

Discussion Contribution to Statistical Science

“Likelihood Inference for Models with Unobservables: Another View”

Geert Molenberghs^{1,2} Michael G. Kenward³ Geert Verbeke^{2,1}

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

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

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

1 Introduction

We are grateful for the opportunity to comment on Professors Lee and Nelder's work. Like their other, related papers, they are replete with important ideas and thought-provoking theses. We will take up just a few of the topics touched upon, and offer some reflections. In Section 2, we consider models with, and inferences for, unobservables. Section 3 revisits one of the counterexamples discussed by the authors, which turns out to have insightful connections with the Cauchy distribution. Connections between generalized estimating equations and fully specified joint distributions are touched upon in Section 4. Finally, the position of the authors' computational proposals among alternative routes is explored in Section 5.

2 The Nature of Unobservables

Although Lee and Nelder claim, in their Section 4.4, that unobservables are often verifiable, because the unobservables are latent variables for observed data, we would like to issue a word of caution. Evidently, using the authors' notation, variance components λ and ϕ are identifiable from the data, whenever there is replication within the i levels (e.g., repeated measures j on subject i , or litter mates j corresponding to dam i). However, such data-based verification is confined, in the strict sense, to the way the *induced marginal model* is accurate. This model is of a compound-symmetry type:

$$\mathbf{Y}_i \sim N(\mathbf{1}\beta, V_i = \lambda J_{n_i} + \phi I_{n_i}). \quad (1)$$

In other words, one can assess from the data whether V_i is an accurate description of the variance-covariance structure, just as one can verify whether the implied constant mean is adequate. But whether this provides an accurate description of the unobservables, is another matter because there is a many-to-one map of hierarchical models to (1).

To see this more clearly, we start from this compound-symmetry setting and take an example of Molenberghs and Verbeke (2009). Consider the random-intercepts model:

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\xi} + b_i + \varepsilon_{ij}, \quad (2)$$

where Y_{ij} is the response for member $j = 1, \dots, n_i$ of cluster $i = 1, \dots, N$, \mathbf{x}_{ij} is a vector of known covariate values, $\boldsymbol{\xi}$ is a vector of unknown regression coefficients, $b_i \sim N(0, \lambda^2)$ is a cluster-specific random effect, assumed to be independently distributed from the residual error components $\varepsilon_{ij} \sim N(0, \nu^2)$. The implied marginal model is obtained by integrating (2) over the random effects. Grouping the Y_{ij} into a vector \mathbf{Y}_i and assembling the rows \mathbf{x}'_{ij} into a matrix X_i , this marginal distribution is

$$\mathbf{Y}_i \sim N(X_i\boldsymbol{\xi}, \lambda^2 J_{n_i} + \nu^2 I_{n_i}), \quad (3)$$

in which I_{n_i} denotes the identity matrix of dimension n_i , and where J_{n_i} equals the $n_i \times n_i$ matrix containing only ones. Evidently, (1) is a particular instance of (3).

Traditionally, there have been two views regarding the variance component λ^2 in the above model. In the first, where the focus is entirely on the resulting marginal model (3), negative values for λ^2 are perfectly acceptable (Nelder 1954, Verbeke and Molenberghs 2000, Sec. 5.6.2), because this merely corresponds to the occurrence of negative within-cluster correlation $\rho = \lambda^2/(\lambda^2 + \nu^2)$. In such a case, the only requirement is that $\lambda^2 + \nu^2 > 0$ and $V_i = \lambda^2 J_{n_i} + \nu^2 I_{n_i}$ is positive definite. In the second view, when the link between the marginal model (3) and its generating hierarchical model (2) is preserved, thereby including the concept of random effects b_i and perhaps even requiring inferences for them, it has been considered imperative to restrict λ^2 to nonnegative values.

In such a view, it is implicit that any hierarchical model, corresponding to compound-symmetry model (3), should be of the form (2). But this is not the case. To see this, we first reiterate that the hierarchical model, corresponding to a given marginal model, is non-unique. This originates from the random effects' latency and is crucial to the theme of Lee and Nelder's current paper.

To illustrate this non-uniqueness, consider the simple but insightful case of two measurements per subject, i.e., $n_i = 2$. We contrast two models, the first one of the form (2), with random intercepts $b_i \sim N(0, \lambda^2)$ and heterogeneous errors $\varepsilon_{ij} \sim N(0, \nu_j^2)$, ($j = 1, 2$). The second takes the form:

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\xi} + b_{0i} + b_{1i}(j - 1) + \varepsilon_{ij}, \quad (4)$$

with two uncorrelated random effects

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{pmatrix} \right]$$

and homogeneous error $\varepsilon_{ij} \sim N(0, \nu^2)$. The marginal means are, obviously, equal. At the same time, the marginal variance-covariance matrix in the first model is:

$$V^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (\lambda^2) \begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} \nu_1^2 & 0 \\ 0 & \nu_2^2 \end{pmatrix} = \begin{pmatrix} \lambda^2 + \nu_1^2 & \lambda^2 \\ \lambda^2 & \lambda^2 + \nu_2^2 \end{pmatrix} \quad (5)$$

and the counterpart for the second model is

$$V^{(2)} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \nu^2 & 0 \\ 0 & \nu^2 \end{pmatrix} = \begin{pmatrix} \lambda_1^2 + \nu^2 & \lambda_1^2 \\ \lambda_1^2 & \lambda_1^2 + \lambda_2^2 + \nu^2 \end{pmatrix}. \quad (6)$$

Evidently, $V^{(1)}$ and $V^{(2)}$ are equivalent, through the linear relationships $\lambda_1^2 = \lambda^2$, $\lambda_2^2 = \nu_2^2 - \nu_1^2$, and $\nu^2 = \nu_1^2$. What this means, in this case, is that the observed heterogeneity in variance can be ascribed to either heterogeneous residual errors or to the presence of a random slope. The fitted marginal model, and hence the data, cannot be used to distinguish between these two scenarios. Thus, one view is that fitting a marginal model comes with an entire equivalence class of hierarchical models that reduce to the given marginal model.

We now show that another simple extension of the distributional assumptions of (2) allows for negative variance components, while maintaining the model's random-intercepts interpretation.

We retain the random-intercepts model (2), but now with the assumption

$$\begin{pmatrix} b_i \\ \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} d & \tau & \dots & \tau \\ \tau & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \tau & 0 & \dots & \sigma^2 \end{pmatrix} \right]. \quad (7)$$

The induced conditional distribution of the measurement error vector, given the random intercept, is

$$\varepsilon_i | b_i \sim N \left[\frac{\tau b_i}{d} \mathbf{j}_{n_i}, \frac{1}{d} (d\sigma^2 I_{n_i} - \tau^2 J_{n_i}) \right]. \quad (8)$$

Here, \mathbf{j}_{n_i} is a n_i -vector of ones. Note that (7) allows for marginally uncorrelated measurement errors that become correlated, *conditional upon the random effect*.

The corresponding marginal model is

$$\mathbf{Y}_i \sim N \left[X_i \boldsymbol{\xi}, (d + 2\tau) J_{n_i} + \sigma^2 I_{n_i} \right]. \quad (9)$$

Starting from a conventional compound-symmetry model, it is clear that $\sigma^2 = \nu^2$ and $d + 2\tau = \lambda^2$. Evidently, d and τ are not jointly identified, pointing to a collection of hierarchical models that all yield the same marginal model and hence are indistinguishable based on the data alone. To define the range of this collection, it is necessary for $\tau^2 \leq d\sigma^2$. Together with $\tau = (\lambda^2 - d)/2$, this leads to the set of allowable solutions

$$d = \lambda^2 + 2\nu^2 + 2\nu\alpha\sqrt{\lambda^2 + \nu^2}, \quad (10)$$

with $\alpha \in [-1, 1]$. From (10), we find

$$\tau = -(\nu^2 + \nu\alpha\sqrt{\lambda^2 + \nu^2}). \quad (11)$$

In other words, we have a decomposition of the variance and covariance:

$$\begin{aligned} \text{var}(Y_{ij}) &= \overset{(\sigma^2)}{\nu^2} + \overset{(d)}{(\lambda^2 + 2\nu^2 + 2\nu\alpha\sqrt{\lambda^2 + \nu^2})} + \overset{(2\tau)}{(-2\nu^2 - 2\nu\alpha\sqrt{\lambda^2 + \nu^2})}, \\ \text{cov}(Y_{ij}, Y_{ik}) &= (\lambda^2 + 2\nu^2 + 2\nu\alpha\sqrt{\lambda^2 + \nu^2}) + (-2\nu^2 - 2\nu\alpha\sqrt{\lambda^2 + \nu^2}). \end{aligned}$$

Observe that, when λ^2 is positive, $\tau = 0$ is among the solutions, i.e., this recovers the conventional random-intercepts model with uncorrelated errors. However, when $\lambda^2 < 0$, then all values for τ are necessarily negative.

This construction reduces the conventional random-intercepts model (2) to a special case of the extended family of hierarchical models, with correlation between random effects and measurement errors. Once again, it is clear that one can make inferences about the random effects, *given* that one is prepared to make strong, and unverifiable assumptions about the hierarchical model, stemming from the many-to-one map from hierarchical models to the implied marginal models. In this sense, the above derivations underscore, not just that every compound-symmetry model can be induced by a hierarchical model, but that an entire collection of random-intercepts models fulfills this role. These differ from each other by the degree to which the random intercepts and measurement errors are correlated.

The above considerations focus on random effects. This is but one example of unobservables. Like Lee and Nelder, Verbeke and Molenberghs (2009) argue that so-called augmented data, in the sense of supplementing the observed data with latent, unobserved structures, is common throughout statistics. Examples include models for incompletely observed data, describing observed and unobserved outcomes alike, random-effects models, latent class models, latent variable models, censored survival data, etc. Heitjan and Rubin (1991) and Zhang and Heitjan (2007) have unified some of these settings in a concept called *coarsening*, broadly referring to the fact that the observed data are coarser than the hypothetically conceived data structures, while models target the latter. Generally, models for augmented structures are identifiable only by virtue of making sometimes strong but always partially unverifiable modeling assumptions. These settings taken together are termed *enriched data* by Verbeke and Molenberghs (2009). Of course, there is a subtle distinction between both concepts. In the coarse-data setting, it is understood that a part of the data would ideally be observed but are not in practice (e.g., actual survival time after censoring, outcomes after dropout, etc.). Augmented data refers rather to the addition of useful but artificial constructs to the data setting, such as random effects, latent classes, latent variables, which are never directly observed. Such augmentations permit simple model development and represent a very powerful tool to succinctly accommodate posited, potentially very complex, often causal, real-world structures, a fact of which Lee and Nelder also make use.

Verbeke and Molenberghs (2009) show that every model for enriched-data settings can be factored as a product of two components: the first one, termed the marginal model, fully identifiable from the observed data; the second one, the conditional distribution of the enriched data given the observed data, entirely arbitrary. The evident consequence is that the identification of such a part can come from assumptions only and points at the same time to the considerable risk for conclusions to be sensitive to such assumptions, and ultimately to the need for conducting sensitivity analyses. It implies that such non-identified parts can be replaced arbitrarily, without altering the fit to the observed data but with potentially huge implications for inferences and substantive conclusions. Put simply, one's inferential conclusions may strongly depend on such unverifiable portions of the model.

In the missing data case, studied in more detail by Molenberghs *et al* (2008), one could identify the second factor by requiring, for example, that it is of the MAR type. This means that every model assuming that the missing data are missing not at random corresponds to another model, producing exactly the same fit to the observed data, but now assuming that the missing data are missing at random. In the context of a conventional linear mixed model, Verbeke and Molenberghs (2009) illustrated the implications of the result by replacing the conditional distribution of the random effects given the data, i.e., the random effects' posterior, by two families of exponential distributions, special cases of the gamma family, for the sake of illustration. This nicely supplements the above compound-symmetry model with correlated random effects and measurement errors.

These results imply that one should not simply adopt a hierarchical model, only because it is convenient, is in common use, etc. Rather, one should carefully reflect on that part of the model that cannot be critiqued by the data. One should strive for (1) better understanding of the dependence of one's inferences on non-verifiable model components and (2) developing sensitivity analysis tools regarding substantive conclusions with respect to data enrichment. Generally speaking, inferences relative to observed data only, such as fixed-effects and variance-component parameter estimates, are unaffected by the choice of enrichment model. However, such aspects as empirical Bayes predictions in linear mixed models, or predictive distributions of unobserved measurements given observed ones, strongly rest on unverifiable modeling assumptions. This points to the need for sensitivity analysis. Rather than fitting a single model and putting blind belief in it, it is more reasonable to consider a discrete or continuous set of alternative model formulations, and assess how key inferences are vulnerable to choices made. Molenberghs and Kenward (2007) discuss avenues for sensitivity analysis.

As soon as one is aware of the lack of identification, there is reasonable latitude in making pragmatic identification choices. For example, with random effects or other latent structures, one could express a preference for conjugate priors (Lee and Nelder 1996, 2001, 2003) because, in the absence of identification, the convenience and appeal of conjugacy may be invoked.

3 Bayarri's Example and a Cauchy-type Family of Distributions

Lee and Nelder, in their Section 4.2, revisit Bayarri's example. There is something rather peculiar about it, because it is of a Cauchy type. We will show this in what follows. Owing to the model's absence of finite moments, it seems natural that an estimation method ought to encounter problems. Indeed, any approach that does purport to provide estimates in such circumstances must raise concerns about its properties.

Consider a Weibull model for repeated measures with gamma random effects:

$$f(\mathbf{y}_i | \boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \lambda \rho \theta_{ij} y_{ij}^{\rho-1} e^{\mathbf{x}'_{ij} \boldsymbol{\xi}} e^{-\lambda y_{ij}^{\rho} \theta_{ij} e^{\mathbf{x}'_{ij} \boldsymbol{\xi}}}, \quad (12)$$

$$f(\boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \frac{1}{\beta_j^{\alpha_j} \Gamma(\alpha_j)} \theta_{ij}^{\alpha_j-1} e^{-\theta_{ij}/\beta_j}. \quad (13)$$

Here, i and j are as in Section 2, θ_{ij} are gamma random effects, \mathbf{x}_{ij} are covariates, $\boldsymbol{\xi}$ regression parameters, ρ the Weibull shape parameters, and α_j and β_j the parameters governing the gamma distribution for the j th component.

Rather than the above two-parameter gamma density, it is customary in a gamma frailty context (Duchateau and Janssen 2007) to set $\alpha_j \beta_j = 1$, for reasons of identifiability.

in line with Bayarri's example, we use the less conventional constraint $\alpha_j = 1$ and $\beta_j = 1/\delta_j$, leading to

$$f(\boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \delta_j e^{-\delta_j \theta_{ij}} \quad (14)$$

and implying that the gamma density is reduced to an exponential one. Details can be found in Molenberghs *et al* (2009).

The moments take the form:

$$E(Y_{ij}^k) = \frac{\delta_j^{k/\rho} k}{\lambda^k} \Gamma\left(1 - \frac{k}{\rho}\right) \Gamma\left(\frac{k}{\rho}\right) \exp\left(-\frac{k}{\rho} \mathbf{x}'_{ij} \boldsymbol{\xi}\right). \quad (15)$$

Reducing the Weibull distribution to the exponential one, i.e., setting $\rho = 1$, we further find:

$$E(Y_{ij}^k) = \frac{\delta_j^k k}{\lambda^k} \Gamma(1 - k) \Gamma(k) \exp\left(-k \mathbf{x}'_{ij} \boldsymbol{\xi}\right). \quad (16)$$

The cases corresponding to (15) and, especially, (16) will allow us to make our point about Lee and Nelder's example. Generally, $\Gamma(\alpha - k/\rho)$ poses a problem when $\alpha - k/\rho$ is a negative integer. For simplicity focusing on a single outcome Y for the case where $\alpha = 1$ and $\beta = 1/\delta$, assembling the linear predictor in μ , and writing $\varphi = \lambda e^\mu$, we find:

$$f(y) = \frac{\varphi \rho y^{\rho-1} \delta}{(\delta + \varphi y^\rho)^2}, \quad (17)$$

$$E(Y^k) = \frac{k}{\rho} \left(\frac{\delta}{\varphi}\right)^{k\rho} \cdot \Gamma(1 - k/\rho) \cdot \Gamma(k/\rho). \quad (18)$$

Note that (17) provides a family of distributions, special cases of the Weibull-gamma model that are termed Weibull-exponential by Molenberghs *et al* (2009). Considering further the exponential case with $\rho = 1$, yields exponential-exponential distributions, with:

$$f(y) = \frac{\varphi \delta}{(\delta + \varphi y)^2}, \quad (19)$$

$$E(Y^k) = k \left(\frac{\delta}{\varphi}\right)^k \cdot \Gamma(1 - k) \cdot \Gamma(k). \quad (20)$$

Clearly, (19) defines a family of distributions without finite moments, similar to the Cauchy distribution, because $\Gamma(1 - k)$ is undefined for $k = 1, 2, \dots$. When $\rho \neq 1$ but fractional, some but not all moments exist, whereas for irrational values of ρ , all moments in (18) are properly defined. Finally, observe that in the general case, there are combinations possible for (α, ρ, k) that would lead to negative integers and hence undefined moments (15).

In the light of the above developments, we are concerned that Lee and Nelder provide us with point estimates for moments that are undefined.

4 The Nature of Generalized Estimating Equations

Lee and Nelder touch upon the use of generalized estimating equations (GEE), as opposed to fully specified probability models. We agree that a comparison between GEE to generalized linear mixed models (GLMM) or hierarchical generalized linear models (HGLM) is like a comparison of apples to oranges, because GEE is an estimating method, which can be applied to random-effects models too (Zeger, Liang, and Albert 1988), and to HGLM for that matter. Therefore, again in agreement with Lee and Nelder, we would like to reiterate that a proper basis of comparison is between marginal and random-effects models. Thus, while part of the literature is sloppy in making comparisons using sloppy categorizations, it should not deflect from the real issues. Nevertheless, we would like to reflect on whether GEE are indeed less appealing because of their alleged lack of a probabilistic basis.

There are two important points in our view. First, a fully specified probability model is not always essential when making inferences about a particular aspect of the model, such as mean functions, as long as the appropriate regularity conditions are satisfied. For example, when estimating mean parameters, it is imperative that the mean exist and be finite.

Second, as Molenberghs and Kenward (2008) pointed out, the lower-order moments that need to be formulated when setting up GEE, correspond to at least one fully specified probabilistic model, even though it may not be of the simple, elegant type, one has in mind *a priori*, such as the Bahadur (1961) or odds-ratio model (Molenberghs and Lesaffre 1994). Their work addresses the concern regarding whether the model portions used in GEE can always be viewed as a partially specified version of a model with full distributional assumptions, or rather whether such a *parent* simply does not exist. To this end, they use the hybrid models (in the sense of being partially marginally and partially conditionally specified) of Fitzmaurice and Laird (1993) and Molenberghs and Ritter (1996). The results by Molenberghs and Kenward (2008) are broadly valid. First, they are valid for a wide class of semi-parametric models where specification is done in terms of (parts of) the exponential-family formulation, including binary, nominal, ordinal, and Poisson outcomes. Second, it is also valid when the outcome vector combines outcomes of different types. Third, using transformations, the result can be applied as well when the semi-parametric specification is not directly in terms of the exponential family, such as logistic regressions for binary data coupled with pairwise correlation, as in classical generalized estimating equations.

5 Computational Approaches

We are convinced that h -likelihood is a tremendously appealing and important addition to the literature. Other computational principles and techniques for maximizing a likelihood with unobservables exist as well. Each one of them has its advantages and drawbacks. For example, Taylor-series-based methods, such as PQL and MQL, and Laplace approximations often lead to substantial bias. This is important to realize because they have been in common use nevertheless, not in the least because they are implemented in standard statistical software, such as the SAS procedure GLIMMIX. The numerical-integration-based methodology, implemented for example in the SAS procedure NLMIXED, is frequently slow and/or extremely sensitive to starting values. See also Molenberghs and Verbeke (2005). Also Bayesian methods, sometimes believed to be free of the issues arising in a likelihood or frequentist context, also have their problems. For one, the sensitivity arising from unobservables, as discussed in Section 2, is equally present in this framework. It is clear to us that none of the computational approaches will be able to claim uniform superiority over all others.

Molenberghs *et al* (2009) provide an overview of computational methods, including some less familiar ones. Their context is a hierarchical model with both normally distributed and conjugate or other random effects. Each of them deals in its own way with the lack of closed-form expression for the marginal likelihood, even though Molenberghs *et al* (2009) derive such closed forms for more settings, such as general Poisson, probit, and Weibull models with random effects.

One approach, very specific to the setting of Molenberghs *et al* (2009), is to integrate analytically over conjugate random effects and then further numerically over the normally distributed random effects.

For the specific case of the marginalized probit model, the computational challenge stems from the presence of a high-dimensional multivariate normal integral in the marginal distribution. Zeger, Liang,

and Albert (1988) derived the marginal mean function, needed for their application of generalized estimating equations as a fitting algorithm for the marginalized probit model. It is one of the first instances of the use of GEE to a non-marginally specified model.

In the same spirit, pseudo-likelihood can be used (Aerts *et al* 2002, Molenberghs and Verbeke 2005). This is particularly useful when the joint marginal distribution is available but cumbersome to manipulate and evaluate, such as in the probit case. This is the idea followed by Renard, Molenberghs, and Geys (2004) for a multilevel probit model with random effects.

Schall (1991) proposed an efficient and general estimation algorithm, based on Harville's (1974) modification of Henderson's (1984) mixed-model equations. Hedeker and Gibbons (1994) and Gibbons and Hedeker (1997) proposed numerical-integration based methods, thus considering neither marginal moments (means, variances) nor marginalized joint probabilities. Guilkey and Murphy (1993) provide a useful early overview of estimation methods and then revert to Butler and Moffit's (1982) Hermite-integration based method, supplemented with Monte Carlo Markov Chain ideas. Also the EM algorithm can be used, in line with Booth *et al* (2003) for the Poisson case. The EM is a flexible framework within which random effects can be considered the 'missing' data over which expectations are taken. Booth *et al* (2003) also considered non-parametric maximum likelihood, in the spirit of Aitken (1999) and Alfò and Aitkin (2000).

A suite of methods is available that employ transformation results, essentially based on transforming the non-normal random effects to normal ones, or vice versa. Liu and Yu (2008) propose a simple transformation of a non-normal random effect to a normal one, at density level, upon which the SAS procedure NLMIXED or similar software can be used. Nelson *et al* (2006) advocate the transformation: $u_i = F_u^{-1}[\Phi(a_i)]$, where F_u is the cumulative distribution function (CDF) of u_i and $\Phi(\cdot)$ is the standard normal CDF, as before. Nelson *et al*'s method, labeled *probability integral transformation* (P.I.T.) comes down to generating normal variates and then inserting these in the model only after transformation, ensuring that they are of the desired nature. Lin and Lee (2008) present estimation methods for the specific case of linear mixed models with skew-normal, rather than normal, random effects.

Quite apart from the choice of estimation method, it is important to realize that not all parameters may be simultaneously identifiable. For example, the gamma-distribution parameters in the Poisson case, α and β , such as in (13), are not simultaneously identifiable when the linear-predictor part is also present, because there is aliasing with the intercept term. Therefore, one can set, for example, β equal to a constant, removing the identifiability problem. It is then clear that α , in the univariate case, or the set of α_j in the repeated-measures case, describe the additional overdispersion, in addition to what stems from the normal random effect(s). A similar phenomenon also plays in the binary case, where both beta-distribution parameters are not simultaneously estimable.

6 Concluding Remarks

We end by sincerely thanking Professors Lee and Nelder for their important, thought-provoking, and practically relevant work in this rapidly evolving area. Their paper has allowed us to elaborate on a number of statistical issues put forward in their paper, such as the implications of formulating models with unobservables, and generating distributions with peculiar moment-properties. It further gave us the opportunity to reflect on some conceptual aspects of generalized estimating equations on the one hand, and elaborate on computational strategies for models of the type discussed here on the

other.

References

- Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002) *Topics in Modelling of Clustered Data*. London: Chapman & Hall.
- Aitkin, M. (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117-128.
- Alfö, M. and Aitkin, M. (2000) Random coefficient models for binary longitudinal responses with attrition. *Statistics and Computing*, **10**, 279-288.
- Bahadur, R.R. (1961) A representation of the joint distribution of responses to n dichotomous items. In: *Studies in Item Analysis and Prediction*, H. Solomon (Ed.). Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.
- Booth, J.G., Casella, G., Friedl, H., and Hobert, J.P. (2003) Negative binomial loglinear mixed models. *Statistical Modelling*, **3**, 179-181.
- Butler, J.S. and Moffit (1982) A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, **50**, 761-765.
- Duchateau, L. and Janssen, P. (2007) *The Frailty Model*. New York: Springer.
- Fitzmaurice, G.M. and Laird, N.M. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141-151.
- Gibbons, R.D. and Hedeker, D. (1997) Random effects probit and logistic regression models for three-level data. *Biometrics*, **53**, 1527-1537.
- Guilkey, D.K. and Murphy, J.L. (1993) Estimation and testing in the random effects probit model. *Journal of Econometrics*, **59**, 301-317.
- Harville, D.A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383-385.
- Hedeker, D. and Gibbons, R.D. (1994) A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **51**, 933-944.
- Henderson, C.R. (1984) *Applications of Linear Models in Animal Breeding*. Guelph, Canada: University of Guelph Press.
- Heitjan, D.F. and Rubin, D.B. (1991) Ignorability and coarse data. *The Annals of Statistics*, **19**,
- Lin, T.I. and Lee, J.C. (2008) Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in Medicine*, **27**, 1490-1507.
- Liu, L. and Yu, Z. (2008) A likelihood reformulation method in non-normal random-effects models. *Statistics in Medicine*, **27**, 3105-3124. 2244-2253.

- Molenberghs, G. and Kenward, M.G. (2008) Generalized estimating equations and their corresponding full models. *Submitted for publication*.
- Lee, Y. and Nelder, J.A. (1996) Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B* **58**, 619–678.
- Lee, Y. and Nelder, J.A. (2001) Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.
- Lee, Y., and Nelder, J.A. (2003) Extended-REML estimators. *Journal of Applied Statistics*, **30**, 845–856.
- Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008) Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B* **70**, 371–388
- Molenberghs, G. and Verbeke, G (2005) *Models for Discrete Longitudinal Data*. New York: Springer.
- Molenberghs, G. and Lesaffre, E. (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* **89**, 633–644.
- Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2009) A family of generalized linear models for repeated measures with normal and conjugate random effects. *Submitted for publication*.
- Molenberghs, G. and Kenward, M.G. (2007) *Missing Data in Clinical Studies*. New York: John Wiley.
- Molenberghs, G. and Ritter, L. (1996) Likelihood and quasi-likelihood based methods for analysing multivariate categorical data, with the association between outcomes of interest. *Biometrics*, **52**, 1121–1133.
- Molenberghs, G. and Verbeke, G. (2009) A note on a hierarchical interpretation for negative variance components. *Submitted for publication*.
- Nelder, J.A. (1954) The interpretation of negative components of variance. *Biometrika* **41**, 544–548.
- Renard, D., Molenberghs, G., and Geys, H. (2004) A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, **44**, 649–667.
- Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–729.
- Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.
- Verbeke, G. and Molenberghs, G. (2009) Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling*, **00**, 000–000.

- Zhang, J. and Heitjan, D.F. (2007) Impact of nonignorable coarsening on Bayesian inference. *Biostatistics*, **8**, 722–743.
- Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.