms of placing probabilities. The only difference is at the extreme tails; *logit* places higher probabilities at the tails (Hedeker and Gibbons, 2006). Nonetheless, substantial and high quality data are needed to detect the difference (Doksum and Gakso, 1990, cited in Hedeker and Gibbons, 2006, pp. 153). *logit* allows direct interpretation of the parameter estimates, as changes in (log) odds ratios. The interpretation is more challenging with *probit*. Nonetheless, (approximate) transitions between the parameter estimates based on these link functions is possible (Agresti, 2002; Griswold et al., 2013). For example, the JKB constant (Johnson et al., 1995, pp. 113-163, cited in Griswold et al., 2013) postulates the following: $\beta_{logit} \cong c * \beta_{probit}$ where $c = (15/16)(\pi/\sqrt{3}) \cong 1.700437$. One advantage of *probit* over *logit* is that it allows explicit form of the linkage between the levels of marginalised random effects models (Heagery and Zeger, 2000; Griswold et al., 2013; Caffo and Griswold, 2006). The use of *probit* link in multivariate modelling dates back to Ashford and Sowden (1970). Some recent examples on longitudinal mixed modelling are Hedeker and Gibbons (2006), Liu and Hedeker (2006), Varin and Czado (2010), amongst others.

Generalized estimating equations (GEE; Liang and Zeger, 1986) have been widely used to estimate the parameters of marginal models, especially for discrete outcome. Nonetheless, they might be inefficient because of being a semi-parametric method, compared to the full likelihood-based methods, e.g. ML and BM. BM are widely used in longitudinal data literature and have their own properties. Some distinguishing features of ML over BM are that parameter estimation requires less computational times, and related procedures are more automatised (Efron, 1986). In this paper, we consider ML for parameter estimation to avoid the computational burden.

Marginalised models with transition structures require solving marginal constraint equations (Heagerty, 2002; Schildcrout and Heagerty, 2007; Ilk and Daniels, 2007; Lee and Mercante, 2010). Common literature for solving these equations has been built on optimisation methods, e.g. Newton-Raphson (N-R) algorithm. This might be computationally cumbersome and might yield convergence problems. In this paper, we consider approximately explicit solutions of marginal constraint equations, and propose the use of the *implicit function theorem* for the first time in the scope of marginally specified models.

Publicly available software for analysis of multivariate longitudinal binary data is still rare. Available options include the SAS macro of Shelton et al. (2004), and the `R` (R Core Development Team, 2015) packages `mmm` (Asar and Ilk, 2013) and `mmm2` (Asar and Ilk, 2014). In this study, we propose the R package `pnmtrem` for first-order marginalised transition random effects models with probit link. The package is available from the Comprehensive R Archive Network (CRAN) at

`http://CRAN.R-project.org/package=pnmtrem`.

The paper is organized as follows. Whilst the general modelling framework is introduced in Section 2, first-order version is discussed in detail in Section 3. In Section 4, we discuss inference for the first-order model. Finite-sample behaviours of the estimator are investigated by a simulation study in Section 5. The first-order model is applied to a real data set in Section 6. Section 7 is a concluding discussion.

2 General framework

Let Y_{itj} denote the jth ($j = 1, \ldots, k$) response of the ith ($i = 1, \ldots, n$) subject at time t ($t = 1, \ldots, T$). Also let $\boldsymbol{X}_{itj}$ denote the associated set of covariates, which might include time-varying and/or time-invariant covariates. The framework of the general model with inverse *probit* link is as follows:

$$P^m_{itj} \equiv P(Y_{itj} = 1|\boldsymbol{X}_{itj}) = \Phi(\boldsymbol{X}_{itj}\boldsymbol{\beta}), \tag{1}$$

$$P^t_{itj} \equiv P(Y_{itj} = 1|y_{i,t-1,j}, .., y_{i,t-p,j}, \boldsymbol{X}_{itj}) = \Phi(\Delta_{itj} + \sum_{m=1}^{p} \gamma_{itj,m} y_{i,t-m,j}), \tag{2}$$

$$P^r_{itj} \equiv P(Y_{itj} = 1|y_{i,t-1,j}, ..., y_{i,t-p,j}, \boldsymbol{X}_{itj}, b_{it}) = \Phi(\Delta^*_{itj} + \lambda_j b_{it}), \tag{3}$$

where $\Phi(\cdot)$ is the distribution function of the standard normal.

In (1), the first level of the framework, $\boldsymbol{\beta}$ are marginal regression parameters. These parameters measure the relationship between covariates and responses, and allow comparing covariate subgroups, e.g. males vs. females, without conditioning on response history and/or random effects. The default setting assumes that intercepts and slopes are shared by different responses, i.e. we postulate $\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}_j$. Nonetheless, one is able to specify different intercepts and slopes for multiple responses by including in X_{itj} indicator variables for responses and interactions of these indicator variables with covariates, respectively. This specification provides model flexibility. We might gain in efficiencies considerably, e.g. when the relationships between covariates and multiple responses are not significantly different (Asar and Ilk, 2014). Another default setting is the assumption of accommodating only the relationship of responses with current covariates, i.e. $P(Y_{itj} = 1|\boldsymbol{X}_{i1j}, \ldots, \boldsymbol{X}_{itj}) = P(Y_{itj} = 1|\boldsymbol{X}_{itj})$. Nonetheless, relationships with lagged covariates might be captured by including covariate history in X_{itj}.

In (2), the second level of the framework, Markov model of order p is used to capture the serial dependence. Here, the mth transition parameters, $\gamma_{itj,m}$, can be written in terms of covariates, i.e. $\gamma_{itj,m} = \boldsymbol{\alpha}_{t,m}\boldsymbol{Z}_{itj,m} = \alpha_{t1,m}Z_{itj1,m} + \ldots + \alpha_{tl,m}Z_{itjl,m}$ for $m = 1, \ldots, p$. $\alpha_{tf,m}$ ($f = 1, \ldots, l$) are time, covariate and order specific transition parameters. They capture the relationships between past and current responses. $\boldsymbol{Z}_{itj,.}$ have a form of design matrix with 1's on the first column, and are typically a subset of X_{itj} with l covariates. The form of $\boldsymbol{Z}_{itj,.}$ permits flexibly specifying the association structures between past and current responses. For example, if one suspects that the lag-1 associations are different for males and females, then gender can be included in $\boldsymbol{Z}_{itj,1}$. Similar to the first level, $\boldsymbol{\alpha}_{t,m}$ are assumed to be shared across multiple responses. Response-specific transition parameters can be specified by including in $\boldsymbol{Z}_{itj,.}$ indicator variables for responses and interactions of these indicator variables with response history and other covariates.

In (3), the third level of the model, multivariate response dependence and individual variations are captured. b_{it}'s are subject and time specific random effects coefficients that measure unobserved heterogeneity between subjects at time t. We assume that $b_{it} \sim$ N(0, σ_t^2). b_{it} can be rewritten as b_{it}=σ_t z_i, where z_i is a standard normal random variable, which is useful in numerical integration. λ_j's are response-specific parameters that scale b_{it} with respect to the jth response, and accounts for multivariate response dependence. We set λ_1 to 1 for identifiability, and estimate λ_j for $j = 2, \ldots$. Note that by specifying b_{it}'s are time-varying, the model assumes that multivariate response dependencies might change across time.

Δ_{itj}'s in (2) are subject, time and response specific intercepts. They take into account the (non-linear) relationship between marginal (P^m_{itj}) and transition probabilities (P^t_{itj}). Similarly, Δ^*_{itj}'s in (3) are subject, time and response specific intercepts that account for the (non-linear) relationship between transition and random effects probabilities (P^r_{itj}).

We assume that conditional mean of responses given all covariates is equal to conditional mean of responses given covariate history, i.e., $E(Y_{itj}|\boldsymbol{X}_{iqj}, q = 1, \ldots, T) = E(Y_{itj}|\boldsymbol{X}_{isj}, s \leq t)$. The assumption is meaningful only for exogenous covariates (covariates that do not depend on response history), but not for endogenous ones (covariates that depend on response history at time t). It is necessary for the validity of the marginal constraint equation, to be introduced later.

3 First-order model

In this study, we focus on lag-1 dependence in (2). The framework for the first-order model becomes

$$P^m_{itj} \equiv P(Y_{itj} = 1|\boldsymbol{X}_{itj}) = \Phi(\boldsymbol{X}_{itj}\boldsymbol{\beta}), \tag{4}$$
$$P^t_{itj} \equiv P(Y_{itj} = 1|y_{i,t-1,j}, \boldsymbol{X}_{itj}) = \Phi(\Delta_{itj} + \gamma_{itj,1} y_{i,t-1,j}), \tag{5}$$
$$P^r_{itj} \equiv P(Y_{itj} = 1|y_{i,t-1,j}, \boldsymbol{X}_{itj}, b_{it}) = \Phi(\Delta^*_{itj} + \lambda_j b_{it}). \tag{6}$$

As before, $b_{it} \sim$ N(0, σ_t^2) and b_{it}=z_i σ_t, $z_i \sim$ N(0,1); λ_1=1; $\gamma_{itj,1} = \boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj,1} = \alpha_{t1,1}Z_{itj1,1} + \ldots + \alpha_{tl,1}Z_{itjl,1}$. Throughout, we call this framework as $t \geq 2$ *model*.

(5) is not valid at baseline ($t = 1$), because there is no history data are available at this time point. Based on this and the assumption that variabilities at $t = 1$ and $t \geq 1$ might be different, we postulate a separate model for $t = 1$:

$$P^m_{i1j} \equiv P(Y_{i1j} = 1|\boldsymbol{X}_{i1j}) = \Phi(\boldsymbol{X}_{i1j}\boldsymbol{\beta}^*), \tag{7}$$
$$P^r_{i1j} \equiv P(Y_{i1j} = 1|\boldsymbol{X}_{i1j}, b_{i1}) = \Phi(\Delta^*_{i1j} + \lambda^*_j b_{i1}), \tag{8}$$

where $b_{i1} \sim$ N(0, σ_1^2) and b_{i1}=z_i σ_1, $z_i \sim$ N(0,1); λ^*_1=1. Throughout we call this model as the *baseline model*.

3.1 Linking levels of the $t \geq 2$ model

Levels of the first-order model (4 - 6) are connected to each other for the model being a valid probabilistic model.

3.1.1 Linking first and second levels

Levels 1 (4) and 2 (5) are linked via the following marginal constraint equation,

$$P(Y_{itj}=1|\boldsymbol{X}_{itj}) = \sum_{y_{i,t-1,j}=0}^{1} P(Y_{itj}=1|y_{i,t-1,j},\boldsymbol{X}_{itj})P(y_{i,t-1,j}|\boldsymbol{X}_{i,t-1,j}), \tag{9}$$

which is equivalent to

$$\Phi(\boldsymbol{X}_{i2j}\boldsymbol{\beta}) = \Phi(\Delta_{i2j})(1-\Phi(\boldsymbol{X}_{i,1,j}\boldsymbol{\beta}^*)) + \Phi(\Delta_{i2j}+\gamma_{i2j,1})\Phi(\boldsymbol{X}_{i,1,j}\boldsymbol{\beta}^*), \tag{10}$$

and

$$\Phi(\boldsymbol{X}_{itj}\boldsymbol{\beta}) = \Phi(\Delta_{itj})(1-\Phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta})) + \Phi(\Delta_{itj}+\gamma_{itj,1})\Phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta}). \tag{11}$$

for $t>2$ and $t=2$, respectively. Hereafter, the discussion will be based on (11). We take the difference between (11) and (10) when necessary.

Since (11) does not permit explicitly writing Δ_{itj} in terms of $\boldsymbol{\beta}$ and $\gamma_{itj,1}$ (or $\boldsymbol{\alpha}_{t,1}$), we use the implicit function theorem (IFT; Krantz and Parks, 2003) for an approximately explicit solution. Application of IFT is as follows.

Let F be a function of $\boldsymbol{X}_{itj}$, $\boldsymbol{X}_{it-1j}$, $\boldsymbol{\beta}$, Δ_{itj}, $\boldsymbol{\alpha}_{t,1}$ and $\boldsymbol{Z}_{itj,1}$ such that (by rewriting (11))

$$F(\boldsymbol{X}_{itj},\boldsymbol{X}_{it-1j},\boldsymbol{\beta},\Delta_{itj},\boldsymbol{\alpha}_{t,1},\boldsymbol{Z}_{itj,1}) = \Phi(\boldsymbol{X}_{itj}\boldsymbol{\beta})-\Phi(\Delta_{itj})(1-\Phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta}))-\Phi(\Delta_{itj}+\boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj,1})\Phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta}) = 0. \tag{12}$$

By IFT with first order implicit differentiation, i.e. first order approximation, Δ_{itj} can be obtained as

$$\Delta_{itj} = -\frac{\left.\frac{\partial F}{\partial \boldsymbol{\beta}}\right|_{(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_{t,10},\Delta_{itj0})}}{\left.\frac{\partial F}{\partial \Delta_{itj}}\right|_{(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_{t,10},\Delta_{itj0})}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \frac{\left.\frac{\partial F}{\partial \boldsymbol{\alpha}_{t,1}}\right|_{(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_{t,10},\Delta_{itj0})}}{\left.\frac{\partial F}{\partial \Delta_{itj}}\right|_{(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_{t,10},\Delta_{itj0})}}(\boldsymbol{\alpha}_{t,1}-\boldsymbol{\alpha}_{t,10}), \tag{13}$$

where

$$\begin{aligned}
\frac{\partial F}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}_{itj}\phi(\boldsymbol{X}_{itj}\boldsymbol{\beta}) + \Phi(\Delta_{itj})(\phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta}))\boldsymbol{X}_{i,t-1,j} - \Phi(\Delta_{itj}+\boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj,1})\phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta})\boldsymbol{X}_{i,t-1,j},\\
\frac{\partial F}{\partial \Delta_{itj}} &= -\phi(\Delta_{itj})(1-\Phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta})) - \phi(\Delta_{itj}+\boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj,1})(\Phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta})),\\
\frac{\partial F}{\partial \boldsymbol{\alpha}_{t,1}} &= -\phi(\Delta_{itj}+\boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj,1})\Phi(\boldsymbol{X}_{i,t-1,j}\boldsymbol{\beta})\boldsymbol{Z}_{itj,1}.
\end{aligned} \tag{14}$$

Here, $\phi(\cdot)$ is the density function of the standard normal, and $\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{t,10}$ and Δ_{itj0} are fixed values around which IFT searches for solution. We set $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_{t,10}$ to $\mathbf{0}$, since null hypotheses for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_{t,1}$ are on equality of these parameters to $\mathbf{0}$. Δ_{itj0} is obtained by solving (12) under $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_{t,10}$ being $\mathbf{0}$. This yields $\Delta_{itj0}=0$ for $t>2$. N-R is used to obtain Δ_{i2j0}. Based on our experience, this has very fast convergence due to the simple form of (12).

3.1.2 Linking second and third levels

Level 2 (5) and level 3 (6) are linked via the following convolution equation:

$$P(Y_{itj}=1|y_{i,t-1,j},\boldsymbol{X}_{itj}) = \int P(Y_{itj}=1|y_{i,t-1,j},\boldsymbol{X}_{itj},b_{it})dF(b_{it}), \tag{15}$$

which is equivalent to

$$\Phi(\Delta_{itj} + \boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj,1}y_{i,t-1,j}) = \int \Phi(\Delta^*_{itj} + \lambda_j b_{it}) f(b_{it}) db_{it}. \tag{16}$$

Following Griswold (2005), we can explicitly obtain Δ^*_{itj} as

$$\Delta^*_{itj} = \sqrt{1 + \lambda_j^2 \sigma_t^2}\ (\Delta_{itj} + \boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj,1}y_{i,t-1,j}). \tag{17}$$

Proof of (17) is given in Appendix A. Δ^*_{itj} is now an explicit and deterministic function of Δ_{itj} (hence, Δ^*_{itj} is a function of $\boldsymbol{X}_{itj}$, $\boldsymbol{X}_{i,t-1,j}$ and $\boldsymbol{\beta}$), $\boldsymbol{\alpha}_{t,1}$ $\boldsymbol{Z}_{itj,1}$, y_{it-1j}, λ_j and σ_t.

3.2 Linking levels of the baseline model

First (7) and second (8) levels of the baseline model are linked via the following convolution equation:

$$P(Y_{i1j} = 1|\boldsymbol{X}_{i1j}) = \int P(Y_{i1j} = 1|\boldsymbol{X}_{i1j}, b_{i1}) dF(b_{i1}). \tag{18}$$

Δ^*_{i1j} can be written as an explicit function of $\boldsymbol{X}_{i1j}$, $\boldsymbol{\beta}^*$, λ^*_j and σ_1 such that

$$\Delta^*_{i1j} = \sqrt{1 + \lambda_j^{*2} \sigma_1^2}\ \boldsymbol{X}_{i1j}\boldsymbol{\beta}^*. \tag{19}$$

Proof of (19) can be easily adapted from the proof of (17).

4 Inference

4.1 Estimation

The likelihood of the first-order model is the product of the likelihood functions of the baseline and $t \geq 2$ models. By re-writing the random effects coefficients as $b_{i1} = \sigma_1 z_i$ and $b_{it} = \sigma_t z_i$, it can be expressed as

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = L_1(\boldsymbol{\theta}_1|\boldsymbol{y}_1) L_2(\boldsymbol{\theta}_2|\boldsymbol{y}_2), \tag{20}$$

where

$$L_1(\boldsymbol{\theta}_1|\boldsymbol{y}_1) = \prod_{i=1}^{N} \int \prod_{j=1}^{k} \left(P^r_{i1j}\right)^{y_{i1j}} \left(1 - P^r_{i1j}\right)^{1-y_{i1j}} \phi(z_i) dz_i, \tag{21}$$

$$L_2(\boldsymbol{\theta}_2|\boldsymbol{y}_2) = \prod_{i=1}^{N} \prod_{t=2}^{T} \int \prod_{j=1}^{k} \left(P^r_{itj}\right)^{y_{itj}} \left(1 - P^r_{itj}\right)^{1-y_{itj}} \phi(z_i) dz_i. \tag{22}$$

Here, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^*, \boldsymbol{\lambda}^*, \sigma_1^2)$ with $\boldsymbol{\lambda}^* = (\lambda^*_2, \ldots, \lambda^*_k)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}, \boldsymbol{\alpha}_{t,1}, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2)$ with $\boldsymbol{\lambda} = (\lambda_2, \ldots, \lambda_k)$ and $\boldsymbol{\sigma}^2 = (\sigma_2^2, \ldots, \sigma_T^2)$, are parameters of the baseline and $t \geq 2$ models, respectively; $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are observed responses at baseline and $t \geq 2$ time points, respectively. $L_1(\boldsymbol{\theta}_1|\boldsymbol{y}_1)$ and $L_2(\boldsymbol{\theta}_2|\boldsymbol{y}_2)$ are connected to each other via $\boldsymbol{\beta}^*$ at $t = 2$ (see (10)). We model $log(\sigma_t)$, instead of σ_t or σ_t^2, due to computational aspects. This transformation helps extending the parameter spaces from $[0, +\infty)$ to $(-\infty, +\infty)$. Estimates and standard errors regarding σ_t or σ_t^2 can be easily obtained using the invariance property of ML estimates and delta method, respectively.

We need to use numerical methods to solve the integrals in (21) and (22), since there is no closed-form solutions. Since these integrals are one-dimensional, we use Gauss-Hermite quadrature with 20-points (Lesaffre and Spiessens, 2001; Agresti, 2002; McCulloch et al., 2008). Similarly, the closed-form solutions based on the first partial derivatives of the log-likelihood are not available. We use Fisher-Scoring (F-S) algorithm to obtain the parameter estimates iteratively. An advantage of F-S algorithm is that it only works with the first partial derivates and does not require the second partial derivatives (Hedeker and Gibbons, 2006, pp. 162-165). Another advantage of the algorithm is that at convergence, inverse of the expected information matrix is a consistent estimator of the large sample variance-covariance matrix of the model parameters. With F-S algorithm, ML estimates are obtained iteratively as

$$\boldsymbol{\theta}_s^{(m+1)} = \boldsymbol{\theta}_s^m + I(\boldsymbol{\theta}_s^m)^{-1} \frac{\partial log\,(L_s(\boldsymbol{\theta}_s^m|\boldsymbol{y}_s))}{\partial \boldsymbol{\theta}_s^m}, \tag{23}$$

where $\boldsymbol{s} = (1,2)$; $s = 1$ corresponds to the baseline model and $s = 2$ corresponds to the $t \geq 2$ model; m represents the F-S step and $I(\boldsymbol{\theta}_s)$ is an empirical information matrix; $\frac{\partial log\left(L_s(\boldsymbol{\theta}_s|\boldsymbol{y}_s)\right)}{\partial \boldsymbol{\theta}_s}$ is the first partial derivative of the log-likelihood, calculations of which can be found in Appendix B. $I(\boldsymbol{\theta}_s)$ is calculated as

$$I(\boldsymbol{\theta}_1) = \sum_{i=1}^{N} h(Y_{i1j}|\boldsymbol{\theta}_1)^{-2} \left(\frac{\partial h(Y_{i1j}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1}\right)\left(\frac{\partial h(Y_{i1j}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1}\right)^T \tag{24}$$

and

$$I(\boldsymbol{\theta}_2) = \sum_{i=1}^{N} \left(\sum_{t=2}^{T} \frac{1}{h(Y_{itj}|\boldsymbol{\theta}_2)} \frac{\partial h(Y_{itj}|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2}\right)\left(\sum_{t=2}^{T} \frac{1}{h(Y_{itj}|\boldsymbol{\theta}_2)} \frac{\partial h(Y_{itj}|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2}\right)^T. \tag{25}$$

Details of $h(Y_{itj}|\boldsymbol{\theta}_s)$ and $\frac{\partial h(Y_{itj}|\boldsymbol{\theta}_s)}{\partial \boldsymbol{\theta}_s}$ for $t = 1, \ldots, T$ can be found in Appendix B. Since σ_1 is time-specific and λ_j^* is response-specific for baseline, and σ_t and $\boldsymbol{\alpha}_{t,1}$ are time-specific and λ_j is response-specific for $t \geq 2$, the calculations of $I(\boldsymbol{\theta}_1)$ and $I(\boldsymbol{\theta}_2)$ for these parameters are different compared to the calculations for $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$. Details can be found in the online supplementary material.

4.2 Prediction

Predicting $b_{it} = \sigma_t z_i$ ($t = 1, \ldots, T$) is equivalent to predicting z_i. We obtain the predictions of z_i's as the modes of log-conditional distributions of z_i's given the data (Heagerty, 1999). This requires solving

$$\left\{\sum_{t=1}^{T}\sum_{j=1}^{k} \frac{\hat{\lambda}_j \hat{\sigma}_t \phi(\hat{d}_{itj})\left(Y_{itj} - \Phi(\hat{d}_{itj})\right)}{\Phi(\hat{d}_{itj})\left(1 - \Phi(\hat{d}_{itj})\right)}\right\} - z_i = 0, \tag{26}$$

with respect to z_i usin N-R algorithm. Here, $\hat{d}_{itj} = \hat{\Delta}_{itj}^* + \hat{\lambda}_j \hat{\sigma}_t z_i$. $\hat{\Delta}_{itj}^*$'s are obtained as in (17) and (19) plugging-in the ML estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

5 Simulation study

We conduct a Monte Carlo simulation study to examine the finite-sample behavoiurs of the marginal mean parameters. In each replications, we simulate data using the first-order model. The data sets include bivariate binary responses, Y_1 and Y_2, and two associated covariates, X_1 and X_2, for 250 subjects with 4 follow-ups. We generate X_1 from Uniform$(0,1)$ as a time-independent variable. X_2 is taken as the response indicator variable. It takes 1 for Y_1, 0 for Y_2. We consider different sets of covariates for baseline and $t \geq 2$. Moreover, we consider varying relationships between the covariates and the responses, i.e. $\boldsymbol{\beta}^* \not\equiv \boldsymbol{\beta}$. We specifically consider $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*) = (-0.5, 0.5)$ for $t = 1$, and $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-0.7, 0.7, 0.2)$ for $t = 2, 3, 4$. By the inclusion of response indicator as a covariate, we allow the responses to have different intercepts. Whilst the intercept is $\beta_0 + \beta_2 = -1 + 0.2 = -0.8$ for Y_1, it is $\beta_0 = -1$ for Y_2. The relationships between X_1 and Y_1, and X_1 and Y_2 are assumed to be the same, i.e. interaction between X_1 and X_2 is not included. In terms association structures, by keeping the marginal mean parameter setting same, we consider the following four cases:

Case 1

$(\lambda_2^*, \lambda_2) = (0.9, 0.95), (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.2, 0.25, 0.3, 0.35), (\alpha_{21,1}, \alpha_{31,1}, \alpha_{41,1}) = (0.3, 0.4, 0.5)$

Case 2

$(\lambda_2^*, \lambda_2) = (1.1, 1.15), (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.2, 0.25, 0.3, 0.35), (\alpha_{21,1}, \alpha_{31,1}, \alpha_{41,1}) = (0.3, 0.4, 0.5)$

Case 3

$(\lambda_2^*, \lambda_2) = (0.9, 0.95), (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.5, 0.55, 0.6, 0.65), (\alpha_{21,1}, \alpha_{31,1}, \alpha_{41,1}) = (0.3, 0.4, 0.5)$

Case 4

$(\lambda_2^*, \lambda_2) = (0.9, 0.95), (\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.2, 0.25, 0.3, 0.35), (\alpha_{21,1}, \alpha_{31,1}, \alpha_{41,1}) = (0.6, 0.7, 0.8)$

with λ_1^* and λ_1 being 1. The relationships between the lag-1 and current responses are assumed to be same for Y_1 and Y_2, i.e. $Z_{itj,1} = [\ 1\ ]$.

Simulated data sets are analysed by the first-order model. The simulation procedure is replicated 500 times for each case. Analysis of a simulated data set (the last one) took 8.9 minutes on a PC with 4.00 GB RAM and 3.00 GHz processor. A simulated data set and the `R` script for data analysis are available in the user manual of the `pnmtrem` package.

Simulation results are displayed in Table 1. We report mean, percentage bias (Bias(%)), empirical standard deviations of the parameter estimates (SD), mean of the standard errors of the parameter estimates (meSE), and coverage probabilities of the corresponding 95% confidence intervals (CP). Parameters are approximately unbiased. Empirical standar deviations of the parameter estimates and the means of the standard error estimates are close to each other. Coverage probabilities are close to the nominal level of 0.95.

6 Example: Iowa Youth and Families Project data set

6.1 Data

We apply the first-order model to the data set from the Iowa Youth and Families Project (IFYP; Elder and Conger, 2000; Ilk, 2008). The project was conducted to investigate long-term effects of the farm crisis that began in 1980's in the U.S. 451 families from eight rural parts of the north central Iowa were selected. 7th graders with two alive and biological parents and a sibling within 4 years old were the target. The focus is on their well-being. The study was started in 1989, and conducted yearly until 1992. Then, it was conducted in 1994, 1995, 1997 and 1999. At each follow-up, both

Table 1: Simulation results based on 500 replications. For the details of the abbreviations, see the text.

		β_0^*	β_1^*	β_0	β_1	β_2
	True	-0.500	0.500	-0.700	0.700	0.200
Case 1	Mean	-0.504	0.513	-0.702	0.704	0.197
	Bias (%)	0.797	2.511	0.284	0.549	-1.557
	SE	0.114	0.199	0.081	0.122	0.068
	meSE	0.120	0.205	0.081	0.125	0.065
	CP	0.956	0.958	0.958	0.962	0.944
Case 2	Mean	-0.497	0.494	-0.700	0.697	0.250
	Bias (%)	-0.650	-1.252	0.022	-0.480	0.081
	SE	0.120	0.206	0.081	0.119	0.066
	meSE	0.119	0.204	0.082	0.125	0.065
	CP	0.950	0.936	0.952	0.956	0.956
Case 3	Mean	-0.497	0.499	-0.706	0.712	0.198
	Bias (%)	-0.546	-0.136	0.923	1.755	-0.924
	SE	0.125	0.220	0.077	0.120	0.063
	meSE	0.123	0.211	0.083	0.129	0.063
	CP	0.946	0.944	0.960	0.962	0.950
Case 4	Mean	-0.499	0.498	-0.703	0.701	0.201
	Bias (%)	-0.183	-0.338	0.358	0.176	0.456
	SE	0.116	0.195	0.082	0.125	0.063
	meSE	0.122	0.211	0.083	0.128	0.063
	CP	0.956	0.964	0.966	0.956	0.956

the parents and children were surveyed. At the beginning of the study, 48% of the 7th graders were male and their average age was 12.7 years.

Three main distress variables, anxiety, hostility and depression, were used to measure emotional statuses of the young people (Table 2). These variables were collected by a list of symptoms, e.g. including nervousness, shakiness, an urge to break things and feeling low in energy etc. The symptoms were then dichomotised (Ilk, 2008). The frequencies of the dichomotised distress variables are given in Table 3. The frequencies of depression were higher compared to those of anxiety and hostility, and the frequencies of the latter variables were close to each other. For instance, almost 93% of them reported at least one depression symptom at 1989, whilst the frequencies of anxiety and hostility were 83.2%. A set of explanatory variables, thought to be related with the distress variables, were also collected (Table 2). These variables include gender, degree of negative life event experiences of the young people, e.g. having a close friend moved away permanently, financial cutbacks, e.g. moving to a cheaper residence, and negative economical event experiences of their families, e.g. such as changing job for a worse one. Amongst the explanatory variables, whilst gender was time-invariant, the others were time-varying.

Transitional structure of our model requires equally-spaced data. Therefore, we analyse the first four follow-ups of the IYFP data set. Indicator variables for distress variables and time are considered as additional explanatory variables, and dummy variables are created for all the categorical covariates (Table 2). We coded the binary explanatory variables as 0 vs. 1 in our initial data analyses. However, the alternative coding of -1 vs. 1 is used due to convergence problems.

Table 2: Variable list of IYFP used in PNMTREM(1).

Variable	Explanation
Responses	
anxiety	whether the young person had symptoms: 0=absence, 1=presence
hostility	whether the young person had symptoms: 0=absence, 1=presence
depression	whether the young person had symptoms: 0=absence, 1=presence
Covariates	
gender	gender of the young person: -1=male, 1=female
NLE1	first indicator variable for negative life event experiences of young people: 1=some, -1=none or many
NLE2	second indicator variable for negative life event experiences of young people: 1=many, -1=none or some
NEE	whether the household had any negative economical event: -1=no, 1=yes
cut1	first indicator variable for financial cutback experiences of the household: 1=between 1 and 5, -1= none or more than 5
cut2	second indicator variable for financial cutback experiences of the household: 1=more than 5, -1= none or between 1 and 5
resp1	first response indicator variable: 1=hostility, -1=anxiety or depression
resp2	second response indicator variable: 1=depression, -1=hostility or anxiety
time1	first indicator variable for follow-up time: 1=1991, -1=1990 or 1992
time2	second indicator variable for follow-up time: 1=1992, -1=1990 or 1991

Table 3: Frequency table of the distress variables across years.

	1989	1990	1991	1992
Anxiety	375 (83.2%)	347 (76.9%)	342 (75.8%)	327 (72.5%)
Hostility	375 (83.2%)	350 (77.6%)	342 (75.8%)	328 (72.7%)
Depression	418 (92.7%)	385 (85.4%)	378 (83.8%)	386 (85.6%)

6.2 Results

We specifically build two models. Whilst the set of explanatory variables are same, the models differ in terms of separating the lag-1 associations amongst the distress variables. Whilst the first model (Model 1 in Table 5) assumes these associations are shared across the responses, i.e. $Z_{itj,1} = [\ 1\ ]$, the second model (Model 2 in Table 5) assumes that the associations are different for the distres variables, i.e. $\boldsymbol{Z}_{itj,1} = [\ 1\ \ resp1\ \ resp2\ ]$. Results for baseline models are presented in Table Table 4. Note that the baseline results of Model 1 and Model 2 are same, since the specifications of the baseline parameter sets are same. Results for $t \geq 2$ models are presented in Table 5.

We compare Model 1 and 2 by likelihood ratio test (LRT), since they are nested. Respective maximised log-likelihoods are -1236.78 ($= -210.78 - 1026$) and -1234.49 ($= -210.78 - 1023.71$). The LRT statistic is 4.58 ($= -2 * (-1026 - (-1023.71))$), with a p-value of 0.60. This indicates that there is not enough evidence to conclude that Model 2 is a better model to analyse the IYFP data set compared to Model 1. Therefore, throughout the paper we only discuss the results of Model 1.

We check existence of multicollinearity problem by variance inflation factor. The largest value is 1.17 (results not shown here). This indicates that multicollinearity is not a problem for the analysis of the IYFP data set. We rely on the findings of Ilk and Daniels (2007) regarding the exogeneity of the time-varying covariates in the IFYP data set.

Table 4: Results for $t = 1989$. $H_0 : \lambda^*_{hostility} = 1$ and $H_0 : \lambda^*_{depression} = 1$; other parameters are tested for 0.

	PNMTREM(1)			
Parameter	Est.	SE	Z	P
β^*_0	1.33	0.07	18.82	0.00
β^*_{gender}	-0.09	0.06	-1.41	0.16
β^*_{NLE1}	0.20	0.12	1.61	0.11
β^*_{NLE2}	0.41	0.12	3.27	0.00
β^*_{NEE}	0.03	0.05	0.72	0.47
β^*_{cut1}	0.08	0.07	1.15	0.25
β^*_{cut2}	-0.003	0.07	-0.04	0.97
β^*_{resp1}	-0.001	0.06	-0.02	0.99
β^*_{resp2}	0.29	0.07	4.17	0.00
$\beta^*_{gender*resp1}$	-0.04	0.06	-0.73	0.47
$\beta^*_{gender*resp2}$	-0.08	0.07	-1.29	0.20
$\lambda^*_{hostility}$	1.10	0.79	0.12	0.91
$\lambda^*_{depression}$	1.04	0.71	0.05	0.96
$\log(\sigma_1)$	-0.41	0.41		
Max. loglik.	-210.78			

6.3 Population-averaged results

At baseline (1989), only the intercept, one of the negative life event indicators (NLE2) and one of the response indicators (resp2) are significant. The estimate of intercept, $\hat{\beta}^*_0 = 1.33$, indicates that young people had high probability of distress at 1989. The estimate of the second response indicator variable, $\hat{\beta}^*_{resp2} = 0.29$, indicates that young people were more likely to report depression compared to anxiety and hostility. Insignificance of the first response indicator (p-value=0.99) indicates that reporting anxiety and hostility were equally likely. These findings are in agreement with the empirical frequencies (Table 3). Young people who had many negative life events were more likely to be distressed ($\hat{\beta}^*_{NLE2} = 0.41$). Pairwise correlations between anxiety, hostility and depression were not significantly different, p-values of $\lambda^*_{hostility}$ and $\lambda^*_{depression}$ were 0.91 and 0.96. The standard deviation estimate of the random effects distribution is 0.66 (= $\exp(-0.41)$), with a standard error of 0.27 (= $\sqrt{0.41^2 * \exp(-0.41 * 2)}$, by the delta method). The standard deviation is significantly different from 0, with a p-value of 0.007. Of note, we modified the p-value following Molenberghs and Verbeke (2007).

For 1990 – 1992, the intercept, gender, both negative life event indicators (NLE1, NLE2), negative economical events experience (NEE), one of the cutbacks indicators (cut1), one of the response indicators (resp2), one of the time indicators (time2) and the interaction between gender and second response indicator (gender * resp2) are significant. The estimate of the intercept ($\hat{\beta}_0 = 0.96$) indicates high probability of distress for 1990 – 1992, which tend to be higher compared to baseline, since $\hat{\beta}^*_0 > \hat{\beta}_0$. Females were more likely to report distress compared to males ($\hat{\beta}_{gender} = 0.18$). Furthermore, they were more likely to report depression ($\hat{\beta}_{gender*resp2} = 0.07$) compared to reporting anxiety or hostility. Note that gender is insignificant at 1989. This finding was also reported in Ge et al. (2001, cited in Ilk, 2008) and Ilk (2008). Experiencing many negative life events and any family-level negative economical events were associated with distress ($\hat{\beta}_{NLE1} = 0.14$, $\hat{\beta}_{NLE2} = 0.38$ and $\hat{\beta}_{NEE} = 0.08$). Reporting depression was more likely compared to reporting anxiety or hostility

Table 5: Results for $t \geq 1990$. $H_0 : \lambda_{hostility} = 1$ and $H_0 : \lambda_{depression} = 1$; other parameters are tested for 0.

	Model 1				Model 2			
Parameter	Est.	SE	Z	P	Est.	SE	Z	P
β_0	0.96	0.05	20.77	0.00	0.96	0.05	19.49	0.00
β_{gender}	0.18	0.03	5.87	0.00	0.18	0.03	5.87	0.00
β_{NLE1}	0.14	0.04	3.09	0.00	0.14	0.05	3.05	0.00
β_{NLE2}	0.38	0.05	7.95	0.00	0.38	0.05	7.90	0.00
β_{NEE}	0.08	0.03	3.08	0.00	0.08	0.03	3.03	0.00
β_{cut1}	0.06	0.03	2.10	0.04	0.07	0.03	2.20	0.03
β_{cut2}	0.02	0.03	0.72	0.47	0.02	0.03	0.73	0.47
β_{resp1}	0.01	0.04	0.28	0.78	0.01	0.04	0.13	0.90
β_{resp2}	0.22	0.04	5.21	0.00	0.22	0.05	4.66	0.00
β_{time1}	-0.07	0.04	-1.75	0.08	-0.08	0.05	-1.75	0.08
β_{time2}	-0.09	0.05	-1.96	0.05	-0.09	0.05	-1.88	0.06
$\beta_{gender*resp1}$	-0.01	0.03	-0.18	0.86	-0.01	0.03	-0.20	0.84
$\beta_{gender*resp2}$	0.07	0.04	2.08	0.04	0.07	0.04	2.07	0.04
$\beta_{resp1*time1}$	-0.002	0.03	-0.07	0.95	-0.02	0.04	-0.42	0.68
$\beta_{resp1*time2}$	0.004	0.04	0.10	0.92	0.003	0.04	0.07	0.94
$\beta_{resp2*time1}$	-0.01	0.04	-0.36	0.72	-0.01	0.04	-0.31	0.75
$\beta_{resp2*time2}$	0.05	0.04	1.15	0.25	0.05	0.04	1.03	0.30
$\alpha_{21,1}$	0.76	0.11	6.62	0.00	0.75	0.17	4.50	0.00
$\alpha_{22,1}$					0.06	0.13	0.43	0.67
$\alpha_{23,1}$					0.11	0.16	0.70	0.48
$\alpha_{31,1}$	0.87	0.10	9.11	0.00	0.86	0.13	6.58	0.00
$\alpha_{32,1}$					0.08	0.11	0.74	0.46
$\alpha_{33,1}$					0.07	0.14	0.48	0.63
$\alpha_{41,1}$	0.90	0.10	9.53	0.00	0.86	0.12	7.03	0.00
$\alpha_{42,1}$					-0.04	0.12	-0.34	0.74
$\alpha_{43,1}$					0.12	0.13	0.93	0.35
$\lambda_{hostility}$	1.03	0.37	0.60	0.94	0.99	0.36	-0.02	0.99
$\lambda_{depression}$	1.21	0.49	0.57	0.68	1.18	0.49	0.36	0.72
$\log(\sigma_2)$	-0.48	0.25			-0.47	0.26		
$\log(\sigma_3)$	-0.62	0.25			-0.59	0.26		
$\log(\sigma_4)$	-0.62	0.26			-0.59	0.26		
Max. loglik	-1026.00				-1023.71			

($\hat{\beta}_{resp2} = 0.22$). On the other hand, reporting anxiety or hostility were not significantly different (p-value of $\beta_{resp1} = 0.78$). Whilst the distress levels were lower at 1992 compared to 1990 and 1991 ($\hat{\beta}_{time2} = -0.09$), the distress levels at 1990 and 1991 were not significantly different from each other (p-value of $\beta_{resp1} = 0.08$). The decrease in the distress probabilities at 1992 was not different for depression, anxiety and hostility; respective p-values for $\beta_{resp1*time2}$ and $\beta_{resp2*time2}$ are 0.92 and 0.25.

The marginal mean parameter estimates based on *probit* link can be interpreted in terms of odds-ratios, using the JKB constant; for details see Introduction. For instance, young people who experienced many negative life events were approximately 2.26 ($= \exp(1.700437 * ((-1 * 0.14 + 1 * 0.38) - (1 * 0.14 - 1 * 0.38)))$) times more likely to be distressed compared to those with some negative life events, and individuals in the latter group were 1.60 ($= \exp(1.700437 * ((1 * 0.14 - 1 * 0.38) - (-1 * 0.14 - 1 * 0.38)))$) times more likely to be distressed compared to those with no negative life events.

The transition parameter estimates are positive and significant: $\hat{\alpha}_{21,1} = 0.76$, $\hat{\alpha}_{31,1} = 0.87$, $\hat{\alpha}_{41,1} = 0.90$ with p-values $< 1 \times 10^{-10}$. These indicate that that young people who were distressed at year before were more likely to be distressed at current year. As indicated by the baseline model, the pairwise correlations between anxiety, hostility and depression were not significantly different; corresponding p-values were 0.94 and 0.68 for hostility and depression, respectively. The standard deviation estimates of the random effects distributions were 0.62 (= $\exp(-0.48)$), 0.54 (= $\exp(-0.62)$) and 0.54 (= $\exp(-0.62)$) at 1990, 1991 and 1992, respectively. Respective standard errors were 0.16, 0.14 and 0.14, and all of these parameters were significant (p-values < 0.0001). These results indicate that the individual variations decreased through time (recall that $\hat{\sigma}_1$=0.66) and close to each other at 1991 and 1992.

6.4 Subject-specific results

We calculate probabilities of reporting anxiety, hostility and depression for each individual at each year. We also calculate the marginal probabilities for comparison. These probabilities are plotted in Figure 1; only the results for depression are shown here due to page limits, others can be found in the online supplementary material. We label observed values by 0 and 1 according to absence and presence depression, respectively. Marginal probabilities range in a narrower interval compared to conditional probabilities. For instance, whilst the range for the marginal probabilities of being depressed at the period of 1990 - 1992 was (0.576, 0.971), it was (0.118, 0.999) for the conditional probabilities. This indicates that marginal probabilities are high even for young people who did not report depression. On the other hand, conditional probabilities leads to correct decisions. For instance, in Figure 1, the 0's were associated with lower conditional probabilities. The associated box-plots reflect the location and scale of the marginal and conditional probabilities. Whereas the the conditional probabilities have a spread distribution with many outliers, the marginal probabilities have a stacked and narrow distribution.

Probabilities of a young person with ID=223 are presented in Table 6. This person was a female, with some negative life event experiences, no negative economical event experiences and cutbacks between 1 and 5, except in 1992 at which her family did not experience any cutbacks. She did not report distress at all. Predicted value of z_{223} is -2.45. This indicates that she was less likely to report distress compared to an average person, i.e. $z_i = 0$. For this person, the conditional probabilities lead to correct inferences compared to the marginal probabilities. For instance, at 1992, whereas the marginal probability of being anxious is 0.64, the conditional probability being anxious is 0.08. We also calculate conditional probabilities assuming that the person is an average person, i.e. setting $z_{223} = 0$. Related results are given under Conditional*. These probabilities are still subject, time and response specific, since Δ^*_{itj} holds subject, time and response specific information. For instance, at 1992, probability of being anxious based on this method is 0.46.

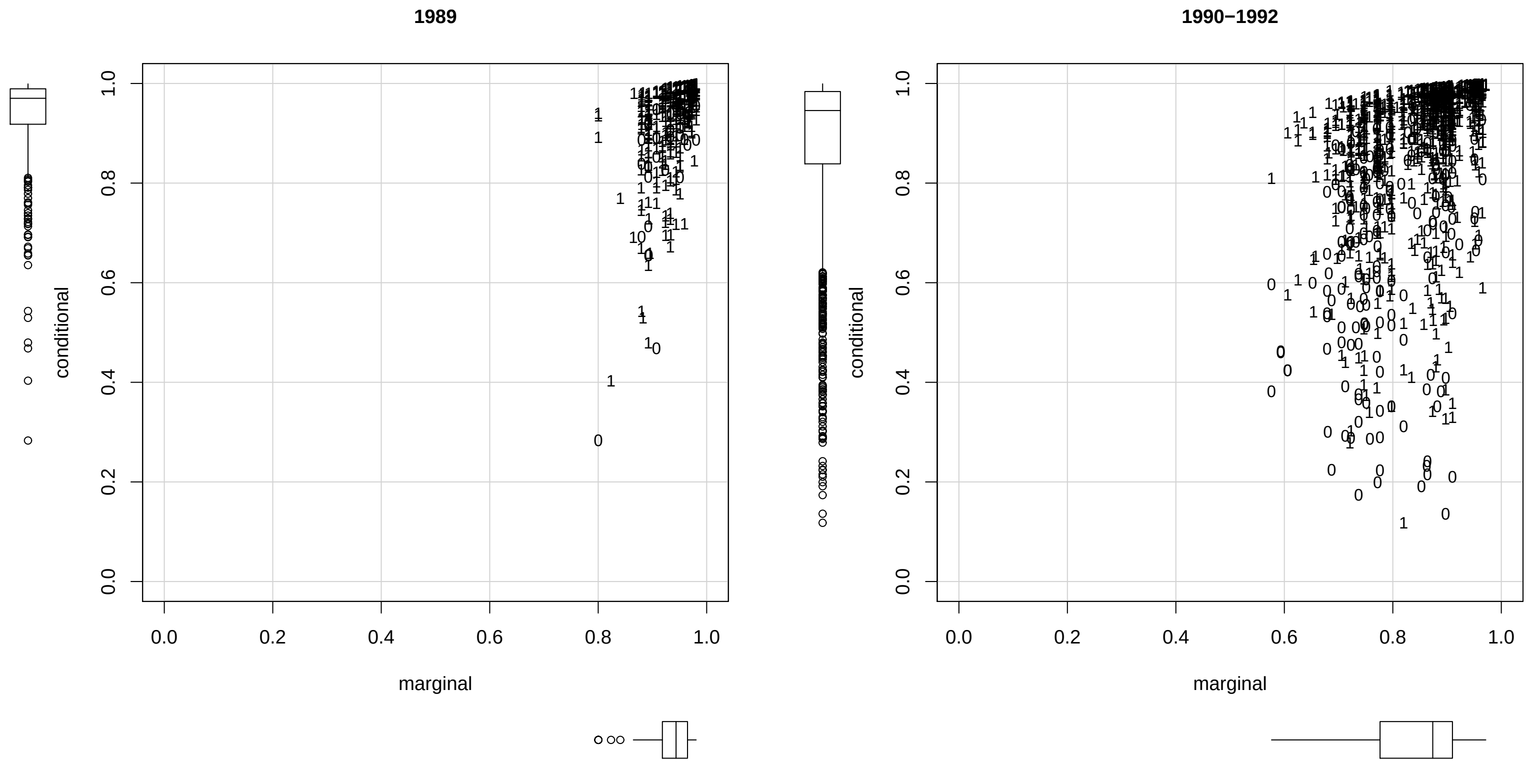

Figure 1: Scatter and box plots of marginal vs. conditional probabilities for depression at 1989 (left panel) and 1990-1992 (right panel).

Table 6: Marginal and conditional probabilities for the young person with ID = 223.

Time	Response	Gender	NLE	NEE	Cutbacks	Observed	Marginal	Conditional	Conditional*
	Anxiety	Female	Some	No	Betw. 1 & 5	Absence	0.82	0.30	0.87
1989	Hostility	Female	Some	No	Betw. 1 & 5	Absence	0.80	0.23	0.85
	Depression	Female	Some	No	Betw. 1 & 5	Absence	0.91	0.47	0.95
	Anxiety	Female	Some	No	Betw. 1 & 5	Absence	0.78	0.09	0.56
1990	Hostility	Female	Some	No	Betw. 1 & 5	Absence	0.78	0.09	0.58
	Depression	Female	Some	No	Betw. 1 & 5	Absence	0.90	0.14	0.77
	Anxiety	Female	Some	No	Betw. 1 & 5	Absence	0.74	0.14	0.59
1991	Hostility	Female	Some	No	Betw. 1 & 5	Absence	0.74	0.13	0.59
	Depression	Female	Some	No	Betw. 1 & 5	Absence	0.86	0.22	0.79
	Anxiety	Female	Some	No	None	Absence	0.64	0.08	0.46
1992	Hostility	Female	Some	No	None	Absence	0.65	0.08	0.47
	Depression	Female	Some	No	None	Absence	0.85	0.19	0.77

Table 7: Frequency table of the stayers. “All” stands for the subjects who reported the same answer for all the distress variables.

	Absence (0)	Presence (1)
Anxiety	15 (3.3%)	215 (47.7%)
Hostility	9 (2%)	221 (49%)
Depression	2 (0.4%)	288 (63.9%)
All	2 (0.4%)	134 (29.7%)

6.5 Diagnostics

Longitudinal binary data sets almost surely include stayers, i.e. subjects who constantly report absence (0) or presence (1) of a binary variable at all time points, for which the subject with ID=223 is an example. The counts and percentages of the stayers in the IYFP data set are given in Table 7. For instance, 29.7% of the subjects reported 1 for all the three distress variables at all the time points. Marginal and conditional anxiety probabilities of the stayers in terms all the distress variables are summarised in Figure 2. Other results can be found the online supplementary material. Whilst the gray lines represent the subjects who always reported 1, the black lines represent the ones who always reported 0. Conditional probabilities are successful at correctly assigning the success probabilities for these subjects; higher probabilities for subjects reporting 1 and lower probabilities for those who reported 0. On the other hand, marginal probabilities are not able to distinguish these subjects.

We also calculate accuracy measures to summarise the predicted probabilities. We specifically use expected proportion of correct prediction (Herron, 1999) and area under the receiver operating characteristics curve (AUROC). Results (not shown here) show that conditional probabilities outperform the marginal probabilities. This difference is apparent especially in terms of AUROC. For instance, while the AUROC value for depression at 1990-1992 is 0.684 for marginal probabilities, it is 0.864 for the conditional probabilities.

7 Discussion and conclusion

In this paper, we have proposed a marginalised model for analysis of multivariate longitudinal binary data. It is an extension of the model proposed by Ilk and Daniels (2007). These authors use *logit* link, and estimate the parameters using BM, specifically Markov Chain Monte Carlo methods. Un-

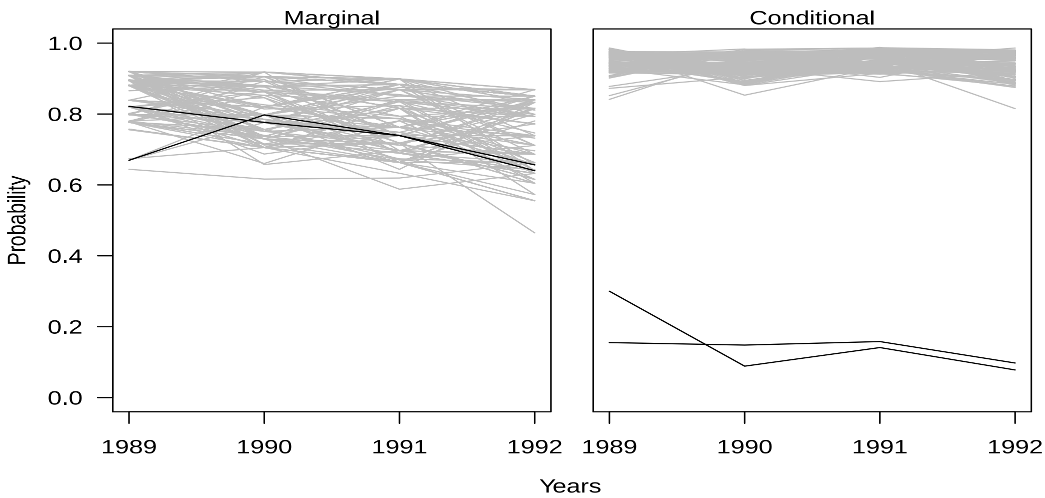

Figure 2: Spagetti plots of marginal (left panel) and conditional (right panel) anxiety probabilities for stayers in terms of all the distress variables. While gray lines represent subjects who reported 1, the black lines represent subjects who reported 0.

automated Fortran codes are available from the personal website of Dr. Ilk. However, their procedure is computationally cumbersome and requires expertise in BM and Fortran. These aspects prohibit the routine use of the model. In this study, we replace *logit* link by *probit*, and use ML for parameter estimation. *probit* link enables us explicitly linking the second and third levels of the model, which is not possible with the *logit* link. On the other hand, parameter estimation with ML takes less time compared to BM. We propose the use of implicit function theorem to solve the marginal constraint equations directly. To the best of our knowledge, this application is proposed for the first time here for marginalised models. We have prepared the publicly available R package `pnmtrem` to fit the proposed model. Currently, the package provides a function for fitting the first-order model. The function considers both parameter estimation and random effects prediction. It has been tested under different conditions. For the details and usage, we refer the readers to the package manual.

We have conducted a simulation study to investigate the properties of the estimator under different scenarios. Results are satisfactory in terms of unbiasedness, efficiency and coverage. We have illustrated the first-order model with an application to the IYFP data set. Both population-averaged and subject-specific inferences have been illustrated. Our findings on the IYFP data analysis coincide with the findings of Ilk (2008). As a separate note, the IYFP data set is available upon request from the authors.

A natural extension of our work here would be fitting higher-order models. The variances of random effects could be modified by a subset of covariates, i.e. $\log(\sigma_t) = \boldsymbol{M}_{itj}\ \boldsymbol{\omega}_t$ where $\boldsymbol{M}_{itj}$ is a set covariates and $\boldsymbol{\omega}_t$ are the associated parameters. Also, the random effects coefficients might be assumed to have a multivariate normal distribution, i.e. $b_{it} \sim N(0, \boldsymbol{D})$ where $\boldsymbol{D}$ is a $T \times T$ matrix. However, all of these extensions require intensive new derivations and implementations. Therefore, we leave them as future work.

Appendices

A. Linking second and third levels of the $t \geq 2$ model

Whilst linking second and third levels of the $t \geq 2$ model, we claim the following

$$\int \Phi(\Delta^*_{itj} + \lambda_j b_{it}) f(b_{it}) db_{it} = \Phi\left(\frac{\Delta^*_{itj}}{\sqrt{1+\lambda_j^2\sigma_t^2}}\right)$$

where $b_{it} \sim N(0, \sigma_t^2)$ and $b_{it} = z_i\sigma_t$, $z_i \sim N(0, 1)$. The related proof, which is modified from Griswold (2005), is given below.

Let $W_i \perp z_i$, where $W_i \sim N(0, 1)$, then,

$$\begin{gathered} W_i/(\lambda_j\sigma_t) \sim N(0, (\lambda_j\sigma_t)^{-2}) \\ W_i/(\lambda_j\sigma_t) - z_i \sim N(0, 1 + (\lambda_j\sigma_t)^{-2}) \\ \frac{W_i/(\lambda_j\sigma_t) - z_i}{\sqrt{1+(\lambda_j\sigma_t)^{-2}}} \sim N(0, 1) \end{gathered}$$

and

$$\begin{aligned} \int \Phi(\Delta^*_{itj} + \lambda_j b_{it}) f(b_{it}) db_{it} &= \int_{-\infty}^{+\infty} \Phi(\Delta^*_{itj} + \lambda_j z_i \sigma_t)\phi(z_i) dz_i \\ &= \int_{-\infty}^{+\infty} P(W_i \leq \Delta^*_{itj} + \lambda_j z_i \sigma_t)\phi(z_i) dz_i \\ &= \int_{-\infty}^{+\infty} P\left(\frac{W_i/(\lambda_j\sigma_t) - z_i}{\sqrt{1 + (\lambda_j\sigma_t)^{-2}}} \leq \frac{\Delta^*_{itj}/(\lambda_j\sigma_t)}{\sqrt{1 + (\lambda_j\sigma_t)^{-2}}}\right)\phi(z_i) dz_i \\ &= P\left(\frac{W_i/(\lambda_j\sigma_t) - z_i}{\sqrt{1 + (\lambda_j\sigma_t)^{-2}}} \leq \frac{\Delta^*_{itj}/(\lambda_j\sigma_t)}{\sqrt{1 + (\lambda_j\sigma_t)^{-2}}}\right) = \Phi\left(\frac{\Delta^*_{itj}}{\sqrt{1 + (\lambda_j\sigma_t)^2}}\right) \end{aligned}$$

B.1 ML estimation of $\boldsymbol{\theta}_1$

Maximizing the log-likelihood function of the baseline model, $L_1(\boldsymbol{\theta}_1|\boldsymbol{y}_1)$, with respect to $\boldsymbol{\theta}_1$ yields

$$\frac{\partial \log\left(L_1(\boldsymbol{\theta}_1|\boldsymbol{y}_1)\right)}{\partial \boldsymbol{\theta}_1} \approx \sum_{i=1}^{N} \frac{1}{h(Y_{i1}|\boldsymbol{\theta}_1)} \frac{\partial h(Y_{i1}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1}, \tag{27}$$

where

$$h(Y_{i1}|\boldsymbol{\theta}_1) \approx \sum_{q=1}^{20} w_q \underbrace{exp\left[\sum_{j=1}^{k}\left(Y_{i1j}\log\left(\Phi(d_{i1jq})\right) + (1 - Y_{i1j})\log\left(1 - \Phi(d_{i1jq})\right)\right)\right]}_{\ell(Y_{i1}|\boldsymbol{\theta}_1)}, \tag{28}$$

$$\frac{\partial h(Y_{i1}|\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \approx \sum_{q=1}^{20} w_q \left\{\ell(Y_{i1}|\boldsymbol{\theta}_1)\left\{\sum_{j=1}^{k}\left[\frac{\partial d_{i1jq}}{\partial \boldsymbol{\theta}_1}\phi(d_{i1jq})\left(\frac{Y_{i1j} - \Phi(d_{i1jq})}{\left(\Phi(d_{i1jq})\right)\left(1 - \Phi(d_{i1jq})\right)}\right)\right]\right\}\right\}, \tag{29}$$

$$d_{i1jq} = \sqrt{1 + \lambda_j^{*2} e^{2c_1}}\ (\boldsymbol{X}_{i1j}\boldsymbol{\beta}^*) + \lambda_j^* e^{c_1} \sqrt{2}\ z_q. \tag{30}$$

Here, $log(\sigma_1)$ is equated to c_1 for simplicity of notation and (z_q, w_q) for $q = 1, \ldots, 20$ are Gauss-Hermite quadrature points and weights, respectively which are available in Abramowitz and Stegun (1972). The derivatives of d_{i1jq} with respect to $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^*, \boldsymbol{\lambda}^*, c_1)$ with $\boldsymbol{\lambda}^* = (\lambda_2^*, \ldots, \lambda_k^*)$ are given below.

$$\begin{aligned}
\frac{\partial d_{i1jq}}{\partial \boldsymbol{\beta}^*} &= \sqrt{1+\lambda_j^{*2} e^{2c_1}}(\boldsymbol{X}_{i1j}) \\
\frac{\partial d_{i1jq}}{\partial \lambda_j^*} &= (1+\lambda_j^{*2} e^{2c_1})^{-1/2} \lambda_j^* e^{2c_1}(\boldsymbol{X}_{i1j}\boldsymbol{\beta}^*) + e^{c_1}\sqrt{2} z_q \\
\frac{\partial d_{i1jq}}{\partial c_1} &= (1+\lambda_j^{*2} e^{2c_1})^{-1/2} \lambda_j^{*2} e^{2c_1}(\boldsymbol{X}_{i1j}\boldsymbol{\beta}^*) + \lambda_j^* e^{c_1}\sqrt{2} z_q
\end{aligned}$$

B.2 ML estimation of $\boldsymbol{\theta}_2$

Similar to the baseline model, maximizing the log-likelihood function of the $t \geq 2$ model with respect to $\boldsymbol{\theta}_2$ yields

$$\frac{\partial log\,(L_2(\boldsymbol{\theta}_2|\boldsymbol{y}_2))}{\partial \boldsymbol{\theta}_2} \approx \sum_{i=1}^{N}\sum_{t=2}^{T} \frac{1}{h(Y_{it}|\boldsymbol{\theta}_2)} \frac{\partial h(Y_{it}|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2}, \tag{31}$$

where

$$h(Y_{it}|\boldsymbol{\theta}_2) \approx \sum_{q=1}^{20} w_q \underbrace{exp\left[\sum_{j=1}^{k}\left(Y_{itj} log\left(\Phi(d_{itjq})\right) + (1-Y_{itj}) log\left(1-\Phi(d_{itjq})\right)\right)\right]}_{\ell(Y_{it}|\boldsymbol{\theta}_2)}, \tag{32}$$

$$\frac{\partial h(Y_{it}|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} \approx \sum_{q=1}^{20} w_q \left\{\ell(Y_{it}|\boldsymbol{\theta}_2)\left\{\sum_{j=1}^{k}\left[\frac{\partial d_{itjq}}{\partial \boldsymbol{\theta}_2}\phi(d_{itjq})\left(\frac{Y_{itj}-\Phi(d_{itjq})}{\left(\Phi(d_{itjq})\right)\left(1-\Phi(d_{itjq})\right)}\right)\right]\right\}\right\}, \tag{33}$$

$$d_{itjq} = \sqrt{1+\lambda_j^2 e^{2c_t}}\left(\Delta_{itj} + \boldsymbol{\alpha}_{t,1}\boldsymbol{Z}_{itj} y_{it-1j}\right) + \lambda_j e^{c_t}\sqrt{2} z_q. \tag{34}$$

Here, $c_t = log(\sigma_t)$ for $t \geq 2$, and (z_q, w_q) for $q = 1, \ldots, 20$ are Gauss-Hermite quadrature points and weights. Also note that explicit solution of Δ_{itj} is given in (13). The derivatives of d_{itjq} with respect to $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}, \boldsymbol{\alpha}_{t,1}, \boldsymbol{\lambda}, \boldsymbol{c})$ with $\boldsymbol{\lambda} = (\lambda_2, \ldots, \lambda_k)$ and $\boldsymbol{c} = (c_2, \ldots, c_T)$ are given below.

$$\begin{aligned}
\frac{\partial d_{itjq}}{\partial \boldsymbol{\beta}} &= \sqrt{1+\lambda_j{}^2 e^{2c_t}}(\boldsymbol{A}_{itj}) \\
\frac{\partial d_{itjq}}{\partial \boldsymbol{\alpha}_{t,1}} &= \sqrt{1+\lambda_j{}^2 e^{2c_t}}(\boldsymbol{B}_{itj} + \boldsymbol{Z}_{itj,1} y_{it-1j}) \\
\frac{\partial d_{itjq}}{\partial \lambda_j} &= (1+\lambda_j{}^2 e^{2c_t})^{-1/2} \lambda_j e^{2c_t}\left(-(\boldsymbol{A}_{itj}\boldsymbol{\beta}_0 + \boldsymbol{B}_{itj}\boldsymbol{\alpha}_{t,10}) + \boldsymbol{A}_{itj}\boldsymbol{\beta} + \boldsymbol{\alpha}_{t,1}(\boldsymbol{B}_{itj} + \boldsymbol{Z}_{itj,1} y_{it-1j})\right) + e^{c_t}\sqrt{2} z_q \\
\frac{\partial d_{itjq}}{\partial c_t} &= (1+\lambda_j{}^2 e^{2c_t})^{-1/2} \lambda_j{}^2 e^{2c_t}\left(-(\boldsymbol{A}_{itj}\boldsymbol{\beta}_0 + \boldsymbol{B}_{itj}\boldsymbol{\alpha}_{t,10}) + \boldsymbol{A}_{itj}\boldsymbol{\beta} + \boldsymbol{\alpha}_{t,1}(\boldsymbol{B}_{itj} + \boldsymbol{Z}_{itj,1} y_{it-1j})\right) + \lambda_j e^{c_t}\sqrt{2} z_q
\end{aligned}$$

where

$$\boldsymbol{A}_{itj} = -\frac{\left.\frac{\partial F}{\partial \boldsymbol{\beta}}\right|_{(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{t,10}, \Delta_{itj0})}}{\left.\frac{\partial F}{\partial \Delta_{itj}}\right|_{(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{t,10}, \Delta_{itj0})}}, \quad \boldsymbol{B}_{itj} = -\frac{\left.\frac{\partial F}{\partial \boldsymbol{\alpha}_{t,1}}\right|_{(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{t,10}, \Delta_{itj0})}}{\left.\frac{\partial F}{\partial \Delta_{itj}}\right|_{(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_{t,10}, \Delta_{itj0})}}$$

References

Abramowitz, K. M. and Stegun, I. A. (1972) *Handbook of mathematical functions*. New York: Dover Publications.

Agresti, A. (2002) *Categorical data analysis*, 2nd edition. New Jersey: John Wiley & Sons.

Asar, Ö. and Ilk, Ö. (2013) mmm: an R package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine*, **112**, 649–654.

Asar, Ö. and Ilk, Ö. (2014) Flexible multivariate marginal models for analyzing multivariate longitudinal data, with applications in R. *Computer Methods and Programs in Biomedicine*, **115**, 135–146.

Ashford, J. and Sowden, R. (1970) Multivariate probit analysis. *Biometrics*, **26(3)**, 535–546.

Caffo, B. and Griswold, M. (2006) User-friendly introduction to link-probit-normal models. *American Statistician*, **60 (2)**, 139–145.

Diggle, P. J., Heagerty, P., Liang, K. -Y. and Zeger, S. L. (2002) *Analysis of longitudinal data*, 2nd edition. Oxford: Oxford University Press.

Doksum, K. A. and Gasko, M. (1990) On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review*, **58**, 243–252.

Efron, B. (1986) Why isn't everyone a Bayesian? *The American Statistician*, **40(1)**, 1–5.

Elder, G. H. and Conger, R. (2000) *Children of the land*. The University of Chicago Press.

Ge, X., Conger, R. D. and Elder, G. H. (2001) Pubertal transition, stressful life events, and the emergence of gender differences in adolescent depressive symptoms. *Developmental Psychology*, **37(3)**, 404–417.

Griswold, M. (2005) *Complex distributions, hmmmm... hierarchical mixtures of marginalized multilevel models*. Ph.D. Thesis, Johns Hopkins University.

Griswold, M., Swihart, B. J., Caffo B. S. and Zeger S. L. (2013) Practical marginalized multilevel models. *Stat*, **2**, 129–142.

Heagerty, J. P. (1999) Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **45(3)**, 688–698.

Heagerty, P. and Zeger, S. L. (2000) Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). Statistical Science, **15(1)**, 1–26.

Heagerty, J. P. and Kurland, B. F. (2001) Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, **88(4)**, 973–985.

Heagerty, J. P. (2002) Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, **58**, 342–351.

Hedeker, D. and Gibbons, R. D. (2006) *Longitudinal data analysis*. New Jersey: John Wiley & Sons.

Herron, M. (1999) Postestimation uncertainty in limited dependent variable models. *Political Analysis*, **8**, 83–98.

Iddi, S. and Molenberghs, G. (2012) A combined overdispersed and marginalized multilevel model. *Computational Statistics and Data Analysis*, **56**, 1944–151.

Ilk, Ö. and Daniels, M. J. (2007) Marginalized transition random effects models for multivariate longitudinal binary data. *The Canadian Journal of Statistics*, **35**, 105–123.

Ilk, Ö. (2008) *Multivariate longitudinal data analysis: models for binary response and exploratory tools for binary and continuous response*. Saarbrücken: Verlag Dr. Müller (VDM).

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous univariate distributions, volume 2*, 2nd edition. New York: John Wiley & Sons.

Krantz, S. G. and Parks, H. R. (2003) *The implicit function theorem: history, theory and applications*. Boston: Birkhäuser.

Lee, K. and Daniels, M. J. (2007) Marginalized models for longitudinal ordinal data with application to quality of life studies. *Statistics in Medicine*, **27** 4359–4380.

Lee, K., Joo, Y., Yoo, J. K. and Lee, J. (2009) Marginalized random effects models for multivariate longitudinal binary data. *Statistics in Medicine*, **28**, 1284–1300.

Lee K. and Mercante D. (2010) Longitudinal nominal data analysis using marginalized models. *Computational Statistics and Data Analysis*, **54**, 208–218.

Lee, K., Joo, Y., Song, J. J. and Harper, D. W. (2011) Analysis of zero-inflated clustered count data: a marginalized model approach. *Computational Statistics and Data Analysis*, **55**, 824–837.

Lee, K., Daniels, M. J. and Joo, Y. (2013) Flexible marginalized models for bivariate longitudinal ordinal data. *Biostatistics*, **14(3)**, 462–476.

Lesaffre, E. and Spiessens, B. (2001) On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, **50(3)**, 325–335.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73(1)**, 13–22.

Liu, L. C. and Hedeker D (2006) A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, **62**, 261–268.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008) *Generalized, linear, and mixed models*, 2nd edition. New Jersey: John Wiley & Sons.

Molenberghs, G. and Verbeke, G. (2007) Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician*, **61(1)**, 22–27.

R Core Development Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schildcrout, J. S. and Heagerty, P. (2007) Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics*, **63**, 322–331.

Shelton, B. J., Gilbert, G. H., Liu, B. and Fisher, M. (2004) A SAS macro for the analysis of multivariate longitudinal binary outcomes. *Computer Methods and Programs in Biomedicine*, **76**, 163–175.

Varin, C. and Czado, C. (2010) A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics*, **11(1)**, 127–138.