Statistica Sinica **14**(2004), 989-1020

MEANINGFUL STATISTICAL MODEL FORMULATIONS FOR REPEATED MEASURES

Geert Molenberghs and Geert Verbeke

Limburgs Universitair Centrum and Catholic University of Leuven

Abstract: When choosing a parametric statistical model two important considerations are mathematical soundness and substantive relevance. In this paper, we illustrate and exemplify that a number of issues arise from these considerations, even in relatively simple settings, such as ordinal regression, linear mixed models, models for cross-classified data and generalized linear mixed models. Many of our points are illustrated with data.

Key words and phrases: Binary data, conditional model, generalized linear mixed model, likelihood ratio test, linear mixed model, logistic regression, marginal model, ordinal data, random effects, score test, variance components.

1. Introduction

Choosing a parametric statistical model is a common task in statistical practice. When choosing a model, it is important to reflect on whether the model is sound from a theoretical point of view and whether it is adequate in terms of the scientific research question of interest. While some authors have approached aspects of this problem from a fundamental, theoretical perspective (McCullagh (2002) and references therein) it is fair to say that the problem receives less attention in everyday practice than it should. We consider a number of simple but key settings in order to make a number of general and specific points about this topic. Many of these points are illustrated using a few simple settings (Section 2).

First, we consider the linear mixed-effects model, that has become a standard tool for analyzing repeated continuous, normally distributed outcomes. While it looks like a relatively straightforward extension of linear regression it is surrounded with a number of problems, some of them arising due to the fact that one can adopt either a hierarchical or a marginal point of view which, while having connections, are different. The implications of this fact for variance component testing ought not to be overlooked (Section 4). Second, when switching from normally distributed to discrete repeated measures (Section 5), one should realize that, even though there are links to the simpler linear mixed model, the situation is dramatically more complicated. In particular, one can choose between a number of relevant but non-equivalent modelling families. Within each

family (marginal, conditional, mixed-effects models), specific issues have to be addressed. Quite a bit of confusion stems from real or apparent connections between the families.

2. Examples

In this section, two examples, used to illustrate various points, are introduced.

2.1. Toenail data

The data come from a randomized, double-blind, parallel group, multicenter study for the comparison of two oral treatments for toenail dermatophyte onychomycosis. Patients with a clinical diagnosis of toe onychomycosis, confirmed by a positive direct microscopy and a positive culture for dermatophytes at a central laboratory, were randomly assigned to treatment A or treatment B. After a twelve week treatment period, there was a follow-up period of 36 weeks. Patients returned to the hospital at months 0 (baseline), 1, 2, 3, 6, 9 and 12. More details can be found in De Backer, De Vroey, Lesaffre, Scheys and De Keyser (1998). One of the outcomes measured at each occasion was the severity of the infection, coded as 0 (not severe) or 1 (severe). The question of interest was whether the rate of severe infections decreased over time, and whether that evolution was different for the two treatment groups. Although 189 patients were initially included in each group, only 118 patients from group A and 108 patients from group B completed the study. However, we ignore this dropout problem for now, and refer to Verbeke and Molenberghs (2000) for an extensive discussion on dropout, and on missing data in general.

2.2. Fluvoxamine study

These data come from a multicenter study involving 315 patients that were treated by fluvoxamine for psychiatric symptoms described as possibly resulting from a dysregulation of serotonine in the brain. Patients with one or more of the following diagnoses were included: depression, obsessive, compulsive disorder and panic disorder. Several covariates were recorded, such as sex and initial severity. After recruitment of the patient in the study, he or she was investigated at three visits. On the basis of about twenty psychiatric symptoms, the therapeutic effect and the side-effects were scored at each visit in an ordinal manner. Side effect is coded as (1) = no; (2) = not interfering with functionality of patient; (3) = interfering significantly with functionality of patient; (4) = the side-effect surpasses the therapeutic effect. Similarly, the effect of therapy is recorded on a four point ordinal scale: (1) no improvement over baseline or worsening; (2)

minimal improvement (not changing functionality); (3) moderate improvement (partial disappearance of symptoms) and (4) important improvement (almost disappearance of symptoms). Thus a side effect results if new symptoms occur, while there is therapeutic effect if old symptoms disappear. These data were used, among others, by Molenberghs and Lesaffre (1994) and Lapp, Molenberghs and Lesaffre (1998).

3. Hierarchical and Marginal Views on the Linear Mixed Models

The linear mixed-effects model (Laird and Ware (1982) and Verbeke and Molenberghs (2000)) is a commonly used tool for, among others, variance component models and for longitudinal data. The model and some of its implications for interpretational meaningfulness will be discussed in this and the next section.

Let $\boldsymbol{Y_i}$ denote the n_i-dimensional vector of measurements available for subject $i = 1, \ldots, N$. A general linear mixed model then assumes that $\boldsymbol{Y_i}$ satisfies

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \tag{1}$$

in which $\boldsymbol{\beta}$ is a vector of population-average regression coefficients called fixed effects, and where $\boldsymbol{b_i}$ is a vector of subject-specific regression coefficients. The $\boldsymbol{b_i}$ describe how the evolution of the ith subject deviates from the average evolution in the population. The matrices X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. The random effects $\boldsymbol{b_i}$ and residual components $\boldsymbol{\varepsilon_i}$ are assumed to be independent with distributions $N(\mathbf{0}, D)$, and $N(\mathbf{0}, \Sigma_i)$, respectively. Inference for linear mixed models is usually based on maximum likelihood or restricted maximum likelihood estimation under the marginal model for $\boldsymbol{Y_i}$, i.e., the multivariate normal model with mean $X_i\boldsymbol{\beta}$, and covariance $V_i = Z_iDZ_i' + \Sigma_i$ (Laird and Ware (1982), Verbeke and Molenberghs (2000)). Thus, we can adopt two *different* views on the linear mixed model. The fully *hierarchical* model is specified by

$$\boldsymbol{Y_i}|\boldsymbol{b}_i \sim N_{n_i}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i), \qquad \boldsymbol{b}_i \sim N(0, D), \tag{2}$$

while the marginal model is given by

$$\boldsymbol{Y_i} \sim N_{n_i}(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i). \tag{3}$$

In practice, one can be interested in the fully hierarchical model (e.g., to use the random effects for individual-level predictions) or in the marginal model only. In the latter case, a hierarchical model formulation might be used as a convenient tool to derive a (parsimonious) covariance structure. However, even though they are often treated as equivalent, there are important differences between the hierarchical and marginal views of the model. Obviously, (2) requires the covariance

matrices Σ_i and D to be positive definite, while in (3) it is sufficient for the resulting matrix V_i to be positive definite.

Different hierarchical models can produce the same marginal model. To see this, consider the case where every subject is measured twice ($n_i = 2$). First, assume that the random-effects structure is confined to a random intercept ($\boldsymbol{b_i}$ is scalar) and residual error structure $\Sigma_i = \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ (Model I):

$$V = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (d) \begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} d+\sigma_1^2 & d \\ d & d+\sigma_2^2 \end{pmatrix}. \tag{4}$$

Second, consider the random effects to consist of a random intercept and a random slope, $\boldsymbol{b_i} = (b_{0i}, b_{1i})'$, mutually uncorrelated, with residual error structure $\Sigma_i = \Sigma = \sigma^2 I_2$ (Model II):

$$V = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} = \begin{pmatrix} d_1+\sigma^2 & d_1 \\ d_1 & d_1+d_2+\sigma^2 \end{pmatrix}. \tag{5}$$

Obviously, (4) and (5) are equivalent: $d_1 = d$, $d_2 = \sigma_2^2 - \sigma_1^2$ and $\sigma^2 = \sigma_1^2$. Thus different hierarchical models can produce the same marginal model, illustrating that a good fit of the marginal model cannot be seen as equally strong evidence for any hierarchical model. Arguably, a satisfactory treatment of the hierarchical model is only possible within a Bayesian context. There is another subtle difference between Models I and II. Model I is a proper hierarchical model if d, σ_1^2 and σ_2^2 are nonnegative. However, this is not sufficient for Model II, since, if $\sigma_2^2 < \sigma_1^2$, d_2 is negative and hence it is then impossible for the resulting D matrix to be positive definite. In the reverse case, both Models I and II are consistent with the marginal model.

Further, there exist marginal models that are not implied by a hierarchical model. The simplest example is found by restricting the random effects in (1) to a random intercept and choosing $\Sigma_i = \sigma^2 I_{n_i}$. The resulting marginal model is given by

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, \tau^2 J_{n_i} + \sigma^2 I_{n_i}), \tag{6}$$

where J_{n_i} equals the $n_i \times n_i$ matrix containing only ones. Regarding the variance component τ^2 in the above model, one can take two views. In the first view, where the focus is entirely on the resulting marginal model (6), negative values for τ^2 are perfectly acceptable (Nelder (1954), Verbeke and Molenberghs (2000, Section 5.6.2)), since this merely corresponds to the occurrence of negative within-cluster correlation $\rho = \tau^2/(\tau^2 + \sigma^2)$. This might occur, for example, in a context of competition such as when littermates compete for the same food resources. In the second view, when the link between the marginal model (6) and its generating hierarchical model (2) is preserved, thereby including the concept of random

effects b_i and perhaps even requiring inference for them, it is imperative to restrict τ^2 to nonnegative values. In other words, a model with negative τ^2 cannot be derived from a hierarchical model.

4. Nonstandard Testing Problems in a Hierarchical View

While both the marginal and hierarchical views are possible, there are important differences regarding statistical inference for variance components. The first situation, which we term the *unconstrained case*, is standard regarding inference for the variance component τ^2. Under the unconstrained parameterization, i.e., the model under which negative values for τ^2 are allowed, classical inferential tools are available for testing the general two-sided hypothesis $H_0 : \tau^2 = 0$ *versus* $H_{A2} : \tau^2 \neq 0$. Wald, likelihood ratio and score tests are then asymptotically equivalent, and the asymptotic null distribution is well known to be χ^2_1 (Cox and Hinkley (1990)). Under the constrained model, i.e., the model where τ^2 is restricted to the non-negative real numbers, the one-sided hypothesis (7) is the only meaningful one.

In the second situation (the *constrained case*), however, one typically needs one-sided tests of the null-hypothesis

$$H_0 : \tau^2 = 0 \qquad \text{versus} \qquad H_{A1} : \tau^2 > 0. \tag{7}$$

As the null-hypothesis is now on the boundary of the parameter space, classical inference no longer holds, and appropriately tailored test statistics need to be developed along with their corresponding (asymptotic) null distributions.

Suppressing dependence on the other parameters, let $\ell(\tau^2)$ denote the log-likelihood as a function of the random-intercepts variance τ^2. Further, let $\widehat{\tau}^2$ denote the maximum likelihood estimate of τ^2 under the unconstrained parameterization. We first consider the likelihood ratio test statistic:

$$T_{LR} = 2 \ln \left[\frac{\max_{H_{1A}} \ell(\tau^2)}{\max_{H_0} \ell(\tau^2)} \right].$$

Two cases, graphically represented in Figure 1, can now be distinguished. Under Case A, $\widehat{\tau}^2$ is positive, and the likelihood ratio test statistic is identical to the one that would be obtained under the unconstrained parameter space for τ^2. Hence conditionally on $\widehat{\tau}^2 \geq 0$, T_{LR} has asymptotic null distribution equal to the classical χ^2_1. Under Case B however, we have that, under H_{1A} as well as under H_0, $\ell(\tau^2)$ is maximized at $\tau^2 = 0$ yielding $T_{LR} = 0$. Under H_0, both cases occur with 50% probability. Hence the asymptotic null distribution of T_{LR} is easily seen to follow a $0.5P(\chi^2_1 > c) + 0.5P(\chi^2_0 > c)$ null distribution, where χ^2_0 denotes the distribution with all probability mass at 0. Hence the asymptotic

null distribution of the one-sided likelihood ratio test statistic is a mixture of two chi-squared distributions, with degrees of freedom 0 and 1, and with equal mixing proportions 1/2. This was one of Stram and Lee's (1994, 1995) special cases. Note that, whenever $\widehat{\tau}^2 \geq 0$, the observed likelihood ratio test statistic is equal to the one under the unconstrained model, but the p-value is half the size of the one obtained from the classical χ^2_1 approximation to the null distribution.

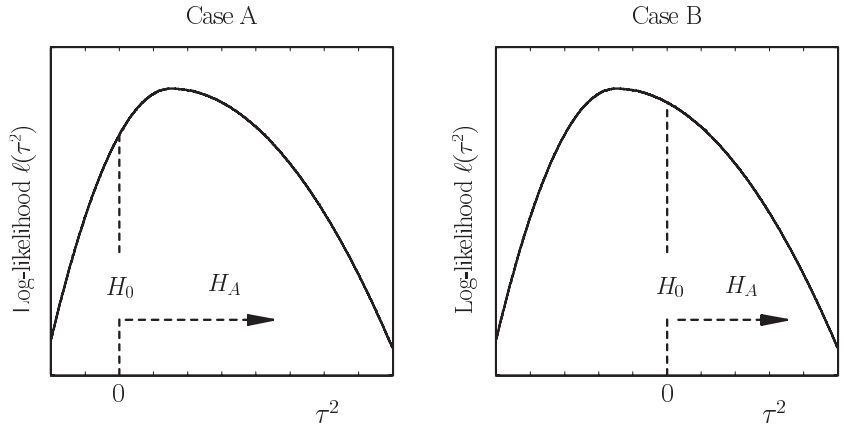

Figure 1. Graphical representation of two different situations, when developing one-sided tests for the variance τ^2 of the random intercepts b_i in model.

Similar to the random-intercepts model, the general marginal model does not require D to be positive definite, while a hierarchical interpretation of the model does. As before, inference under the unconstrained model for the variance components in D can be based on the classical chi-squared approximation to the null distribution for the likelihood ratio test statistic. Under the constrained model, Stram and Lee (1994, 1995) have shown that the asymptotic null distribution for the likelihood ratio test statistic for testing a null hypothesis which allows for k correlated random effects versus an alternative of $k+1$ correlated random effects (with positive semi-definite covariance matrix D_{k+1}), is a mixture of a χ^2_k and a χ^2_{k+1}, with equal probability 1/2. For more general settings, e.g., comparing models with k and $k+k'$ ($k' > 1$) random effects, the null distribution is a mixture of χ^2 random variables (Shapiro (1988) and Raubertas, Lee and Nordheim (1986)), the weights of which can only be calculated analytically in special cases. Shapiro's (1988) results provide a few important special cases not studied by Stram and Lee (1994). For example, if the null hypothesis allows for k *un*correlated random effects (with a diagonal covariance matrix D_k) versus the alternative of $k+k'$ uncorrelated random effects (with diagonal covariance

matrix $D_{k+k'}$), the null distribution is a mixture of the form

$$\sum_{m=0}^{k'} 2^{-k'} \binom{k'}{m} \chi^2_m.$$

Verbeke and Molenberghs (2003), using results by Silvapulle and Silvapulle (1995), have shown that similar results are obtained when a score test is used instead of a likelihood ratio test. To provide insight, we again consider the random-intercepts model and then sketch the general result. The usual form of the test statistic in the scalar case is

$$T_S = \Big[\frac{\partial \ell(\tau^2)}{\partial \tau^2}\Big|_{\tau^2=0}\Big]^2 \Big[-\frac{\partial^2 \ell(\tau^2)}{\partial \tau^2 \partial \tau^2}\Big|_{\tau^2=0}\Big]^{-1}. \tag{8}$$

Nuisance parameters are suppressed and replaced by their MLE's. The classical score test (8) implicitly assumes a two-sided alternative. Hence, the test statistic itself needs to be redefined appropriately in order to be able to discriminate between positive and negative alternative values for τ^2. The same two cases as for the likelihood ratio test can be considered (see Figure 1). Under Case A, $\widehat{\tau}^2$ is positive, and the positive score $\partial \ell(\tau^2)/\partial \tau^2$ at zero is evidence against H_0 in favor of the one-sided alternative H_{A1}. Hence (8) can be used as test statistic, provided that $\widehat{\tau}^2 \geq 0$. This implies that, conditionally on $\widehat{\tau}^2 \geq 0$ and under H_0, our test statistic asymptotically follows the classical χ^2_1 distribution. Under Case B, however, the score at $\tau^2 = 0$ is negative, and therefore cannot be used as evidence against H_0 in favor of H_{A1}. Hence, whenever $\widehat{\tau}^2$ is negative, (8) is no longer meaningful as test statistic. Considering that a negative score at zero supports the null hypothesis, a meaningful modified test statistic is obtained from restricting (8) to the case where $\widehat{\tau}^2 \geq 0$ and setting it to zero in case $\tau^2 < 0$. It is easily seen that the asymptotic null distribution is, again, $0.5P(\chi^2_1 > c) + 0.5P(\chi^2_0 > c)$. This heuristic but insightful argument can be formalized and generalized to vector valued settings. The above heuristic arguments have suggested that employment of score tests for testing variance components under the constrained parameterization requires replacing the classical score test statistic by an appropriate one-sided version. This is where the general theory of Silvapulle and Silvapulle (1995) on one-sided score tests proves very useful. They consider models parameterized through a vector $\boldsymbol{\theta} = (\boldsymbol{\lambda}', \boldsymbol{\psi}')'$, where testing a general hypothesis of the form $H_0 : \boldsymbol{\psi} = \mathbf{0}$ *versus* $H_A : \boldsymbol{\psi} \in \mathcal{C}$ is of interest. In our context, the alternative parameter space $\mathcal{C}$ equals the nonnegative real numbers (e.g., when testing (7), or the set of positive semi-definite covariance matrices D). In general, Silvapulle and Silvapulle (1995) allow $\mathcal{C}$ to be a closed and convex cone in Euclidean space, with vertex at the origin. The advantage

of such a general definition is that one-sided, two-sided, and combinations of one-sided and two-sided hypotheses are included.

Adopt the following notation. Let $\boldsymbol{S}_N(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ be the score vector and Hessian matrix of the log-likelihood function. Further, decompose $\boldsymbol{S}_N$ as $\boldsymbol{S}_N = (\boldsymbol{S}'_{N\lambda}, \boldsymbol{S}'_{N\psi})'$, let $H_{\lambda\lambda}(\boldsymbol{\theta})$, $H_{\lambda\psi}(\boldsymbol{\theta})$ and $H_{\psi\psi}(\boldsymbol{\theta})$ be the corresponding blocks in $H(\boldsymbol{\theta})$, and define $\boldsymbol{\theta}_H = (\boldsymbol{\lambda}', \mathbf{0}')'$. $\boldsymbol{\theta}_H$ can be estimated by $\widehat{\boldsymbol{\theta}}_H = (\widehat{\boldsymbol{\lambda}}', \mathbf{0}')'$, in which $\widehat{\boldsymbol{\lambda}}$ is the maximum likelihood estimate of $\boldsymbol{\lambda}$ under H_0. Finally, let $\boldsymbol{Z}_N$ be equal to $\boldsymbol{Z}_N = N^{-1/2}\boldsymbol{S}_{N\psi}(\widehat{\boldsymbol{\theta}}_H)$. A one-sided modified score statistic can now be defined as

$$T_S := \boldsymbol{Z}'_N H^{-1}_{\psi\psi}(\widehat{\boldsymbol{\theta}}_H)\boldsymbol{Z}_N \;-\; \inf\left\{(\boldsymbol{Z}_N - \boldsymbol{b})' H^{-1}_{\psi\psi}(\widehat{\boldsymbol{\theta}}_H)(\boldsymbol{Z}_N - \boldsymbol{b}) | \boldsymbol{b} \in \mathcal{C}\right\}. \quad (9)$$

Note that the modified score statistic, heuristically defined in the case of the random-intercepts model, is a special case of (9). Indeed when $\widehat{\tau}^2$ is positive the score at zero is positive, and therefore in $\mathcal{C}$, such that the infimum in (9) becomes zero. For $\widehat{\tau}^2$ negative, the score at zero is negative as well and the infimum in (9) is attained for $\boldsymbol{b} = \mathbf{0}$, resulting in $T_S = 0$.

It follows from Silvapulle and Silvapulle (1995) that, under suitable regularity conditions, as $N \to \infty$, the modified likelihood ratio and score test statistics satisfy $T_{LR} = T_S + o_p(1)$. This indicates that the equivalence of the score and likelihood ratio tests not only holds in the classical two-sided but also in the modified one-sided cases. Moreover, what is known about the null distribution in the case of the likelihood ratio test immediately carries over to the score test case. This result corrects the common belief that, even when variance components are on the boundary of the parameter space, the score test deserved no special treatment. Verbeke and Molenberghs (2003) provide an empirical illustration. In practice, calculation of (9) requires some extra programming work and, even though it is not insurmountable, one may therefore be inclined to resort to likelihood ratio testing.

5. Discrete Repeated Measures

In the previous section we discussed a number of issues arising from the use of the linear mixed effects model. In particular, we focused on complexities stemming from the difference between a marginal and a hierarchical (random-effects) interpretation of such a model. Marginal and random-effects models are two important sub-families of models for repeated measures. Several authors, such as Diggle, Heagerty, Liang and Zeger (2002) and Aerts, Geys, Molenberghs and Ryan (2002) distinguish between three such families. Still focusing on continuous outcomes, a marginal model is characterized by the specification of a marginal mean function

$$E(Y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}, \quad (10)$$

whereas a random-effects model focuses on the expectation, conditional upon the random-effects vector

$$E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i. \tag{11}$$

Finally, a third family of models conditions a particular outcome on the other responses or a subset thereof. In particular, a simple first-order stationary transition model focuses on expectations of the form

$$E(Y_{ij}|Y_{i,j-1}, \ldots, Y_{i1}, \boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \alpha Y_{i,j-1}. \tag{12}$$

The fixed-effects (random-effects) covariate vector $\boldsymbol{x}_{ij}$ ($\boldsymbol{z}_{ij}$) groups all covariates that are used in the model for the measure at occasion j. In line with the linear mixed model sections, we often group the outcomes Y_{ij} into a vector $\boldsymbol{Y}_i$. In such cases the covariates, when explicitly usesd, are grouped into matrices X_i and Z_i.

As seen before, random-effects models imply a simple marginal model in the linear mixed model case. This is due to the elegant properties of the multivariate normal distribution. In particular, the expectation (10) follows from (11) by either (a) marginalizing over the random effects or (b) by conditioning upon the random-effects vector $\boldsymbol{b_i} = \boldsymbol{0}$. Hence, the fixed-effects parameters $\boldsymbol{\beta}$ have both a marginal as well as a hierarchical model interpretation. Finally, when a conditional model is expressed in terms of residuals rather than outcomes directly, it also leads to particular forms of the general linear mixed effects model.

Such a close connection between the model families does not exist when outcomes are non-Gaussian. We consider each of the model families in turn, then point to some particular issues arising within them or when comparisons are made between them. We first review some general concepts from univariate generalized linear models, with emphasis on logistic regression.

5.1. Generalized linear models, exponential family, and logistic regression

For the analysis of binary response variables, one of the most commonly used tools is logistic regression (Agresti (1990)). There are at least three obvious reasons for this. First, it is considered an extension of linear regression. Second, it fits within the theory of generalized linear models. Third, especially in a biometrical context, the interpretation of its parameters in terms of odds ratios is considered convenient. When the latter is less of a concern, such as in econometric applications, one frequently encounters probit regression.

Consider a response variable Y_i, measured on subjects $i = 1, \ldots, N$, together with covariates $\boldsymbol{x}_i$. A generalized linear model minimally specifies the mean $E(Y_i) = \mu_i$ and links it to a linear predictor in the covariates $\eta(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$,

where $\eta(.)$ is the so-called link function. Further, the variance of Y_i is then linked to the mean model by the mean-variance link $\text{Var}(Y_i) = \phi v(\mu_i)$, where $v(.)$ is a known variance function and ϕ is a scale or overdispersion parameter. Such a specification is sufficient to implement moment-based estimation methods, e.g., iteratively reweighted least squares or quasi likelihood (McCullagh and Nelder (1989)). In case full likelihood is envisaged, the above framework can be seen to be derived from the general exponential family definition

$$f(y|\theta_i, \phi) = \exp\left\{\phi^{-1}[y\theta_i - \psi(\theta_i)] + c(y, \phi)\right\} \tag{13}$$

with θ_i the natural parameter and $\psi(.)$ a function satisfying $\mu_i = \psi'(\theta_i)$ and $v(\mu_i) = \psi''(\theta_i)$.

In the case of a binary outcome Y_i, the model can be written as

$$f(y_i|\theta_i, \phi) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left\{y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i)\right\},$$

and hence the Bernoulli model and, by extension logistic regression, fits within this framework. In particular,

$$\theta_i = \text{logit}(\mu_i) = \mu_i/(1-\mu_i) = \text{logit}[P(Y_i = 1|\boldsymbol{x}_i)], \tag{14}$$

$\mu = e^\theta/(1+e^\theta)$ and $v(\mu) = \mu(1-\mu)$. In case one opts for a probit link, the logit in (14) is replaced by the inverse of the standard normal distribution Φ^{-1}, i.e., the probit function. This model cannot be put within the exponential family context. Hence, the choice for logistic regression is often based on the mathematical convenience entailed by the exponential family framework. Now, it has been repeatedly shown (Agresti (1990)) that the logit and probit link functions behave very similarly, in the sense that for probabilities other than extreme ones (say, outside the interval $[0.2; 0.8]$) both forms of binary regression provide approximately the same parameter estimates, up to a scaling factor equal to $\pi/\sqrt{3}$, the ratio of the standard deviations of a logistic and a standard normal variable.

The beauty and elegance of the exponential family framework should not disguise that there are fundamental differences with linear regression. First, the normal densities, explicitly or implicitly underlying linear regression, exhibit a separation between mean and variance; this is radically different in most commonly used generalized linear models. Second, the link function introduces a form of non-linearity that is absent in linear regression, complicating, for example, model selection.

In spite of these remarks, logistic regression has found its way into everyday statistical practice. Perhaps due to this familiarity, the model has been extended to a number of different settings, including longitudinal data. We show such extensions are less straightforward in the non-Gaussian case than when the linear mixed model is used.

5.2. Marginal models

In marginal models, the parameters characterize the marginal probabilities of a subset of the outcomes, without conditioning on the others. Advantages and disadvantages of conditional and marginal modeling have been discussed in Diggle, Heagerty, Liang and Zeger (2002), and Fahrmeir and Tutz (2001). The specific context of clustered binary data has received treatment in Aerts, Geys, Molenberghs and Ryan (2002). Apart from full likelihood approaches, non-likelihood methods, such as generalized estimating equations (Liang and Zeger (1986)) or pseudo-likelihood (le Cessie and van Houwelingen (1994) and Geys, Molenberghs and Lipsitz (1998)) have been considered.

Bahadur (1961) proposed a marginal model, accounting for the association via marginal correlations. Ekholm (1991) proposed a so-called success probabilities approach. George and Bowman (1995) proposed a model for the particular case of exchangeable binary data. Ashford and Sowden (1970) considered the multivariate probit model, for repeated ordinal data, thereby extending univariate probit regression. Molenberghs and Lesaffre (1994) and Lang and Agresti (1994) have proposed models which parameterize the association in terms of marginal odds ratios. Dale (1986) defined the bivariate global odds ratio model, based on a bivariate Plackett distribution (Plackett (1965)). Molenberghs and Lesaffre (1994, 1999) extended this model to multivariate ordinal outcomes. Their 1994 method involves solving polynomials of high degree, while in 1999 generalized linear models theory is exploited, together with an adaptation of the iterative proportional fitting algorithm. Lang and Agresti (1994) exploit the equivalence between direct modeling and imposing restrictions on the multinomial probabilities, using undetermined Lagrange multipliers. Alternatively, the cell probabilities can be fitted using a Newton iteration scheme, as suggested by Glonek and McCullagh (1995). We consider some of these models in turn.

5.3. Some marginal models for repeated binary data

Let the binary response Y_{ij} indicate outcome j for individual i. Let

$$\varepsilon_{ij} = \frac{Y_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1-\mu_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1-\mu_{ij})}},$$

where y_{ij} is an actual value of the binary response variable Y_{ij}. Further, let $\rho_{ijk} = E(\varepsilon_{ij}\varepsilon_{ik})$, $\rho_{ijk\ell} = E(\varepsilon_{ij}\varepsilon_{ik}\varepsilon_{i\ell})$, $\ldots$, $\rho_{i12\ldots n_i} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in_i})$. The parameters ρ_{ijk} are classical Pearson type correlation coefficients. The general Bahadur model can be represented by the expression $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, where

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}},$$

$$c(\boldsymbol{y}_i) = 1 + \sum_{j<k} \rho_{ijk} e_{ij} e_{ik} + \sum_{j<k<\ell} \rho_{ijk\ell} e_{ij} e_{ik} e_{i\ell} + \ldots + \rho_{i12\ldots n_i} e_{i1} e_{i2} \ldots e_{in_i}.$$

Thus, the probability mass function is the product of the independence model $f_1(\boldsymbol{y}_i)$ (combining n_i logistic regressions) and the correction factor $c(\boldsymbol{y}_i)$. The factor $c(\boldsymbol{y}_i)$ can be viewed as a model for overdispersion. The Bahadur model has a very tractable form, a clear advantage over other, more implicitly defined, models. However, a practical drawback is the fact that the correlation between two responses is highly constrained when the higher order correlations are removed. Such a decision is often made to keep the computations, as well as the modeling exercise, within reasonable limits. Even when higher order parameters are included, the parameter space of marginal parameters and correlations is known to be of a very peculiar shape. For detailed studies, see Kupper and Haseman (1978), Prentice (1988) and Declerck, Aerts and Molenberghs (1998). In conclusion, this model is very appealing at first sight, since it combines logistic regression for the univariate marginal distributions with seemingly very interpretable correlation coefficients. However our intuition about correlation coefficients largely comes from the normal distribution, where there is a total separation between the mean parameters and the dependence parameters. In the present case, they are heavily constrained, not only by themselves but also by the marginal parameters. This makes computations and interpretation difficult. Thus, while the correlation is undoubtedly a meaningful parameter in the case of normally distributed outcomes, it can be highly questionable in the context of binary outcomes.

Ekholm (1991), using $\mu_{ijk} = P[Y_{ij} = 1, Y_{ik} = 1|\boldsymbol{x}_i]$, considered $\eta_{ijk} = \text{logit}(\mu_{ijk}) = \ln(\mu_{ijk}) - \ln(1 - \mu_{ijk})$, with similar definition for higher orders. While such an approach seems symmetric and therefore appealing, this is only seemingly so. For example, while both $P(Y_{ij} = 1|\boldsymbol{x}_i)$ and $P(Y_{ij} = 0|\boldsymbol{x}_i)$ are linear in the covariates on the logit scale, this is not true for $P[Y_{ij} = 0, Y_{ik} = 1|\boldsymbol{x}_i]$, $P[Y_{ij} = 1, Y_{ik} = 0|\boldsymbol{x}_i]$, or $P[Y_{ij} = 0, Y_{ik} = 0|\boldsymbol{x}_i]$, in spite of $P[Y_{ij} = 1, Y_{ik} = 1|\boldsymbol{x}_i]$ being modeled linearly on the logit scale. Moreover, the range of μ_{ijk} is restricted by the values for the univariate probabilities, the so-called Fréchet bounds: $\max(0, \mu_{ij} + \mu_{ik} - 1) \leq \mu_{ijk} \leq \min(\mu_{ij}, \mu_{ik})$, implying complicated restrictions on the parameters. Even when valid probabilities are obtained, interpretation of such coefficients is problematic. This is another instance of a model that seems appealing but engenders a lot of practical and interpretational difficulties.

5.4. Some marginal models for repeated ordinal data

While we already encountered problems with marginal models for repeated binary data, issues are magnified with ordinal outcomes. We introduce a modeling formalism in this section, then introduce conditional models in the next,

after which we will be in a position to discuss the meaningfulness of one relative to the other.

The outcome for cluster i is a series of measurements Y_{ij} $(j = 1, \ldots, n_i)$. Assume that Y_{ij} can take on c_j distinct ordered values $k_j = 1, \ldots, c_j$. It is convenient to define so-called cumulative multi-indicator functions: $z_i(\boldsymbol{k}) = z_i(k_1, \ldots, k_{n_i}) = I(\boldsymbol{y}_i \leq \boldsymbol{k})$, the multi-indicator being one if every component of the vector-valued inequality is satisfied and zero otherwise. The corresponding probability is denoted by $\mu_i(\boldsymbol{k})$. The choice to use cumulative indicators is in agreement with the ordinal nature of the outcomes. Setting one or more of the indices k_j equal to their maximal value c_j has the effect of marginalizing over the corresponding outcome. Doing this for all but one index results in the univariate indicators $z_{ijk} = I(y_{ij} \leq k)$ and their corresponding marginal probability μ_{ijk}. The ordering needed to stack the multi-indexed counts and probabilities into a vector will be done by dimensionality.

We can now complete the model by choosing appropriate link functions. For the vector of links $\boldsymbol{\eta}_i$ we consider a function mapping the C_i-vector $\boldsymbol{\mu}_i$ $(C_i = c_1 \cdot c_2 \cdot \ldots \cdot c_{T_i})$ to

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_i(\boldsymbol{\mu}_i), \tag{15}$$

a C_i'-vector. Often, $C_i = C_i'$, and $\boldsymbol{\eta}_i$ and $\boldsymbol{\mu}_i$ have the same ordering. A counterexample is provided by the probit model, where the number of link functions is smaller than the number of mean components, as soon as $n_i > 2$.

We consider particular choices of link functions. The univariate logit link becomes $\eta_{ijk} = \ln(\mu_{ijk}) - \ln(1 - \mu_{ijk}) = \text{logit}(\mu_{ijk})$. The probit link is $\eta_{ijk} = \Phi_1^{-1}(\mu_{ijk})$, with Φ_1 the univariate standard normal distribution. Next, full specification of the association requires addressing the form of pairwise and higher-order probabilities. First we consider pairwise associations. Let us denote the bivariate probabilities pertaining to the j_1th and j_2th outcomes by $\mu_{i,j_1j_2,k_1k_2} = \mu_i(c_1, \ldots, c_{j_1-1}, k_1, c_{j_1+1}, \ldots, c_{j_2-1}, k_2, c_{j_2+1}, \ldots, c_{n_i})$. The Dale model is based on the marginal global odds ratio defined by

$$\psi_{i,j_1j_2,k_1k_2} = \frac{(\mu_{i,j_1j_2,k_1k_2})(1 - \mu_{ij_1k_1} - \mu_{ij_2k_2} + \mu_{i,j_1j_2,k_1k_2})}{(\mu_{ij_2k_2} - \mu_{i,j_1j_2,k_1k_2})(\mu_{ij_1k_1} - \mu_{i,j_1j_2,k_1k_2})}, \tag{16}$$

and is usefully modeled on the log scale. Higher order global odds ratios are defined similarly.

The multivariate probit model also fits within the class defined by (15). For three categorical outcome variables, the inverse link is specified by $\mu_{ijk} = \Phi_1(\eta_{ijk})$, $\mu_{i,j_1j_2,k_1k_2} = \Phi_2(\eta_{ij_1k_1}, \eta_{ij_1k_2}, \eta_{i,j_1j_2,k_1k_2})$ and $\mu_{i,123,k_1k_2k_3} = \Phi_3(\eta_{i1k_1}, \eta_{i2k_3}, \eta_{i3k_3}, \eta_{i,12,k_1k_2}, \eta_{i,13,k_1k_3}, \eta_{i,23,k_2k_3})$. The association links $\eta_{i,ts,k\ell}$ represent any transform (e.g., Fisher's z-transform) of the polychoric correlation coefficient. It is common practice to keep each correlation constant throughout a table, rather

than having it depend on the categories: $\eta_{i,j_1j_2,k_1k_2} \equiv \eta_{i,j_1j_2}$. Relaxing this requirement may still give a valid set of probabilities, but the correspondence between the categorical variables and a latent multivariate normal variable is lost. A choice may be driven by whether one wants a good description of the cell probabilities or inference in terms of the underlying latent variables (e.g., to quantify association structures). Finally, observe that univariate links and bivariate links (representing correlations) fully determine the joint distribution.

We return to these models in Section 5.7, after introducing conditional models. However, we would like to assert that both the global odds ratio model, building upon univariate logits, and the multivariate probit model often provide meaningful models where a choice of one versus the other, even though there are differences, is less pronounced. This will be rather different in the context of random-effects models (Section 5.9).

5.5. Conditional models

In a conditional model the parameters describe a feature (probability, odds, logit,...) of (a set of) outcomes, given values for the other outcomes (Cox (1972)). The best known example is undoubtedly the log-linear model. Rosner (1984) described a conditional logistic model. Due to the popularity of marginal (especially generalized estimating equations) and random-effects models for correlated binary data, conditional models have received relatively little attention. Diggle, Heagerty, Liang and Zeger (2002, pp.142-143) criticized the conditional approach because the interpretation of the covariate effect on the probability of one outcome is conditional on the responses of other outcomes for the same individual, outcomes of other individuals and the cluster size.

We consider the model proposed by Cox (1972). The probability mass function is given by

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i;\boldsymbol{\Theta}_i) = \exp\Big\{\sum_{j=1}^{n_i}\theta_{ij}y_{ij} + \sum_{j<j'}\omega_{ijj'}y_{ij}y_{ij'} + \cdots + \omega_{i1\cdots n_i}y_{i1}\cdots y_{in_i} - A(\boldsymbol{\Theta}_i)\Big\}. \tag{17}$$

The θ parameters can be thought of as "main effects", whereas the ω parameters are association parameters. Models that do not include all interactions are derived by replacing the vector of ω parameters by one of its subvectors. A useful special case is found by setting all three and higher order parameters equal to zero. This is a member of the quadratic exponential family discussed by Zhao and Prentice (1990). Thélot (1985) studied the case where $n_i = n = 2$. If $n_i = n = 1$, the model reduces to ordinary logistic regression. The parameters $\omega_{ijj'}$ can be interpreted as conditional odds ratios, i.e., the odds ratio between outcomes at

occasions j and j', conditional upon all other outcomes being zero. Given the exponential family nature of the model, parameter estimation is particularly easy.

Model (17) is usually not meaningful when the cluster sizes n_i are unequal. Indeed, when $n_i = 1$ then $\theta_{i1} = \text{logit}[P(Y_{ij} = 1]$ while, when $n_i = 2$, $\theta_{i1} = \text{logit}[P(Y_{ij} = 1|Y_{ij} = 0]$. Thus, the same parameter would change its interpretation depending on the cluster size. When $n_i = n$ for all i, and the design is balanced (i.e., measurement occasions are common to all clusters), then the model is mathematically principled. The question then is whether the investigator is interested in a response to a conditional question rather than to, for example, a marginal one. A marginal question might be whether the probability of side effects in the fluvoxamine study increases or decreases with time; a conditional question might consider the probability of side effects at the second occasion, given there were none at the first occasion.

5.6. Generalized estimating equations

The main issue with full likelihood approaches is the computational complexity they entail. When we are mainly interested in first-order marginal mean parameters and pairwise interactions, a full likelihood procedure can be replaced by quasi-likelihood methods (McCullagh and Nelder (1989)), solely expressing the mean response as a function of covariates and writing the variance as a function of the mean, up to possibly unknown scale parameters. Wedderburn (1974) first noted that likelihood and quasi-likelihood theories coincide for exponential families and that the quasi-likelihood "estimating equations" provide consistent estimates of the regression parameters $\boldsymbol{\beta}$ in any generalized linear model, even for choices of link and variance functions that do not correspond to exponential families.

For correlated data, Liang and Zeger (1986) proposed *generalized estimating equations* (GEE or GEE1), requiring only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the association structure. The method combines estimating equations for the regression parameters $\boldsymbol{\beta}$ with moment-based estimation of the correlation parameters entering the working assumptions. Prentice (1988) extended their results to allow joint estimation of probabilities and pairwise correlations. Lipsitz, Laird and Harrington (1991) modified the latter estimating equations, replacing correlations by odds ratios. When adopting GEE1 one does not use information of the association structure to estimate the main effect parameters. As a result, it can be shown that GEE1 yields consistent main effect estimators, even when the association structure is misspecified. However, severe misspecification may seriously affect the efficiency of the GEE1 estimators. In addition, GEE1 should be avoided when some scientific interest is placed on the association parameters.

Liang, Zeger and Qaqish (1992) proposed a second-order extension (GEE2), fully specifying the association model.

5.7. Marginal versus conditional models and global odds ratios

Having introduced marginal and conditional models, we are now in a position to discuss points of meaningfulness of one relative to the other. It will be clear from the briefest comparison, that fitting a marginal model is typically more involved than fitting the conditional model of the previous section. Most marginal models have constrained parameter spaces. This is often cited as an interpretational disadvantage. However, the same is true for the multivariate normal model since the covariance matrix has to be positive definite. Exactly the same constraint applies to the multivariate probit model and similar but less tractable constraints apply to the Dale model. In contrast, the parameters of (17) can take on any value in the Euclidean space whilst still producing valid probabilities. Also, marginal models differ one from the other in terms of the severity of the restrictions. While in the Bahadur model the association parameter is restricted, even when $n_i = n = 2$, this is not the case in the Dale model where the odds ratio can range over the entire parameter space $[0, +\infty]$. Restrictions in the higher dimensional case exist but are rather weak.

One of the main interpretational advantages of marginal models is their upward compatibility or reproducibility (Liang, Zeger and Qaqish (1992)). This means that when a marginal model (e.g., the Dale, probit, or Bahadur model) is used to model a response vector, the appropriate sub-model applies to any subvector of the response vector. Such a sub-vector still follows a model of the same structure, with as parameter vector the corresponding sub-vector. In particular, the univariate margins of the marginal models discussed above are typically of the logistic type, the probit model the obvious exception.

Marginal models should be chosen whenever there are marginal research questions, e.g., pertaining to one or a few occasions, or the evolution between them (e.g., the time evolution of the response in the toenail data). They are also useful when not only the strength of association between occasions, but also a quantification of this association is of interest. Of course, when the number of measurement occasions within a subject grows, such models become intractable from a likelihood perspective. One can then resort to alternative approaches, such as generalized estimating equations or pseudo-likelihood.

We view the odds ratio as a meaningful measure of association between repeated categorical outcomes. It can be defined in several ways. Model (17) is based on conditional odds ratios, whereas the multivariate Dale model is based on marginal odds ratios. Apart from a marginal-conditional dimension, there is also a local-global dimension to the discussion. The odds ratios are local in the

log-linear model and global in the Dale model. Lapp, Molenberghs and Lesaffre (1998) provide some support for the use of global odds ratios rather than local ones for cross-classified ordinal data. They do so based on a comparison of the Dale model with Goodman's (1981) association models. Consider a single $J \times K$ contingency table (no covariates). Log local cross-ratios are given by

$$\ln \theta^*_{jk} = \ln \left(\frac{\text{pr}(Y_1 = j, Y_2 = k)\text{pr}(Y_1 = j+1, Y_2 = k+1)}{\text{pr}(Y_1 = j, Y_2 = k+1)\text{pr}(Y_1 = j+1, Y_2 = k)} \right) = \ln \frac{\mu^*_{jk}\mu^*_{j+1,k+1}}{\mu^*_{j,k+1}\mu^*_{j+1,k}},$$

with $j = 1, \ldots, J-1$ and $k = 1, \ldots, K-1$. Then for

$$\mu^*_{jk} = \alpha_j \beta_k e^{\phi \lambda_j \nu_k}, \tag{18}$$

$j = 1, \ldots, J$; $k = 1, \ldots, K$, α_j and β_k are main effect parameters while λ_j, ν_k and ϕ describe the association structure. Indeed, the local cross-ratios are $\ln \theta^*_{jk} = \phi(\lambda_j - \lambda_{j+1})(\nu_k - \nu_{k+1})$. Identifiability constraints have to be imposed on the parameters in (18). This model is also called the row-column model (RC model). The predictor function can be represented as $\boldsymbol{\eta} = \ln \boldsymbol{\mu}^* = g(\boldsymbol{\xi})$, with $\boldsymbol{g}(\boldsymbol{\xi})$ defined by

$$g_{ij}(\boldsymbol{\xi}) = \ln \alpha_j + \ln \beta_k + \phi \lambda_j \nu_k. \tag{19}$$

Predictor function (19) is non-linear and a combination of main effects and association parameters. An alternative association parameterization is additive in the log cross-ratios: $\ln \theta^*_{jk} = \delta_{1j} + \delta_{2k}$ and induced by

$$\mu^*_{jk} = \alpha_j \beta_k \gamma^k_{1j} \gamma^j_{2k}. \tag{20}$$

For this parameterization, (19) changes to $g_{jk}(\boldsymbol{\xi}) = \ln \alpha_j + \ln \beta_k + k \ln \gamma_{1j} + j \ln \gamma_{2k}$. Note that this predictor is linear in the parameters.

Goodman (1981) generalizes (18) to

$$\mu^*_{jk} = \alpha_j \beta_k \exp \left(\sum_{\ell=1}^{4} \phi_\ell \lambda_{\ell j i} \nu_{\ell k} \right), \tag{21}$$

where λ_{1j} and λ_{3j} are linear functions of the index j and ν_{1k} and ν_{2k} are linear in k. The others are allowed to be non-linear. He shows that the log cross-ratios can be written as

$$\ln \theta^*_{jk} = \eta + \eta^J_j + \eta^K_k + \zeta^J_j \zeta^K_k. \tag{22}$$

This so-called R+C+RC model allows the inclusion of additive effects on the association.

Although the above models provide an elegant description of the association in contingency tables, a disadvantage of the RC family is their cumbersome

forms for the marginal models, especially when there are substantive marginal questions.

To contrast it with the above model, let us return to the Dale model for the specific case of bivariate ordinal data. The model is defined in terms of marginal cumulative logits and global cross-ratios. The cumulative logits η_{1j} and η_{2k} ($j = 1, \ldots, J-1$; $k = 1, \ldots, K-1$), together with the global cross-ratios

$$\ln \psi_{jk} = \ln \left(\frac{\text{pr}(Y_1 \leq j, Y_2 \leq k)\text{pr}(Y_1 > j, Y_2 > k)}{\text{pr}(Y_1 \leq i, Y_2 > k)\text{pr}(Y_1 > j, Y_2 \leq k)} \right) = \ln \frac{\mu_{jk}(1-\mu_{Jk}-\mu_{jK}+\mu_{jk})}{(\mu_{jK}-\mu_{jk})(\mu_{Jk}-\mu_{jk})}, \tag{23}$$

define the joint probabilities. Should it be thought reasonable, then local cross-ratios:

$$\ln \psi^*_{jk} = \ln \frac{\mu^*_{jk}(1-\mu^*_{j+1,k}-\mu^*_{j,k+1}+\mu^*_{jk})}{(\mu^*_{j,k+1}-\mu^*_{jk})(\mu^*_{j+1,k}-\mu^*_{jk})} \tag{24}$$

can be used instead, meaningful for nominal data, but less so for ordinal data since the property of collapsibility is lost (pooling adjacent categories without remaining parameters changing their meaning).

At (23), we pay particular attention to

$$\ln \psi_{jk} = \phi + \rho_{1j} + \rho_{2k} + \sigma_{1j}\sigma_{2k}, \tag{25}$$

including row and column effects, and interactions between rows and columns. This model is identified, e.g., by imposing $\rho_{1J} = \rho_{2K} = \sigma_{1J} = \sigma_{2K} = 0$ and $\sigma_{11} = 1$.

The Goodman and Dale models differ in two important respects. First, the association in the RC model is in terms of local cross-ratios, while the Dale model is based on global cross-ratios. Second, and more importantly, the marginal probabilities of the RC model are complicated functions of the model parameters, whereas the Dale model is expressed directly in terms of the marginal logits, facilitating completely general models. For example, a genuine marginal model can be constructed, with an association function of the RC type. Depending on the data problem, one can opt for local or for global cross-ratios. Lapp, Molenberghs and Lesaffre (1998) have shown that this choice is supported by a very good fit for this kind of model to a range of applications. The global cross-ratio can lead to interesting interpretations of the association structure itself, an often neglected feature, illustrated in the next section.

In spite of the close connection between an RC model and an underlying normal density (Lapp, Molenberghs and Lesaffre (1998)) and the absence of this connection with a fully marginal model, the latter category provides a versatile way of exploring the association structure of cross-classified data, whether of nominal or of ordinal type. We infer from the examples that they often yield

parsimonious descriptions of the association structure. Further, marginal association models are easily extended to marginal regression models to include covariate effects. Both families extend to multi-way tables as well.

Table 1. Fluvoxamine Data. Cross-classification of (a) initial severity and side effects at the second occasion; (b) therapeutic effect at second and third occasions; (c) side effects at the second and third occasions; (d) side effects and therapeutic effect at the second occasion.

Severity	1	2	3	4
1	1	0	1	0
2	21	28	5	5
3	62	62	15	7
4	41	31	6	2
5	1	5	0	1

(a) Side 2

Ther. 2	1	2	3	4
1	13	2	0	0
2	37	40	8	4
3	13	58	18	4
4	1	13	36	21

(b) Therapeutic 3

Side 2	1	2	3	4
1	105	14	0	0
2	34	80	7	1
3	2	7	10	2
4	3	1	0	2

(c) Side 3

Side 2	1	2	3	4
1	8	40	40	40
2	7	45	51	25
3	2	9	8	9
4	2	1	3	9

(d) Therapeutic 2

5.8. Illustration: fluvoxamine data

We illustrate the points of view developed in the previous section using cross-classifications from the fluvoxamine study (Table 1). A summary of model fits is given in Table 2.

Table 1(a) shows a complete lack of association and hence the independence model is accepted for both the Dale and the RC model. Of course, the deviance for the independence model in both families is equal. Initial severity measures symptoms present at baseline, whereas side effects measures symptoms induced by the therapy. Thus the independence model implies that incidence and intensity of side effects do not depend on initial conditions. Since the R+C+RC model is overparameterized, and thus coincides with the saturated model, it is not included in Table 2.

For Table 1(b) we find a strong association main effect with the Dale model. The constant global cross-ratio is high: $\hat{\psi} = \hat{\psi}_{ij} = \exp(2.52) = 12.43$. The fit improves by 7.68 on 2 degrees of freedom if we add a row effect. This model deserves our preference. For the RC family, there is certainly a strong constant

association effect, but the fit is not yet acceptable. A fully satisfactory fit is provided by the row and column association model.

Table 2. Fluvoxamine Data. Deviance χ^2 Goodness-Of-Fit statistics for Dale and RC Models, fitted to the data in Table 1. The models with an acceptable fit are indicated by an asterisk.

	Table 2		Table 3		Table 4		Table 5	
Description	df	χ^2	df	χ^2	df	χ^2	df	χ^2
	Dale Models							
Independence	12	*14.20	9	141.95	9	158.15	9	17.12
Constant Association	11	*11.71	8	*11.48	8	18.27	8	17.12
Row Effects Only	8	*8.34	6	*3.80	6	14.49	6	*9.78
Column Effects Only	9	*11.37	6	*10.26	6	*12.29	6	16.74
Row and Column Effects	6	*8.03	4	*1.29	4	*2.05	4	*9.31
Row, Column, Interactions	2	*0.22	1	*0.31	1	*0.35	1	*0.94
Saturated Model	0	0.00	0	0.00	0	0.00	0	0.00
	RC Models							
Independence	12	*14.20	9	141.95	9	158.15	9	17.12
Constant Association	11	*12.04	8	19.46	8	48.66	8	16.71
Row Effects Only	8	*8.21	6	12.90	6	18.84	6	*11.69
Column Effects Only	9	*11.88	6	14.35	6	45.12	6	15.14
Row and Column Effects	6	*2.22	4	*5.16	4	10.48	4	*1.44
Saturated Model	0	0.00	0	0.00	0	0.00	0	0.00

There is also a clear global association main effect in Table 1(c), having a dramatic effect on model fit, which is further improved by adding row and column effects. Associations are shown in Table 3(c). Some of the observed cross-ratios are infinite, due to observed zero cells. But for one, all associations are very high. High associations in the upper right corner are due to high correlation between side-effects assessments over time; also, they tend to go down. It is remarkable that no RC model fits the data well (Table 2). In conclusion, a marginal model such as the Dale model fits the data better than a model from the RC family. Should one choose to remain within the RC family, then a model of a more elaborate nature might be needed. Related model (20) yields an acceptable fit: $\chi^2 = 6.33$ on 4 degrees of freedom ($P = 0.1760$).

Both Tables 1(b) and 1(c) are cross-classifications of an ordinal variable, recorded at two subsequent measurement times. In both cases, a parsimonious global association model explains the data well. It seems to be much harder to fit these data with local association models.

For Table 1(d), the row effects model is the most parsimonious one that provides an acceptable fit, although caution might dictate keeping column effects

and interactions in the model. Fitted frequencies for both models are shown in Table 1(d). Table 3(c) shows the global cross-ratios for the data of Table 1(d), together with the predicted values under both models. We observe two patterns in Table 3(c). First, the association increases along the main diagonal. This means that the association between the variables $I(\text{SIDE2} \leq 1)$ and $I(\text{THER2} \leq 1)$ is smaller than the association between the variables $I(\text{SIDE2} \leq 3)$ and $I(\text{THER2} \leq 3)$. Also, the association becomes "negative" (i.e., smaller than 1 on the cross-ratio scale) for pairs such as $I(\text{SIDE2} \leq 3)$ and $I(\text{THER2} \leq 1)$. The best RC model is the row and column model. The fitted model is also presented in Table 1(d). All RC models are based on model (18).

Table 3. Fluvoxamine Data. Global cross ratios fitted to the data in Tables 1(c) and 1(d).

Side 2	1	2	3
Observed			
1	21.15	$+\infty$	$+\infty$
2	6.00	31.37	41.74
3	1.17	6.05	43.17
Row and Column Effects			
1	21.07	116.88	760.06
2	5.70	31.65	205.37
3	1.20	6.67	43.26

(c) Side 3

Side 2	1	2	3
Observed			
1	0.97	0.95	0.74
2	0.61	1.33	2.12
3	0.41	2.57	4.26
Column Effects Only			
1	0.86	0.86	0.86
2	1.77	1.77	1.77
3	3.24	3.24	3.24
Row, Column, Interaction			
1	0.92	0.86	0.80
2	0.55	1.55	1.92
3	0.37	2.17	4.00

(d) Therapeutic 2

5.9. Random-effects models

Unlike for correlated Gaussian outcomes, the parameters of the random effects and marginal models for correlated non-Gaussian data describe different types of effects of the covariates on the response probabilities (Neuhaus (1992)). The choice between marginal and random effects strategies should depend heavily on the scientific goals. Marginal models evaluate the overall risk as a function of covariates. With a subject-specific approach, the response rates are modeled as a function of covariates and parameters, specific to a subject, rendering interpretation of fixed-effect parameters conditional on a constant level of the random-effects parameter. Marginal comparisons make no use of within-subject comparisons for within-subject varying covariates and are therefore not useful to assess within-subject effects (Neuhaus, Kalbfleisch and Hauck (1991)).

Whereas the linear mixed model is the most popular choice in the case of Gaussian response variables, there are more options in general. Stiratelli, Laird and Ware (1984) assume the parameter vector to be normally distributed. This idea has been carried further in the work on so-called *generalized linear mixed models* (Breslow and Clayton (1993)). Skellam (1948) introduced the beta-binomial model, in which the response probability of any response of a particular subject comes from a beta distribution. Hence, this model can also be viewed as a random effects model. We consider these in turn.

5.10. The beta-binomial model

Skellam (1948) and Kleinman (1973) assume the success probability P_i of a response within cluster (subject) i to come from a beta distribution with parameters α_i and β_i:

$$\frac{p^{\alpha_i-1}(1-p)^{\beta_i-1}}{B(\alpha_i,\beta_i)}, \qquad 0 \leq p \leq 1,$$

where $B(.,.)$ denotes the beta function. Conditional on P_i, the number of successes Z_i in the ith cluster follows a binomial distribution with mean $\mu_i = n_i\pi_i = n_i\alpha_i/(\alpha_i+\beta_i)$ and variance $\sigma_i^2 = n_i\pi_i(1-\pi_i)[(1+n_i\theta_i)/(1+\theta_i)]$ with $\theta_i = 1/(\alpha_i+\beta_i)$. It can be shown that the intra-cluster correlation is $\rho_i = (\alpha_i+\beta_i+1)^{-1}$.

Generalized linear model ideas can be applied to model the mean parameter π_i (e.g., using a logit link) and the correlation parameter ρ_i (e.g., using Fisher's z transform).

The beta-binomial is, just as the linear mixed model, an example of a model that can be given a hierarchical as well as a marginal intepretation. In particular, the hierarchical view can be adopted to conveniently arrive at a marginal model.

5.11. Generalized linear mixed models

Perhaps the most commonly encountered subject-specific model is the generalized linear mixed model. Assume the data setting is the same as in Section 3. A general framework for mixed-effects models for longitudinal data can be expressed as follows. Assume that $\boldsymbol{Y_i}$ (possibly appropriately transformed) satisfies

$$\boldsymbol{Y_i}|\boldsymbol{b_i} \sim F_i(\boldsymbol{\theta}, \boldsymbol{b_i}), \tag{26}$$

i.e., conditional on $\boldsymbol{b_i}$, $\boldsymbol{Y_i}$ follows a pre-specified distribution F_i, possibly depending on covariate matrices X_i and Z_i (suppressed from notation), and parameterized through a vector $\boldsymbol{\theta}$ of unknown parameters, common to all subjects. Further, $\boldsymbol{b_i}$ is a q-dimensional vector of subject-specific parameters, called random effects, assumed to follow a so-called mixing distribution G which may depend on a vector $\boldsymbol{\psi}$ of unknown parameters, i.e., $\boldsymbol{b_i} \sim G(\boldsymbol{\psi})$. The $\boldsymbol{b_i}$ reflect the between-unit

heterogeneity in the population with respect to the distribution of $\boldsymbol{Y_i}$. In the presence of random effects, conditional independence (upon $\boldsymbol{b_i}$) is often assumed.

In general, unless a fully Bayesian approach is followed, inference is based on the marginal model for $\boldsymbol{Y_i}$ which is obtained from integrating out the random effects over their distribution $G(\boldsymbol{\psi})$ (Fahrmeir and Tutz (2001)). If $f_i(\boldsymbol{y_i}|\boldsymbol{b_i})$ and $g(\boldsymbol{b_i})$ denote the density functions corresponding to the distributions F_i and G, respectively, we have the marginal density function of $\boldsymbol{Y_i}$ as

$$f_i(\boldsymbol{y_i}) = \int f_i(\boldsymbol{y_i}|\boldsymbol{b_i}) g(\boldsymbol{b_i}) d\boldsymbol{b_i}, \tag{27}$$

which depends on the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. Assuming independence of the units, estimates of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\psi}}$ can be obtained from maximizing the likelihood function built from (27), and inferences immediately follow from classical maximum likelihood theory.

It is important to realize that the random-effects distribution G is crucial in the calculation of the marginal model (27). One approach is to leave G unspecified and to use non-parametric maximum likelihood (NPML, McLachlan and Peel (2000)) estimation, which maximizes the likelihood over all possible distributions G. The resulting estimate $\widehat{G}$ is discrete with finite support. Depending on the context, this may or may not be a realistic reflection of the true heterogeneity between units. One therefore often assumes G to be of a parametric form, such as a (multivariate) normal. Depending on F_i and G, the integration in (27) may or may not be analytically possible. Proposed solutions are based on Taylor series expansions of $f_i(\boldsymbol{y_i}|\boldsymbol{b_i})$, or on numerical approximations of the integral, such as (adaptive) Gaussian quadrature (Pinheiro and Bates (1995)).

Although one is usually primarily interested in estimating the parameters in the marginal model, it is often necessary to calculate estimates for the random effects $\boldsymbol{b_i}$ as well, e.g., for predictive purposes or to detect special profiles, outlying individuals, or groups of individuals evolving differently in time. Inference for the random effects is often based on their posterior distribution $f_i(\boldsymbol{b_i}|\boldsymbol{y_i})$, given by

$$f_i(\boldsymbol{b_i}|\boldsymbol{y_i}) = \frac{f_i(\boldsymbol{y_i}|\boldsymbol{b_i})\ g(\boldsymbol{b_i})}{\int\ f_i(\boldsymbol{y_i}|\boldsymbol{b_i})\ g(\boldsymbol{b_i})\ d\boldsymbol{b_i}}, \tag{28}$$

in which the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are replaced by estimates obtained from maximizing the marginal likelihood. The mean or mode corresponding to (28) can be used as point estimates for $\boldsymbol{b_i}$, yielding empirical Bayes (EB) estimates.

There are two major differences with the linear mixed model. First, the marginal distribution of $\boldsymbol{Y_i}$ can no longer be calculated analytically, complicating

the computation of the MLE for $\boldsymbol{\beta}$, D, and the parameters in all Σ_i. As a result, the marginal covariance structure does not immediately follow, such that it is not always clear in practice what assumptions a specific model implies with respect to the underlying variance function and the underlying correlation structure in the data.

A second difference is related to the interpretation of the fixed effects $\boldsymbol{\beta}$. Under the linear model (1), that the fixed effects have a subject-specific as well as a marginal interpretation: the elements in $\boldsymbol{\beta}$ reflect the effect of specific covariates, conditionally on $\boldsymbol{b_i}$, as well as marginalized over these random effects. Under non-linear mixed models, this does not generally hold. The fixed effects now only reflect the conditional effect of covariates and the marginal effect is not easily obtained anymore, as $\mathrm{E}(\boldsymbol{Y_i})$ is given by $\mathrm{E}(\boldsymbol{Y_i}) = \int \boldsymbol{y_i} \int f_i(\boldsymbol{y_i}|\boldsymbol{b_i}) g(\boldsymbol{b_i}) d\boldsymbol{b_i} d\boldsymbol{y_i}$, which, in general, is *not* of the form $h(X_i, Z_i, \boldsymbol{\beta}, \mathbf{0})$.

Only for very particular models, can (some of) the fixed effects still be interpreted as marginal covariate effects. For example, consider the model where, apart from an exponential link function, the mean is linear in the covariates, and the only random effects in the model are intercepts. More specifically, this corresponds to the model with $h(X_i, Z_i, \boldsymbol{\beta}, \boldsymbol{b_i}) = \exp(X_i\boldsymbol{\beta} + Z_i b_i)$, in which Z_i is a vector containing only ones. The expectation of $\boldsymbol{Y_i}$ is now given by

$$\mathrm{E}(\boldsymbol{Y_i}) = \mathrm{E}\left[\exp(X_i\boldsymbol{\beta} + Z_i b_i)\right] = \exp(X_i\boldsymbol{\beta})\ E\left[\exp(Z_i b_i)\right], \tag{29}$$

which shows that, except for the intercept, all parameters in $\boldsymbol{\beta}$ have a marginal interpretation.

The generalized linear mixed model (GLMM, Breslow and Clayton (1993) and Wolfinger and O'Connell (1993)) is the most frequently used random-effects model for discrete outcomes. A general formulation is as follows. Conditionally on random effects $\boldsymbol{b_i}$, it assumes that Y_{ij} are independent, with density function of the form (13) with mean $\mathrm{E}(Y_{ij}|\boldsymbol{b_i}) = a'(\eta_{ij}) = \mu_{ij}(\boldsymbol{b_i})$ and variance $\mathrm{Var}(Y_{ij}|\boldsymbol{b_i}) = \phi a''(\eta_{ij})$, and with linear predictor $h(\boldsymbol{\mu_i}(\boldsymbol{b_i})) = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}$. The linear mixed model is a special case with identity link function. The random effects $\boldsymbol{b_i}$ are assumed to be sampled from a (multivariate) normal distribution with mean $\mathbf{0}$ and covariance matrix D. When the link function is chosen to be of the logit form and the random effects are assumed to be normally distributed, the familiar logistic-linear GLMM follows.

The non-linear nature of the model again implies that the marginal distribution of $\boldsymbol{y_i}$ is, in general, not easily obtained. An exception to this occurs when the probit link is used. Further, as was also the case for non-linear mixed models, the parameters $\boldsymbol{\beta}$ have no marginal interpretation, except for some very particular models such as count data with log link (Liang, Zeger and Qaqish (1992)).

As an important example, consider the binomial model for binary data with the logit canonical link function, and where the only random effects are intercepts b_i. It can be shown that the marginal mean $\boldsymbol{\mu_i} = \mathrm{E}(Y_{ij})$ satisfies $h(\boldsymbol{\mu_i}) \approx X_i\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* = [c^2\mathrm{Var}(b_i)+1]^{-1/2}\boldsymbol{\beta}$, in which c equals $16\sqrt{3}/15\pi$ (Wang, Lin, Gutierrez and Carroll (1998)). Hence, although the parameters $\boldsymbol{\beta}$ in the generalized linear mixed model have no marginal interpretation, they do show a strong relation to their marginal counterparts. As a consequence, larger covariate effects are obtained under the random-effects model in comparison to the marginal model.

5.12. Marginal versus random-effects models

Fitting several marginal models often produces similar parameter estimates and standard errors. This is totally different when models across model families are considered. This has led to a lot of confusion, including discussions as to the nature of this bias. Such a discussion is ill-founded, since the parameters underlying marginal, random-effects, and conditional models, are *different* at the population level. Only in some cases (e.g., the linear mixed model) are there easy connections between them. Thus, intuition borrowed from linear mixed models can be misleading.

To see this, consider a binary outcome variable and assume a random-intercept logistic model with linear predictor $\mathrm{logit}[P(Y_{ij} = 1|t_{ij}, b_i] = \beta_0 + b_i + \beta_1 t_{ij}$, where t_{ij} is the time covariate. The conditional means $E(Y_{ij}|b_i)$, as functions of t_{ij}, are given by

$$E(Y_{ij}|b_i) = \frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}, \tag{30}$$

whereas the marginal average evolution is obtained from averaging over the random effects:

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E\left[\frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}\right] \neq \frac{\exp(\beta_0 + \beta_1 t_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij})}. \tag{31}$$

A graphical representation of both (30) and (31) is given in Figure 2. This implies the interpretation of the parameters in both types of model is completely different. A schematic display is given in Figure 3. Depending on the model family (marginal or random-effects), one is led to either marginal or hierarchical inference. In general, the parameter $\boldsymbol{\beta}^M$ in a marginal model is different from the parameter $\boldsymbol{\beta}^{RE}$ even when the latter is estimated using marginal inference. Some of the confusion results from the equality of these parameters in the linear mixed model. When a random-effects model is considered, the marginal mean profile can be derived, but it will generally not produce a simple parametric form.

In Figure 3 this is indicated by putting the corresponding parameter between quotes. While this issue arises in the logistic random-effects model, it does not in the probit version since then the marginal model is of closed form and again of probit type (Renard, Molenberghs and Geys (2004)).

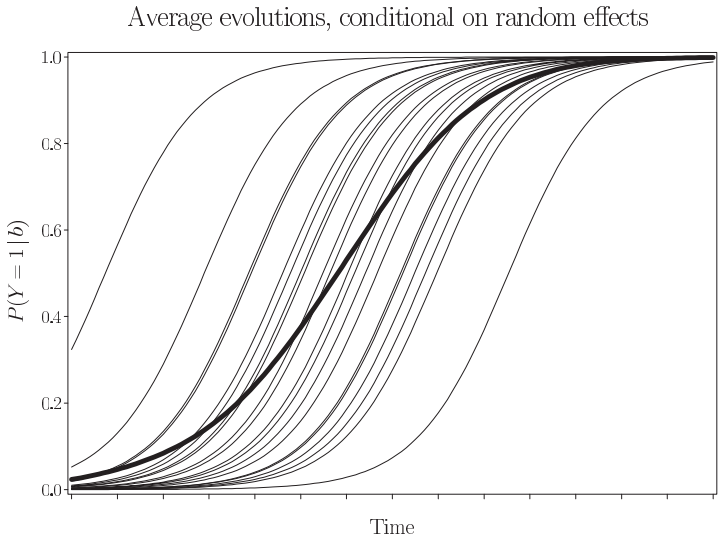

Figure 2. Graphical representation of a random-intercept logistic curve, across a range of levels of the random intercept, together with the corresponding marginal curve.

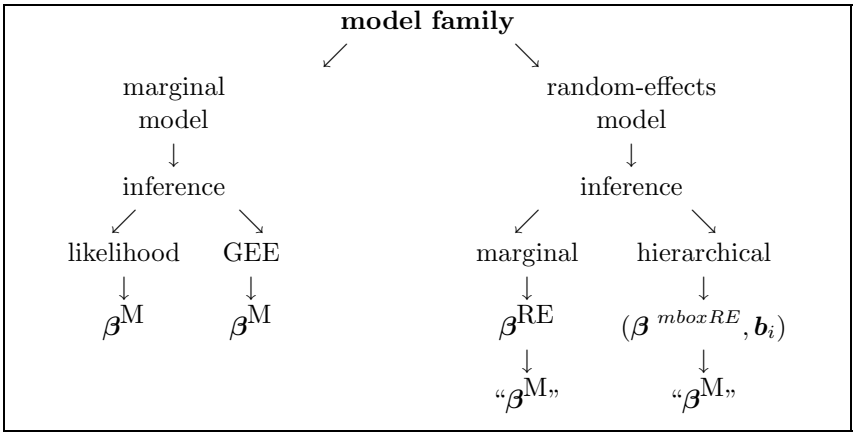

Figure 3. Representation of model families and corresponding inference. A superscript ‘M’ stands for marginal, ‘RE’ for random effects. A parameter between quotes indicates that marginal functions but no direct marginal parameters are obtained.

This discussion points to the need to carefully reflect on the choice of the model for the responses and the distribution of the random effects. The choice for logistic-normal random-effects models is based on the combination of the familiar logistic model with linear mixed model ideas. However, some of the nice properties of the logistic model do not carry over to the random-effects setting. For the logistic case with random intercepts, the following approximate relationship holds between the marginal and random-effects parameters:

$$\frac{\widehat{\boldsymbol{\beta}}^{\mathrm{RE}}}{\widehat{\boldsymbol{\beta}}^{\mathrm{M}}} = \sqrt{c^2\sigma^2 \ + \ 1} \ > \ 1, \tag{32}$$

where σ^2 is the variance of the random intercepts and $c^2 = 16\sqrt{3}/15\pi$.

5.13. Illustration: toenail data

Table 4 displays parameter estimates (standard errors) for a marginal model (GEE with unstructured working assumptions) and a random-effects model (GLMM). The logit function, conditional upon the random intercept, takes the form:

$$\mathrm{logit}[P(Y_{ij} = 1|T_i, t_{ij}, b_i)] = \beta_0 + b_i + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij} \tag{33}$$

with t_{ij} the time of measurement j on subject i, $T_i = 0$ in Group A and $T_i = 1$ otherwise. The random intercepts b_i are assumed normal with mean 0 and variance σ^2. There is a huge difference between the parameter estimates. Of course, (32) equals 2.56, well in line with Table 4.

Table 4. Toenail Data. Parameter estimates (standard errors) for a generalized linear mixed model (GLMM) and a marginal model (GEE), as well as the ratio between both sets of parameters.

	GLMM	GEE	
Parameter	Estimate (s.e.)	Estimate (s.e.)	Ratio
Intercept group A	−1.63 (0.44)	−0.72 (0.17)	2.26
Intercept group B	−1.75 (0.45)	−0.65 (0.17)	2.69
Slope group A	−0.40 (0.05)	−0.14 (0.03)	2.87
Slope group B	−0.57 (0.06)	−0.25 (0.04)	2.22
Random int. var.		4.02	

In Figure 4, the marginal evolutions obtained with GEE, are very similar to those obtained from marginalizing a GLMM. In contrast, within a GLMM, the marginal evolutions differ sharply from the evolutions conditional upon the random effect being equal to $b_i = 0$. These observations are in agreement with (31) and the difference between the $\boldsymbol{\beta}^{\mathrm{M}}$ and $\boldsymbol{\beta}^{\mathrm{RE}}$ parameters in Figure 3.

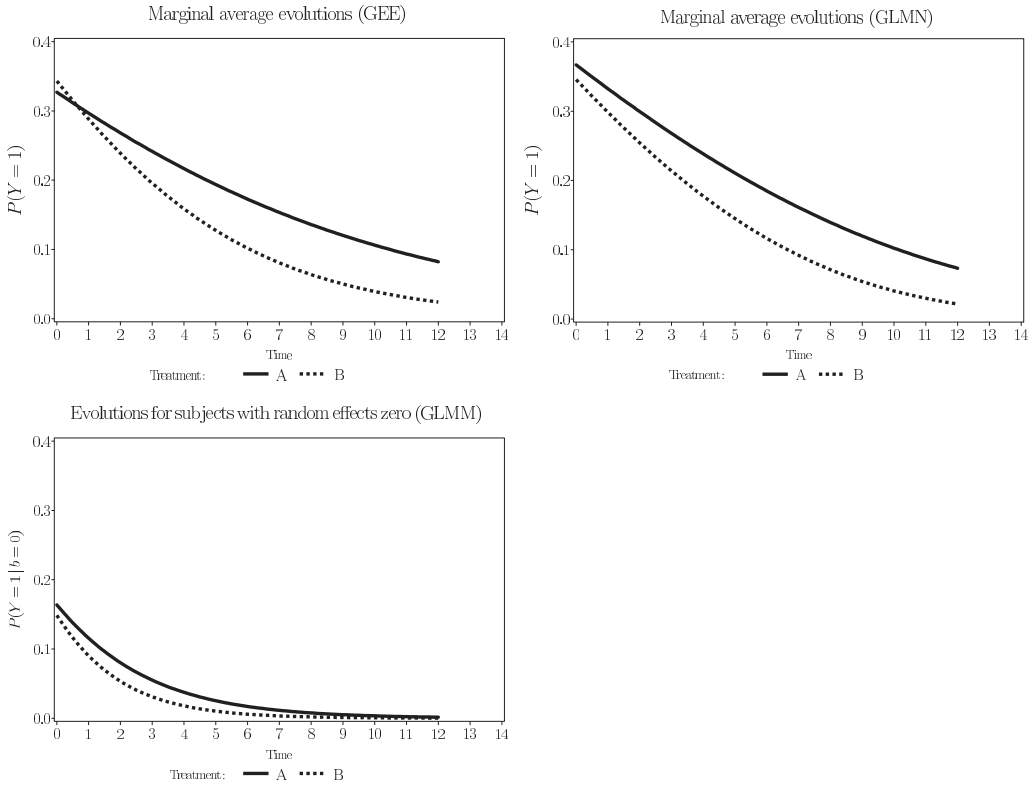

Figure 4. Toenail Data. Treatment-arm specific evolutions. (a) Marginal evolutions as obtained from a marginal (GEE) model, (b) marginal evolutions as obtained from integrating out a GLMM, and (c) evolutions for an "average" subject from a GLMM, i.e., with $b_i = 0$.

6. Concluding Remarks

Through a number of simple yet commonly used settings, we have illustrated that one needs to reflect very carefully on the mathematical and substantive meaning behind a parametric model of choice.

In a repeated measures setting with normally distributed outcomes, the linear mixed model is the most commonly used tool. Nevertheless, the model is not free from issues. First, one has to reflect carefully on the differences between a hierarchical and a marginal point of view. This choice is important, not only for parameter interpretation, but also for inferences on variance components. When outcomes are non-normal, one has to reflect very carefully upon the differences between the marginal, conditional, and random-effects families. In each of the families, a number of models have been formulated, many of which reduce to logistic regression in the case of independence. Nevertheless, there are dramatic differences between them and ideally the substantive question to be answered

should drive the choice of model family and ultimately the particular model chosen within such a family. Among the marginal models, the Bahadur model, the success-probability model, and the George-Bowman folded logistic model suffer from serious drawbacks. The multivariate probit and odds-ratio models have some promise; they provide flexible ways of modelling both the individual outcomes as well as the association between them. Unfortunately, they become computationally intractable for large clusters, but generalized estimating equations and pseudo-likelihood methods come to the rescue. Conditional models are easier from a computational point of view but we have illustrated they suffer from serious problems in terms of meaningfulness, especially but not only when cluster sizes are unequal.

Within the random-effects family, the generalized linear mixed model for binary data with logit link has become very popular. Nevertheless, the combination of a logit link with normally distributed random effects poses unique computational and interpretational challenges. Indeed, it is important to understand the main differences between the linear mixed model and the generalized linear mixed model, particularly if of logistic-linear type. In the first case, all properties of the normal distribution can be invoked, while in the second case one typically resorts to the exponential family. In the normal distribution, there is no mean-variance link, while such a link plays a prominent place in most exponential family models. In addition, the link function is linear in the first case and usually non-linear in the second case. In the linear mixed model case, the sources of variability all enter the same linear predictor as additive terms. However, there is no additive relationship between them in other settings. To see this, consider the logistic-linear model. An outcome can be written, with obvious notation, as $Y_{ij} = \mu_{ij} + \varepsilon_{ij}$. Thus, while the measurement error is linked linearly to the outcome, the random-effects variability enters non-linearly since the linear predictor is coupled to the mean μ_{ij} via the link function. Thus, not only model fitting is more involved in the generalized linear mixed model case, also a number of interpretational differences follow, including a different meaning for the regression parameters in both types of models. Whether marginal or hierarchical inference is chosen in the GLMM case, the resulting parameters refer to non-marginal population quantities. Marginalization is possible, but will generally provide functions of an intractable form, the use of which is primarily graphical. Some of these issues are alleviated when the beta-binomial model or the probit random-effects model is chosen. In the first case, the parameters have, at the same time, a marginal and a random-effects interpretation, while in the second case the link between both sets of parameters exists in closed form. When outcomes are of the count type, the Poisson-normal model enjoys a simple relationship between the random-effects model and the induced marginal model.

Acknowledgements

We gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

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

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Albert, A. and Lesaffre, E. (1981). Multiple group logistic discrimination. *Comput. Math. Appl.* **12**, 209-224.

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* (Edited by H. Solomon). Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press, Stanford, California.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113-120.

Cox, D. R. and Hinkley, D. V. (1990). *Theoretical Statistics*. Chapman and Hall, London.

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

De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I. and De Keyser, P. (1998). Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: a double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day. *J. Amer. Acad. Dermat.* **38**, S57-63.

Declerck, L., Aerts, M. and Molenberghs, G. (1998). Behaviour of the likelihood ratio test statistic under a Bahadur model for exchangeable binary data. *J. Statist. Comput. Simulation* **61**, 15-38.

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

Ekholm, A. (1991). Fitting regression models to a multivariate binary response. In *A Spectrum of Statistical Thought: Essays in Statistical Theory, Economics, and Population Genetics in Honour of Johan Fellman* (Edited by G. Rosenqvist, K. Juselius, K. Nordström and J. Palmgren), 19-32. Swedish School of Economics and Business Administration, Helsinki.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, Heidelberg.

George, E. O. and Bowman, D. (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics* **51**, 512-523.

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

Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B* **57**, 533-546.

Goodman, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **76**, 320-334.

Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. *J. Amer. Statist. Assoc.* **68**, 46-54.

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicology experiments. *Biometrics* **34**, 69-76.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89**, 625-632.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38** 963-974.

Lapp, K., Molenberghs, G. and Lesaffre, E. (1998). Local and global cross ratios to model the association between ordinal variables. *Comput. Statist. Data Anal.* **28**, 387-411.

le Cessie, S. and van Houwelingen, J. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics* **51**, 600-614.

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

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc. Ser. B* **54**, 3-40.

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.

McCullagh, P. (2002). What is a statistical model? *Ann. Statist.* **30**, 1225-1310.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.* **89**, 633-644.

Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statist. Medicine* **18**, 2237-2255.

Nelder, J. A. (1954). The interpretation of negative components of variance. *Biometrika* **41**, 544-548.

Neuhaus, J. M. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statist. Meth. Medical Res.* **1**, 249-273.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991). A comparison of cluster-specific and marginal approaches for analyzing correlated binary data. *Internat. Statist. Rev.* **59**, 25-35.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Statist.* **4**, 12-35.

Plackett, R. L. (1965). A class of bivariate distributions. *J. Amer. Statist. Assoc.* **60**, 516-522.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Raubertas, R. F., Lee, C. I. C. and Nordheim, E. V. (1986). Hypothesis tests for normal means constrained by linear inequalities. *Comm. Statist. Theory Method* **15**, 2809-2833.

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comput. Statist. Data Anal.* **44**, 649-667.

Rosner, B. (1984). Multivariate methods in opthalmology with applications to other paired-data. *Biometrics* **40**, 1025-1035.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *Internat. Statist. Rev.* **56**, 49-62.

Silvapulle, M. J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *J. Amer. Statist. Assoc.* **90**, 342-349.

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. *J. Roy. Statist. Soc. Ser. B* **10**, 257-261.

Stiratelli, R., Laird, N. and Ware, J. H. (1984). Random-effects model for serial observations with binary response. *Biometrics* **40**, 961-971.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171-1177.

Stram, D. A. and Lee, J. W. (1995). Correction to: variance components testing in the longitudinal mixed effects model. *Biometrics* **51**, 1196.

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

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics. Springer-Verlag, New-York.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254-262.

Wang, N., Lin, X., Gutierrez, R. G. and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *J. Amer. Statist. Assoc.* **93**, 249-261.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *J. Statist. Comput. Simulation* **48**, 233-243.

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

Biostatistics, Center for Statistics, Limburgs Universitair Centrum, tUL, Universitaire Campus, B-3590 Diepenbeek, Belgium.

E-mail: geert.molenberghs@luc.ac.be

Biostatistical Centre, Catholic University of Leuven, U.Z. St.-Rafaël, Kapucijnenvoer 35, B-3000 Leuven, Belgium.

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

(Received January 2003; accepted June 2003)