*Statistical Science*
2000, Vol. 15, No. 1, 1–26

# Marginalized Multilevel Models and Likelihood Inference

**Patrick J. Heagerty and Scott L. Zeger**

*Abstract.* Hierarchical or "multilevel" regression models typically parameterize the mean response conditional on unobserved latent variables or "random" effects and then make simple assumptions regarding their distribution. The interpretation of a regression parameter in such a model is the change in possibly transformed mean response per unit change in a particular predictor having controlled for all conditioning variables including the random effects. An often overlooked limitation of the conditional formulation for nonlinear models is that the interpretation of regression coefficients and their estimates can be highly sensitive to difficult-to-verify assumptions about the distribution of random effects, particularly the dependence of the latent variable distribution on covariates. In this article, we present an alternative parameterization for the multilevel model in which the marginal mean, rather than the conditional mean given random effects, is regressed on covariates. The impact of random effects model violations on the marginal and more traditional conditional parameters is compared through calculation of asymptotic relative biases. A simple two-level example from a study of teratogenicity is presented where the binomial overdispersion depends on the binary treatment assignment and greatly influences likelihood-based estimates of the treatment effect in the conditional model. A second example considers a three-level structure where attitudes toward abortion over time are correlated with person and district level covariates. We observe that regression parameters in conditionally specified models are more sensitive to random effects assumptions than their counterparts in the marginal formulation.

*Key words and phrases:* Generalized linear model, latent variable, logistic regression, random effects model.

## 1. INTRODUCTION

Multilevel modeling (Goldstein, 1995a) refers to a class of multivariate statistical techniques developed for the analysis of data collected in dependent groups or "clusters." Such data arise naturally in many scientific disciplines. For example, in a longitudinal study a cluster might consist of repeated measurements over time on an individual. In teratologic applications, where birth defects in laboratory animals exposed to a pharmaceutical substance are studied, clusters comprise offspring litters. Such studies may yield a multivariate response for each offspring, indicating the presence or absence of various types of malformation. In sociologic applications, clusters can be classrooms, sampling districts or communities. The distinguishing feature of clustered data is that observations within a cluster are usually more similar to one another than are observations from different clusters. When this variation cannot be explained by measured covariates, we require statistical methods for the analysis of correlated measurements.

In many multilevel problems the scientific focus is on the relationship between the vector of responses for cluster $i$, $\mathbf{Y}_i = \text{vec}(Y_{ij})$, and corresponding covariates $\mathbf{X}_i$. Regression modeling of $E(Y_{ij}|\mathbf{X}_i)$ provides a flexible method for characterizing systematic variation and for testing relationships be-

*Patrick J. Heagerty is Assistant Professor, Department of Biostatistics, University of Washington, Seattle, Washington 98195. Scott L. Zeger is Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205.*

tween the response and predictors. However, proper assessment of the statistical evidence, including calculation of standard errors for regression parameters, requires that the clustering be properly accounted for in the model and/or the estimation method used. Several statistical models have been developed for regression with clustered data including hierarchical models using likelihood inference, for example, generalized linear mixed models (GLMM) (Zeger and Karim, 1991; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Goldstein, 1995a) and marginal models fitted by generalize estimating equations (GEE) (Liang and Zeger, 1986).

A "marginal" model is one in which $E(Y_{ij} \mid \mathbf{X}_i)$ is directly modeled. Marginal models for multivariate continuous and categorical data were proposed by Plackett (1965) and Dale (1986). Liang and Zeger (1986) developed marginal regression methods without requiring assumption about the complete joint distribution of the response vectors. Such models have proven useful for categorical response data since few joint probability models for multivariate categorical data permit tractable modeling of the marginal means $E(Y_{ij} \mid \mathbf{X}_i)$. For example, log-linear models (Bishop, Feinberg and Holland, 1975) provide a flexible, valid multivariate model; but the canonical parameterization is in terms of the expectation of one response $Y_{ij}$ conditional on the other responses $Y_{ik}$, $k \neq j$. However, the estimation methods used by Liang and Zeger (1986) are not likelihood based, instead relying on estimating functions and empirical variances for point and interval estimation. This approach affords robustness to misspecification of the multivariate dependence structure, yet it sacrifices use of likelihood-based procedures such as profile likelihood functions or likelihood ratio tests and does not yield estimates of multivariate probabilities. In many situations, likelihood-based methods may be preferred, motivating the development of alternative models.

Fitzmaurice and Laird (1993) and Azzalini (1994) showed how classical multivariate methods for categorical data could be modified to permit regression modeling of $E(Y_{ij} \mid \mathbf{X}_i)$. Fitzmaurice and Laird (1993) reparameterized the likelihood of a canonical log-linear model for balanced binary data in terms of the marginal means and the higher-order, canonical association parameters. Similarly, Azzalini (1994) reparameterized Markov models to allow regression modeling of the induced marginal means. Each of these approaches starts with an underlying probability model with parameters that describe $E(Y_{ij} \mid Y_{ik}\colon k \neq j)$. In the log-linear model, each response is conditioned on all other responses; in the Markov model, each response is conditioned on all previous responses in time. Thus, Fitzmaurice and Laird (1993) and Azzalini (1994) have both "marginalized" models (i.e., reparameterized a model in terms of the marginal mean and additional dependence parameters) for conditional means given other responses, or "response conditional models." In Section 2 we describe these approaches in more detail.

Generalized linear mixed models (Stiratelli, Laird and Ware, 1984; Zeger and Karim, 1991; Breslow and Clayton, 1993) comprise another class of models for dependent data. They use latent variables or random effects $\mathbf{b}_i$ to introduce correlation among observation within a cluster. In a GLMM, we typically construct a mean model for the response variables conditional on both measured covariates and unobserved latent variables, $E(\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{b}_i)$. The consequence of this choice is subtle but can critically impact both parameter interpretation (Zeger, Liang, and Albert, 1988; Grabard and Korn, 1994) and the robustness of estimates to the specification of the distribution of $\mathbf{b}_i$ (Neuhaus, Hauck and Kalbfleisch, 1992). In Section 2 we consider the interpretation of marginal and conditional parameters, and in Section 4 we evaluate bias that may arise in regression estimates when the distribution of $\mathbf{b}_i$ is misspecified.

Several authors have discussed appropriate domains of application of marginal models fitted by GEE and generalized linear mixed models for the analysis of dependent data (e.g., Zeger, Liang and Albert, 1988; Graubard and Korn, 1994; Pendergast et al., 1996). The goal of this manuscript is to demonstrate that the choice of mean parameterization can, and often should, be separated from the choice of a multivariate probability model and that likelihood methods can be used for either marginal mean models or more traditional conditional mean models. Specifically, we build on results of Heagerty (1999) showing that an underlying latent variable model can be used for either marginal mean models or conditional mean models. By so doing, a comparison of the merits of models for marginal versus conditional means can be made without confounding the choice by the parameterizations of within-cluster association or by different estimation methods. The marginal and conditional methods that we explore are directly comparable because both methods assume an underlying latent variable structure and both can be estimated by maximum likelihood. Our marginal mean model can be considered a "marginalization" of latent variable models in a spirit similar to the marginalization of response conditional models by Fitzmaurice and Laird (1993) and Azzalini (1994).

Finally, we illustrate some of the potential advantages of adopting a full probability model for use with a marginal mean structure. Section 5 considers two multilevel examples with binary response variables. The first example is a now classic data set from teratology and has a single level of clustering. The second example is from sociology and has three nested levels: observation within individuals within sampling districts. We briefly introduce these examples:

EXAMPLE 1. *Teratology data*. Weil (1970) presents two-level data on whether individual rat pups survive the first 21 days of life after their mother (level 2) was exposed to a given dose of teratogen. Let $Y_{ij}$, $j = 1, 2, \ldots, N_i$, represent the indicator of 21-day survival for individual pups born to animal $i$. The scientific question concerns the impact on birth outcomes for pups with maternal exposure to a teratogen ($X_i = 1$) relative to pups with unexposed mothers ($X_i = 0$). Estimates of the marginal means $E(Y_{ij} \mid X_i)$ can be used to compare the rates of survival between pups with exposed and unexposed mothers. These data have previously been analyzed by Liang and Hanfelt (1994) to demonstrate the sensitivity of beta-binomial likelihood inference to the assumed correlation model. This lack of robustness motivated the investigation presented in Section 4 where the sensitivity of our proposed likelihood-based inference to the proper specification of the latent variable distribution is similarly evaluated.

EXAMPLE 2. *British Social Survey data*. A second example considers a three-level logistic model where repeated binary measurements (level 1) are obtained on individuals (level 2) who are clustered into sampling districts (level 3). In this sociology example the binary response $Y_{ijk}$ represents the view at time $k$ of participant $j$ in district $i$ toward governmental regulation of abortion (McGrath and Waterton, 1986). Scientific interest is in the correlation between attitudes and measured demographic covariates such as gender and religion. The only level-1 covariate is the year of the measurement, whereas all demographic characteristics are level-2 covariates. A single level-3 covariate is derived as the district mean for a level-2 covariate, illustrating the potential importance of separating between-cluster covariate differences from within-cluster differences (Neuhaus and Kalbfleisch, 1998). A marginal regression model provides a description of the systematic variation across different subsets of the population in the proportion favoring no regulation of abortion. Marginalized latent variable models provide an additional summary of the magnitude of random "within-group" heterogeneity and permit flexible likelihood-based inference regarding the marginal mean structure.

## 2. MODELS

In this section we present an overview of several approaches to generalized linear modeling of multivariate discrete data. Traditional approaches can be classified as "response conditional models" (log-linear and Markov models), latent variable models and direct marginal models. We then discuss how marginalization of conditionally specified multivariate models can be used to permit likelihood-based, marginal regression analysis.

We use notation for three-level data where $\mathbf{Y}_i = \text{vec}(Y_{ijk})$ denotes a vector of response variables for cluster $i$. Notation for examples with fewer or greater numbers of levels is obvious. Let $\mathbf{X}_i$ denote covariates associated with cluster $i$ and let $\mathbf{X}_{ijk}$ denote covariates for observation $k$ within subcluster $j$. We decompose covariates into level-1, level-2 and level-3 covariates, $\mathbf{X}_{ijk} = (\mathbf{X}_{1,ijk}, \mathbf{X}_{2,ij}, \mathbf{X}_{3,i})$.

We are interested in the specification, interpretation and estimation of a parameter $\boldsymbol{\theta}$ which describes the joint distribution of $\mathbf{Y}_i$ given covariates $\mathbf{X}_i$. For certain models it will be natural to partition this parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ into the mean parameter $\boldsymbol{\beta}$, which specifies the first moment, and the association parameter or variance components $\boldsymbol{\alpha}$.

### 2.1 Direct Marginal Specification

The "marginal modeling approach" builds separate regressions for first, second and higher moments of the joint distribution $[\mathbf{Y}_i \mid \mathbf{X}_i]$. The *marginal* expectation of the response can be linked to covariates using a generalized linear model,

$$E(Y_{ijk} \mid \mathbf{X}_i) = \mu_{ijk},$$
$$g(\mu_{ijk}) = \eta(\mathbf{X}_{ijk}) = \mathbf{X}_{ijk}\boldsymbol{\beta}^M.$$

Additional models are then specified for the second moment $E(Y_{ijk}, Y_{ij'k'} \mid \mathbf{X}_i)$, and possibly for the higher moments. For example, Dale (1986) parameterized the joint distribution of two binary variables in terms of their marginal means and their pairwise odds ratio $\Psi(Y_{ij}, Y_{ik})$, defined as

$$\Psi(Y_{ij}, Y_{ik}) = \frac{\left[P(Y_{ij} = 1, Y_{ik} = 1)P(Y_{ij} = 0, Y_{ik} = 0)\right]}{\left[P(Y_{ij} = 1, Y_{ik} = 0)P(Y_{ij} = 0, Y_{ik} = 1)\right]}.$$

Molenberghs and Lesaffre (1994), Glonek and McCullagh (1995) and Heagerty and Zeger (1996) extend direct marginal models to vectors of binary or ordinal responses specifying separate regression

models for the marginal means, the pairwise odds ratios and the higher-order contrasts among log-odds ratio to complete the likelihood function.

In these marginal models, the mean (or first moment) regression parameters represent the change in expected response, such as prevalence with binary outcomes, per unit change in a given predictor without conditioning on the other responses or any latent variables. Correlation among elements of $\mathbf{Y}_i$ given $\mathbf{X}_i$, even if reasonably attributed to shared unobservable latent variables, are accounted for by a separate correlation model or pairwise association regression. For example, with binary $\mathbf{Y}_i$, the pairwise log-odds ratio $\log \Psi(Y_{ij}, Y_{ik})$ can be simultaneously regressed on predictors (Lipsitz, Laird and Harrington, 1991; Carey, Zeger and Diggle, 1993; Heagerty and Zeger, 1996).

There are several advantages of a direct marginal approach. First, the interpretation of regression coefficients in either the mean or the odds ratio model does not depend on the dimension of $\mathbf{Y}_i$ as it does in certain response conditional models (see Sections 2.2.1 and 2.4.1 for illustration). Hence clusters of different size can be easily accommodated by marginal models as they can in latent variable models. Second, the interpretation of mean parameters $\boldsymbol{\beta}^M$ is invariant with respect to specification of the association, or of higher-order models. Two data analysts with the same mean regression but different association models have exactly the same target of estimation, $\boldsymbol{\beta}^M$. In this sense, the mean model is "separable" from the remainder of the model for the joint distribution. In Section 2.3 we illustrate that the property of mean separability does not hold for response or latent variable conditional models.

Marginal models describe the dependence of the means, and of the association among responses within a cluster, on measured predictors. In some applications investigators seek to determine whether the observed associations are caused by hypothesized dependence of one response on others or by unobserved latent variables. Marginal models do not address such questions directly.

Because marginal models separately parameterize the mean and higher-order moments, it is possible to estimate mean parameters without specifying the complete joint distribution of $\mathbf{Y}_i$. Liang and Zeger (1986) introduced one approach, GEE, which is an application of optimal estimating functions (Godambe, 1960) to the regression problem. Use of GEE is a natural extension of quasilikelihood (Wedderburn, 1974; McCullagh and Nelder, 1989) to the multivariate response setting where one must contend with additional nuisance parameters. An alternative formulation of this semiparametric approach to marginal regression models was proposed by Gourieroux, Monfort and Trognon (1984). Even though full likelihood specification is not necessary to estimate mean regression parameters in a marginal model, it is always possible and often desirable.

### 2.2 Response Conditional Models

There are two main classes of models for multivariate data that can naturally be viewed as models for the expected value of one response conditional on subsets of the other responses from the same cluster. They are effective for modeling associations but do not admit simple models for the marginal means.

2.2.1 *Log-linear models*. Log-linear models (Bishop, Feinberg and Holland, 1975) have been widely used for the analysis of cross-classified discrete observations. Balanced binary vectors $\mathbf{Y}_i$ $(Y_{i1}, Y_{i2}, \ldots, Y_{in})$ for $i = 1, 2, \ldots, N$ can be considered as a cross-classification of the $n$ component responses. A log-linear model is constructed directly for the multivariate probabilities,

$$\begin{aligned}\log P_{\boldsymbol{\theta}_i}(Y_{i1}, \ldots, Y_{in}) &= \theta_i^{(0)} + \sum_j \theta_{ij}^{(1)} Y_{ij} + \sum_{j<k} \theta_{ijk}^{(2)} Y_{ij} Y_{ik} \\ &\quad + \sum_{j<k<l} \theta_{ijkl}^{(3)} Y_{ij} Y_{ik} Y_{il} + \cdots + \theta_i^{(n)} Y_{i1} \cdots Y_{in}.\end{aligned}$$

Here the canonical parameter vector $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i^{(1)}, \boldsymbol{\theta}_i^{(2)}, \ldots, \theta_i^{(n)})$ is unconstrained and $\theta_i^{(0)}$ is a normalizing constant. Given covariates $\mathbf{X}_i$ it is possible to allow $\boldsymbol{\theta}_i$ to depend on $\mathbf{X}_i$ or to extend the log-linear model to describe $\log P(\mathbf{Y}_i, \mathbf{X}_i)$ when $\mathbf{X}_i$ is also discrete. However, in either case the log-linear model results in complicated functions for the marginal expectations $E(Y_{ij}|\mathbf{X}_i)$ because these are obtained as sums over the response variable joint distribution,

$$E\left(Y_{ij}|\mathbf{X}_i\right) = \sum_{Y_{ik},\, k \neq j} P_{\boldsymbol{\theta}_i}\left(Y_{i1}, Y_{i2}, \ldots, \underline{Y_{ij} = 1}, \ldots, Y_{in} \,|\mathbf{X}_i\right),$$

yielding mixtures of exponential functions of the canonical parameters $\boldsymbol{\theta}_i$. In a log-linear model, the natural (canonical) univariate regressions are for the conditional expectations,

$$\begin{aligned}\operatorname{logit} E\left(Y_{ij} \mid Y_{ik} : k \neq j\right) &= \theta_{ij}^{(1)} + \sum_k \theta_{ijk}^{(2)} Y_{ik} \\ &\quad + \sum_{k<l} \theta_{ijkl}^{(3)} Y_{ik} Y_{il} + \cdots + \theta_i^{(n)} \prod_{l \neq j} Y_{il}.\end{aligned}$$

Therefore, although log-linear models are well suited for describing multivariate dependencies or for modeling joint and conditional distributions, they do not directly facilitate multivariate generalized linear regression modeling of the marginal means.

2.2.2 *Transition models*. When a cluster of response variables $(Y_{i1}, Y_{i2}, \ldots, Y_{in})$ is naturally ordered, for example in time, it is possible to construct a multivariate model by decomposing the joint distribution into a sequence of predictive distributions,

$$P_{\boldsymbol{\theta}_i}(Y_{i1}, Y_{i2}, \ldots, Y_{in}) = P_{\boldsymbol{\theta}_i}(Y_{i1}) \prod_{j=2}^{n} P_{\boldsymbol{\theta}_i}\big(Y_{ij} \mid Y_{ik}\colon k < j\big).$$

These models have been referred to as transition models (e.g., Diggle, Liang and Zeger, 1994) or discrete Markov models (see MacDonald and Zucchini, 1997, for a recent survey) and are useful for modeling the expected value of a response conditional on both covariates and the history of the series. Again, it is straightforward to allow the parameters $\boldsymbol{\theta}_i$ to depend on covariates $\mathbf{X}_i$ but difficult to obtain simple expressions for $E(Y_{ij}|\mathbf{X}_i)$ since sums over the joint distribution of times $1, 2, \ldots, j$ are required.

### 2.3 Latent Variable Models

Another way to model the joint distribution $[\mathbf{Y}_i \,|\mathbf{X}_i]$ is to postulate the existence of unobserved latent variables which are shared by, and hence introduce correlation among, the elements of $\mathbf{Y}_i$. The observed data likelihood is constructed by integrating over the latent variable distribution,

$$P_{\boldsymbol{\theta}}[\mathbf{Y}_i \,|\mathbf{X}_i] = \int P_{\boldsymbol{\theta}}[\mathbf{Y}_i \,|\mathbf{X}_i, \mathbf{b}_i] f_{\boldsymbol{\theta}}(\mathbf{b}_i \,|\mathbf{X}_i)\, d\mathbf{b}_i.$$

There are two common assumptions to simplify this model: conditional independence among responses,

$$(1) \qquad P_{\boldsymbol{\theta}}[\mathbf{Y}_i \,|\mathbf{X}_i, \mathbf{b}_i] = \prod_{j,k} P_{\boldsymbol{\theta}}\big[Y_{ijk}|\mathbf{X}_i, \mathbf{b}_i\big];$$

and homogeneous latent variable distribution,

$$(2) \qquad f_{\boldsymbol{\theta}}(\mathbf{b}_i|\mathbf{X}_i) = f_{\boldsymbol{\theta}}(\mathbf{b}_i).$$

In this article we adopt assumption (1) because it forms our basis for structuring the correlation among responses within clusters. Assumption (2) is a strong one and we consider a more general regression structure for the random effects variance components.

The most common partition of $\boldsymbol{\theta}$ is into $\boldsymbol{\beta}^C$ and $\boldsymbol{\alpha}$, where

$$P_{\boldsymbol{\theta}}[\mathbf{Y}_i|\mathbf{X}_i] = \int \left\{ \prod_{j,k} P_{\boldsymbol{\beta}^C}\big[\mathbf{Y}_{ijk}|\mathbf{X}_i, \mathbf{b}_i\big] \right\} f_{\boldsymbol{\alpha}}(\mathbf{b}_i \,|\mathbf{X}_i)\, d\mathbf{b}_i.$$

Here $\boldsymbol{\beta}^C$ are canonical regression parameters in a GLM for the *conditional* expectation of the response:

$$E\big(Y_{ijk} \,|\mathbf{X}_i, \mathbf{b}_i\big) = \mu^b_{ijk},$$
$$g\big(\mu^b_{ijk}\big) = \Delta(\mathbf{X}_{ijk}) + b_{ijk} = \mathbf{X}_{ijk}\boldsymbol{\beta}^C + b_{ijk}.$$

Assumptions about $b_{ijk}$ commonly used in practice include the following: mixed models where $b_{ijk} = \mathbf{Z}_{ijk}\mathbf{u}_{ij}$ for $\mathbf{Z}_{ijk}$ a subset of $\mathbf{X}_{ijk}$ and $\mathbf{u}_{ij}$ is a $q \times 1$ vector of random effects; nested clusters where $b_{ijk} = b_{ij} + b_i$; and serial or spatial models where $b_{ijk}$ represents an autocorrelated stochastic process (Diggle, 1988). In our notation, the parameter $\boldsymbol{\alpha}$ identifies the specific distribution of $\mathbf{b}_i$ from within its parametric family.

Multilevel models are popular in the empirical sciences for several reasons. First, it is often reasonable to posit that shared, unobserved variables influence the response, thereby making observations within clusters correlated with one another. Simple latent variable assumptions can lead to relatively complex within-cluster associations. Second, the multilevel regression parameter $\beta^C_j$ has a desirable causal interpretation (Holland, 1986) as the change in (possibly transformed) expected response per unit change in $X_j$, holding the other observed variables and unobserved latent factors fixed. Third, these models make possible the estimation of cluster-specific regression coefficients, for example intercepts, with estimates that use information from subjects within a particular cluster but which also borrow information from other clusters. These shrinkage or empirical Bayes estimates (Efron and Morris, 1973) are often superior to competitors which rely only on a cluster's own data.

In the conditional mean parameterization, the regression contrasts $\boldsymbol{\beta}^C$ measure the change in transformed mean per unit change in a covariate, controlling for all other variables including the latent variables $b_{ijk}$. Because the latent variable assumptions determine what values of $b_{ijk}$ are equivalent, these assumptions also determine the interpretation of the parameter $\boldsymbol{\beta}^C$. For example, consider a simple "pre–post" design in which daily binary measurements are taken on each person for one week during which they are on placebo and for a second week during which the same subject receives active therapy. For this scenario we could use the following model: $\text{logit}(\mu^b_{ij}) = \beta^C_0 + \beta^C_1 X_{ij} + b_{ij}$, where $X_{ij} = 0$ for $j = -7, -6, \ldots, -1$ and $X_{ij}$

$= 1$ for $j = 1, 2, \ldots, 7$. If we further assume Gaussian random intercepts, $b_{ij} = b_{i0} \sim N(0, \sigma^2)$, then we obtain $g(\mu_{ik}^b) - g(\mu_{ij}^b) = (\beta_0^C + \beta_1^C X_{ik} + b_{i0}) - (\beta_0^C + \beta_1^C X_{ij} + b_{i0}) = \beta_1^C$ for $k > 0$ (post) and $j < 0$ (pre) representing the change in an individual's log odds comparing a day on treatment to a day off treatment. Such interpretations lead Zeger, Liang and Albert (1988) to refer to $\beta_1^C$ as the "subject-specific" effect of treatment. However, a subject-specific interpretation relies on the assumption that $b_{ij}$ is constant over time. A random intercepts assumption can be viewed as a special case of a more general serially dependent stochastic process model where $\text{cov}(b_{ij}, b_{ik}) = \sigma^2 \rho^{|j-k|}$ and $\rho = 1$. If we relax the random intercepts model to allow dependence to decay as the time separation increases, $\rho < 1$, then the parameter $\beta_1^C$ no longer measures the change in a subject's log odds comparing a day on treatment to a day on placebo, because controlling for the individual no longer ensures that the latent variables $b_{ij}$ and $b_{ik}$ are equal. Thus, a simple change in the latent variable assumptions now makes $\beta_1^C$ both subject and time specific.

### 2.4 Marginalized Response Conditional Models

In Section 2.2.1 and 2.2.2 we discussed response conditional models and commented that these formulations are attractive for describing dependence among the elements of $\mathbf{Y}_i$. In this section we briefly review approaches that have modified the natural parameterization of the response conditional models to allow likelihood-based estimation of marginal mean regression parameters.

2.4.1 *Marginalized canonical models*. Canonical log-linear models provide an unconstrained method for modeling multivariate dependencies yet do not directly allow regression models for the marginal means. Fitzmaurice and Laird (1993) marginalized these models to permit likelihood-based regression estimation of the marginal means by transforming the canonical parameter $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i^{(1)}, \boldsymbol{\theta}_i^{(2)}, \ldots, \theta_i^{(n)})$ into the mixed parameter $\boldsymbol{\theta}_i^* = (\boldsymbol{\mu}_i, \boldsymbol{\theta}_i^{(2)}, \ldots, \theta_i^{(n)})$, where $\boldsymbol{\mu}_i = \text{vec}(\mu_{ij})$, $\mu_{ij} = E(Y_{ij} \mid \mathbf{X}_i)$. In their approach the underlying log-linear model parameters $(\boldsymbol{\theta}_i^{(2)}, \ldots, \theta_i^{(n)})$ are used to describe the covariance of the response vector while the expectations of the response variables are directly modeled via the marginal means. Fitzmaurice and Laird (1993) showed how iterative proportional fitting (Deming and Stephan, 1940) can be used to transform from the mixed parameter $\boldsymbol{\theta}_i^*$ to the canonical parameter $\boldsymbol{\theta}_i$ to evaluate the likelihood function.

A related use of log-linear models that also permits marginal regression models is presented by Lang and Agresti (1994) and Glonek and McCullagh (1995). Each of these approaches is limited to applications with small or moderate cluster sizes due to computational demands. In addition, the methods of Fitzmaurice and Laird (1993) are effectively limited to balanced data since the canonical association parameters $(\boldsymbol{\theta}_i^{(2)}, \ldots, \theta_i^{(n)})$ must be separately modeled and estimated for each cluster size $n$.

2.4.2 *Marginalized transition models*. Azzalini (1994) showed how to marginalize first-order Markov chains for discrete responses. Let $(Y_1, Y_2, \ldots, Y_T)$ denote serial binary observations, let $\mathbf{Z}_T = \{\mathbf{X}_t\}_{t=1}^T$ denote the collection of covariates for all times $t = 1, 2, \ldots, T$ and assume that $\mathbf{Z}_T$ is ancillary. Azzalini (1994) specified the conditional expectations by

$$\text{logit } E\big(Y_t \mid Y_j : j < t, \mathbf{Z}_T\big) = \Delta_t + \alpha Y_{t-1},$$

where $\alpha$ is the log-odds ratio measuring the association between any pair of successive observations. However, rather than directly parameterize $\Delta_t$, Azzalini (1994) modeled the corresponding marginal means using $\text{logit } E(Y_t \mid \mathbf{Z}_T) = \mathbf{X}_t \boldsymbol{\beta}^M$. Given specification of both $\boldsymbol{\beta}^M$ and $\alpha$, joint probabilities $P_{\boldsymbol{\theta}}(Y_t, Y_{t-1} \mid Z_T)$ as well as the likelihood contributions $P_{\boldsymbol{\theta}}(Y_t \mid Y_{t-1}, \mathbf{Z}_t)$ can be obtained analytically. This approach leaves $\Delta_t$ an implicitly defined function of $\boldsymbol{\theta} = (\boldsymbol{\beta}^M, \alpha)$ and covariates $\mathbf{Z}_T$.

### 2.5 Proposed Model: Marginalized Latent Variable Model

In this section we propose a class of marginally specified multivariate generalized linear models, or "marginalized latent variable models." Our approach parallels the marginalization of response conditional models developed by Fitzmaurice and Laird (1993) and by Azzalini (1994), and generalizes the work of Heagerty (1999). We begin with a regression structure for the marginal means $\mu_{ijk}$:

$$g(\mu_{ijk}) = \mathbf{X}_{ijk} \boldsymbol{\beta}^M. \tag{3}$$

The second component of the model describes the dependence among measurements within a cluster by conditioning on a latent variable rather than on other response variables:

$$g\big(\mu_{ijk}^b\big) = \Delta(\mathbf{X}_{ijk}) + b_{ijk}. \tag{4}$$

Finally, we assume that the elements of the response vector $\mathbf{Y}_i$ are conditionally independent

given $\mathbf{b}_i = \text{vec}(b_{ijk})$ and that the distribution of $\mathbf{b}_i$ is completely specified by the parameter $\boldsymbol{\alpha}$.

Our formulation in (3) and (4) is an alternative to the generalized linear mixed model (Breslow and Clayton, 1993) which directly parameterizes the conditional mean function $\Delta(\mathbf{X}_{ijk}) = \mathbf{X}_{ijk}\boldsymbol{\beta}^C$. There is a critical distinction between the marginal parameter $\boldsymbol{\beta}^M$ and the conditional parameter $\boldsymbol{\beta}^C$. The conditional regression coefficient $\boldsymbol{\beta}^C$ contrasts the expected response for different values of the measured covariates $\mathbf{X}_{ijk}$ for equivalent values of the latent variable $b_{ijk}$. The marginal coefficient does not attempt to control for the unobserved $b_{ijk}$. For example, a marginal gender contrast compares the mean among men to the mean among women, while a conditional gender contrast compares the mean among men with $b_{ijk} = b^*$ to the mean among women who also have $b_{ijk} = b^*$. Interpretation of $\hat{\boldsymbol{\beta}}^C$ can be particularly difficult for multilevel models with level-2 and level-3 covariates since no direct matching of $b_{ijk}$ is observed for these contrasts. See Graubard and Korn (1994) for further discussion.

Our marginalized model in (3)–(4) also permits conditional statements via the implicitly defined $\Delta(\mathbf{X}_{ijk})$, recognizing their dependence on model assumptions. The parameter $\Delta(\mathbf{X}_{ijk})$ is a function of both the marginal linear predictor $\eta(\mathbf{X}_{ijk}) = \mathbf{X}_{ijk}\boldsymbol{\beta}^M$ and the random effects distribution $F_{\boldsymbol{\alpha}}(b_{ijk})$ and is defined as the solution to the integral equation that links the marginal and conditional means:

$$\mu_{ijk} = E\left(\mu^b_{ijk}\right), \tag{5}$$

$$h\left(\mathbf{X}_{ijk}\boldsymbol{\beta}^M\right) = \int h\left[\Delta(\mathbf{X}_{ijk}) + b_{ijk}\right] dF_{\boldsymbol{\alpha}}(b_{ijk}), \tag{6}$$

where $h = g^{-1}$. In the common case where $b_{ijk} \sim N[0, \sigma(\mathbf{X}_{ijk})]$, we can rewrite $b_{ijk} = \sigma(\mathbf{X}_{ijk})z$, where $z \sim N(0, 1)$, and the integral equation becomes

$$h\left[\eta(\mathbf{X}_{ijk})\right] = \int h\left[\Delta(\mathbf{X}_{ijk}) + \sigma(\mathbf{X}_{ijk})z\right]\phi(z)\,dz,$$

where $\phi$ is the standard normal density function. Given $\eta(\mathbf{X}_{ijk})$ and $\sigma(\mathbf{X}_{ijk})$ the integral equation can be numerically solved for $\Delta(\mathbf{X}_{ijk})$. See Heagerty (1999) for details of the linkage between $\eta(\mathbf{X}_{ijk})$ and $\Delta(\mathbf{X}_{ijk})$ when $h = \text{logit}^{-1}$.

For certain link function and mixing distribution combinations the transformation between conditional and marginal mean can be obtained analytically. For example, using a probit link function and Gaussian random effects, $b = \sigma(X)z$ for $z \sim N(0, 1)$, yields the relationship

$$\begin{aligned}\Phi[\eta(X)] &= E\{\Phi[\Delta(X) + \sigma(X)z]\} \\ &= \Phi\left[\frac{\Delta(X)}{\sqrt{1 + \sigma^2(X)}}\right],\end{aligned}$$

showing that the marginal linear predictor $\eta(X)$ is a rescaling of the conditional linear predictor $\Delta(X)$. If the variance of the latent variable is independent of $X$, then the marginal and conditional model structures will be the same (i.e., linear, or additive in multiple covariates); however, if $\sigma(X)$ depends on covariates, then the marginal and conditional models will have different functional forms. A key example where heterogeneity or overdispersion is assumed to depend on covariates is in teratologic applications where the intralitter correlation is a function of the dose $X$ (see Aerts and Claeskens, 1997, for a recent example using the beta-binomial model).

By introducing the marginally specified model, we allow a choice as to whether the marginal mean structure or the conditional mean structure is the focus of modeling when using a latent variable formulation. There exists a general correspondence between $\eta(X)$ and $\Delta(X)$ so that the distinction becomes purely one of where simple regression structure is usefully assumed and what summaries will be presented through the estimated regression coefficients. The choice between marginal or conditional regression models can now be determined by the scientific objectives of the analysis rather than by the availability of only conditional multilevel models.

2.5.1 *Marginalized model parameter interpretation*. For any marginal regression model, the parameter $\boldsymbol{\beta}^M$ contrasts the means for subgroups defined by measured covariates. For example, if we consider a two-level logistic model with a single cluster level binary covariate $X_{2,i}$, then $\beta^M = \text{logit}\, E(Y_{ij} \mid X_{2,i} = 1) - \text{logit}\, E(Y_{i'j'} \mid X_{2,i'} = 0)$ measures the variation in the log odds of success "between groups." In the logistic-normal model, we explicitly assume that there exists individual heterogeneity. But for the group contrast $\beta^M$, we average over this distribution within each group.

The interpretation of the parameters $\boldsymbol{\alpha}$ that specify the random effects distribution will depend on the particular model used. In the case of Gaussian random effects, the variance components have simple interpretations as measures of within-group variation. In a marginally specified logistic-normal model with $b_{ij} = b_{2,i}$ and $b_{2,i} \sim N(0, \sigma^2)$, we can substitute $b_{ij} = \sigma z_i$, where $z_i \sim N(0, 1)$, and reex-

press the model for random individual variation using the conditional expectation: logit $E(Y_{ij}|\mathbf{X}_i, z_i) = \Delta(\mathbf{X}_{2,i}) + \sigma z_i$. This representation shows that the variance component $\sigma$ may be interpreted as a regression coefficient for a standardized omitted covariate, with $\sigma$ contrasting individuals with equal $\Delta(\mathbf{X}_{2,i})$ whose $z_i$ differ by one unit. Since $\Delta(\mathbf{X}_{2,i})$ is determined by $\mathbf{X}_{2,i}$ and the parameters $(\boldsymbol{\beta}^M, \sigma)$, subgroups defined by $\mathbf{X}_{2,i}$ are the same as subgroups defined by $\Delta(\mathbf{X}_{2,i})$. Therefore, $\sigma$ measures the magnitude of variation in the log odds "between individuals" within a group, where the group is defined by the measured covariates.

One motivation for adopting the conditionally specified logistic-normal model is that cluster-specific effects can be estimated (Zeger, Liang and Albert, 1988). Since we adopt a model for the marginal mean, $\beta^M$ cannot be given an individual level intervention interpretation. However, we are able to compute the corresponding conditional log odds $\Delta(X_{ijk})$ based on the marginally specified logistic-normal model and can therefore provide $\Delta(X_{ijk} + 1) - \Delta(X_{ijk})$ as an estimate of the change in log odds at the individual level. Thus, although our regression focus is on the marginal mean, the use of an underlying logistic-normal model yields estimates of individual level effects as model summaries.

## 3. LIKELIHOOD ESTIMATION FOR MARGINALIZED MULTILEVEL MODELS

In this section we summarize likelihood estimation for a marginally specified latent variable model. Likelihood inference for GLMMs has been an active research area recently (McCulloch, 1997; Booth and Hobert, 1999). Evaluation of the likelihood function usually requires numerical multivariate integration over the distribution of $\mathbf{b}_i = \text{vec}(b_{ijk})$. Adoption of a marginal mean regression adds some complexity since it requires calculation of the implicitly defined conditional mean parameter $\Delta(\mathbf{X}_{ijk})$. However, the conditional mean parameter can be easily obtained as the solution to an integral equation using only one-dimensional numerical integration since linkage of the marginal and conditional means only requires integration over the univariate marginal distribution of $b_{ijk}$.

The likelihood contribution from measurements $\mathbf{Y}_i = \text{vec}(Y_{ijk})$ can be constructed given the assumptions of conditional independence given $\mathbf{b}_i$ and the assumption that $[\mathbf{b}_i|\mathbf{X}_i]$ follows a mixing distribution known up to a finite parameter $\alpha$. First we discuss a general response model–mixing distribution combination and then give details for the three-level logistic-normal model.

### 3.1 General Hierarchical Models

Let the distribution $[Y_{ijk}|\mathbf{X}_{ijk}, b_{ijk}]$ be a member of the exponential family with conditional canonical parameter $\theta^b_{ijk}$, and scale parameter $\phi$. Then with a canonical link function we have that $\theta^b_{ijk} = \Delta(\mathbf{X}_{ijk}) + b_{ijk}$. Assuming that response variables are conditionally independent given the latent variables, a marginal likelihood function for the observed data $\mathbf{Y}_i$ is given by

$$P_{\boldsymbol{\theta}}(\mathbf{Y}_i|\mathbf{X}_i) = \int_{\mathbf{b}_i} \exp\left\{ \sum_{j,k} \frac{\theta^b_{ijk} Y_{ijk} - c\left(\theta^b_{ijk}\right)}{\phi} + d(Y_{ijk}, \phi) \right\} \cdot dF_{\boldsymbol{\alpha}}(\mathbf{b}_i|\mathbf{X}_i).$$

In general this integral cannot be obtained analytically and numerical methods are required for likelihood evaluation and parameter estimation.

A growing literature exists on approaches to maximizing the generalized linear mixed model likelihood function. Some of the approaches include approximate maximum likelihood (ML) solutions (Stiratelli, Laird and Ware, 1984; Goldstein, 1991; Breslow and Clayton, 1993), Monte Carlo EM algorithms (McCulloch, 1997; Booth and Hobert, 1999), Monte Carlo Newton–Raphson algorithms (McCulloch, 1997), direct use of numerical integration (Hedeker and Gibbons, 1994; Gibbons and Hedeker, 1997) and Markov chain Monte Carlo (MCMC) approaches for posterior inference with Bayesian models that also include proper priors on the regression and variance component parameters (Zeger and Karim, 1991; Gilks, Richardson and Spiegelhalter, 1996). To date, all of these algorithms have only addressed the conditionally specified model where the conditional mean function is directly parameterized, $\Delta(\mathbf{X}_{ijk}) = \mathbf{X}_{ijk}\boldsymbol{\beta}^C$.

Modification of existing algorithms to fit the marginalized model is possible if the implicitly defined canonical parameters $\theta^b_{ijk}$, or equivalently $\Delta(\mathbf{X}_{ijk})$, can be obtained as a function of $(\boldsymbol{\beta}^M, \boldsymbol{\alpha})$. This is achieved through numerical solution of the convolution equation (6) that connects the marginal and conditional mean functions. The partial derivatives of $\Delta(\mathbf{X}_{ijk})$ are also required and are obtained via implicit differentiation of the convolution equation. Details are provided in the Appendix.

To numerically evaluate the convolution equation, one can use either a general method such as Gauss–Hermite quadrature (Abramowitz and Stegun, 1972) or a specialized method tailored to the link–mixture combination; for example, see Monahan and Stefanski's (1992) method for the logistic-normal model.

### 3.2 Multilevel Logistic-Normal Model

Our examples focus on two-level and three-level binary logistic models. Heagerty (1999) describes likelihood estimation for the two-level marginalized logistic-normal model. In this section we extend those results to a three-level model and describe how multilevel likelihood evaluation can be simplified using nested quadrature.

Let $\mathbf{b}_i = \text{vec}(b_{ijk})$, where $b_{ijk} = b_{2,ij} + b_{3,i}$ represents shared but unobserved level-2 and level-3 effects. We further assume that $b_{2,ij} = \sigma_2(\mathbf{X}_{2,ij})z_{ij}$, $z_{ij} \sim N(0,1)$, and $b_{3,i} = \sigma_3(\mathbf{X}_{3,i})z_i$, $z_i \sim N(0,1)$, are mutually independent. Let $\boldsymbol{\alpha}$ be parameters that identify $\sigma_2(\mathbf{X}_{2,ij})$ and $\sigma_3(\mathbf{X}_{3,i})$.

The observed data likelihood for a level-3 cluster $i$ is a mixture over the level-2 and level-3 random effects distributions. By the assumption that response variables are independent conditional on $\mathbf{b}_i$, the likelihood function is

$$L_i(\boldsymbol{\beta}^M, \boldsymbol{\alpha})$$

$$(7) \qquad = \int_{b_{3,i}} \prod_j \left[ \int_{b_{2,ij}} \prod_k P\big(Y_{ijk} = y_{ijk} \,|\, \mathbf{X}_{ijk}, \cdot b_{2,ij}, b_{3,i}\big)\, dF_{b_{2,ij}} \right] dF_{b_{3,i}}$$

$$(8) \qquad = \int_{z_i} \prod_j \left\{ \int_{z_{ij}} \prod_k h_{ijk}(z_{ij}, z_i)^{y_{ijk}} \cdot \big[1 - h_{ijk}(z_{ij}, z_i)\big]^{1-y_{ijk}} \phi(z_{ij})\, dz_{ij} \right\} \cdot \phi(z_i)\, dz_i,$$

where

$$h_{ijk}(z_{ij}, z_i) = h\big[\Delta(\mathbf{X}_{ijk}) + \sigma_2(\mathbf{X}_{2,ij})z_{ij} + \sigma_3(\mathbf{X}_{3,i})z_i\big].$$

Simple univariate integration allows evaluation of the inner integrals in equation (8) (over the distribution of $z_{ij}$), denoted by $L_{ij}(z_i)$, using

$$L_{ij}(z_i) \approx \sum_t w_t \exp\left\{ \sum_k y_{ijk} \log h_{ijk}(z_t^*, z_i) + (1 - y_{ijk}) \log\big[1 - h_{ijk}(z_t^*, z_i)\big] \right\},$$

where $(w_t, z_t^*)$ represent the quadrature weights and evaluation points. In this integral, the variable $z_i$ is an offset in the conditional linear predictor. The outer integral in (8) can then be evaluated by

$$L_i(\boldsymbol{\beta}^M, \boldsymbol{\alpha}) = \int_{z_i} \exp\left[ \sum_j \log L_{ij}(z_i) \right] \phi(z_i)\, dz_i \approx \sum_s w_s \exp\left[ \sum_j \log L_{ij}(z_s^*) \right]$$

so that nested quadrature can be employed for evaluation of the likelihood function. Calculation of the score equations and the information matrix similarly requires numerical integration and requires derivatives of $\Delta(\mathbf{X}_{ijk})$, $\sigma_2(\mathbf{X}_{2,ij})$ and $\sigma_3(\mathbf{X}_{3,i})$ with respect to $(\boldsymbol{\beta}^M, \boldsymbol{\alpha})$. Details regarding the derivatives of $\Delta(\mathbf{X}_{ijk})$ are given in the Appendix. Gibbons and Hedeker (1997) provide further algorithm details for conditionally specified three-level probit and logistic models.

### 3.3 Empirical Bayes Estimation of $b_{ijk}$

Given point estimates for the mean parameters $\boldsymbol{\beta}^M$ and the variance components $\boldsymbol{\alpha}$, we can estimate the random effect $b_{ijk}$ by $\tilde{b}_{ijk}$, the mode of the posterior distribution $[b_{ijk} \,|\, \mathbf{Y}_i, \mathbf{X}_i]$, fixing the parameters at their estimated values. Finding the posterior mode corresponds to solving the posterior score equations for $\mathbf{b}_i$ given by

$$\mathbf{0} = \sum_{j,k} \left( \frac{\partial \mu_{ijk}^b}{\partial b} \right) \text{var}\big(Y_{ijk} \mid b_{ijk}\big)^{-1} \cdot \Big\{ Y_{ijk} - h\big[\hat{\Delta}(\mathbf{X}_{ijk}) + b_{ijk}\big] \Big\} - \hat{\boldsymbol{\Sigma}}_i^{-} \mathbf{b}_i,$$

where $\boldsymbol{\Sigma}_i = \text{cov}(\mathbf{b}_i)$.

Empirical Bayes estimates allows estimation of conditional means $\mu_{ijk}^b$. In Section 5 we show how these can be compared to the marginal means $\mu_{ijk}$ to communicate the relative magnitudes of systematic variation attributable to covariates, and unmeasured variation represented by the latent variables.

## 4. BIAS DUE TO MODEL MISSPECIFICATION

One advantage to using estimating equations to estimate marginal regression parameters is that inference can be made robust to misspecification of the dependence model. Liang and Zeger (1986) show that the working model used for $\text{cov}(\mathbf{Y}_i)$ does not impact the consistency of $\hat{\boldsymbol{\beta}}^M$, the root of the estimating equations. It is therefore important to assess the sensitivity of likelihood-based estimation methods to their additional assumptions. In this section we assume that the mean model is correctly specified and focus on the impact of misspecification of the latent variable distribution.

Neuhaus, Hauck and Kalbfleisch (1992) have studied the impact of assuming $b_{2,i} \sim F$ when in truth $b_{2,i} \sim G$. They show that the logistic-normal maximum-likelihood estimate (MLE) of a conditionally specified mean parameter has bias of less than 20% when random effects are nonnormally distributed. These results suggest that likelihood-based mean estimates using a Gaussian latent variable model may be moderately insensitive to distributional assumptions.

We explore a potentially more serious form of bias that arises due to incorrectly assuming that the variance of the random effects is independent of the covariates. In particular, we consider a two-level model where the random effects variance differs according to a level-2 binary covariate. We assess the impact of this form of model misspecification on both the marginally specified and the conditionally specified mean parameters. Our investigation is motivated in part by observations made while conducting the analysis for the example presented in Section 5.1.

Specifically, suppose that

$$\text{logit } E\big(Y_{ij} \mid X_{1,ij}, X_{2,i}, b_{2,i}\big) = \Delta(\mathbf{X}_i) + b_{2,i},$$

$$V(b_{2,i} \mid X_{2,i} = 0) = \sigma_0^2$$

and

$$V(b_{2,i} \mid X_{2,i} = 1) = \sigma_1^2$$

and that we incorrectly assume $V(b_{2,i} \mid X_i) = \sigma^2$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_0, \sigma_1)$ and let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}, \sigma, \sigma)$ be an element of the subspace where $\sigma_0 = \sigma_1$. White (1982) shows that the misspecified MLE $\boldsymbol{\theta}^*$ converges to the value $\boldsymbol{\theta}^*$ such that

$$\sum_i E_{\boldsymbol{\theta}}\left\{\frac{\partial}{\partial \boldsymbol{\theta}^*} \log P(\mathbf{Y}_i; \mathbf{X}_i, \boldsymbol{\theta}^*)\right\} = \mathbf{0}.$$

To find the value $\boldsymbol{\theta}^*$ corresponding to a given $\boldsymbol{\theta}$, we use Monte Carlo integration to compute the $2^{n_i}$ probability vector $P(\mathbf{Y}_i; \mathbf{X}_i, \boldsymbol{\theta})$ and then numerically solve for $\boldsymbol{\theta}^*$. That is, we compute the probability $P(\mathbf{Y}_i = \mathbf{y}; \mathbf{X}_i, \boldsymbol{\theta})$ for every possible binary response vector $\mathbf{y}$ and use these as weights to solve the score equations for pseudodata composed of each of the $2^{n_i}$ possible response vectors (Rotnitzky and Wypij, 1994).

We consider values for $\sigma_0$ and $\sigma_1$ that range between 0.0 and 3.0 in the conditionally specified model

$$\text{logit } E\big(Y_{ij} \mid X_{1,ij}, X_{2,i}, b_{2,i}\big) = \beta_0^C + \beta_1^C X_{1,ij} + \beta_2^C X_{2,i} + b_{2,i},$$

where $j = 1, 2, 3, 4$, $X_{1,ij} = 0$ for $j = 1, 2$ and $X_{1,ij} = 1$ for $j = 3, 4$, and $X_{2,i} = 0$ for half the clusters and $X_{2,i} = 1$ for the other half. Although we studied a range of different parameter values we present results for a single value $\boldsymbol{\beta}^C = (-2.0, 1.0, 0.5)$ to illustrate bias as a function of $(\sigma_0, \sigma_1)$.

Figure 1a shows the relative bias in the level-2 covariate estimate $\widehat{\beta_2}^C$ due to incorrectly assuming that the heterogeneity is constant. For example, when $\sigma_0 = 1.0$ and $\sigma_1 = 2.0$ the MLE for $\beta_2^C$ converges to 1.34, a relative bias of $+34\%$. The bias can be potentially large, ranging from $+80\%$ to $-75\%$ for the range of parameter values consid-

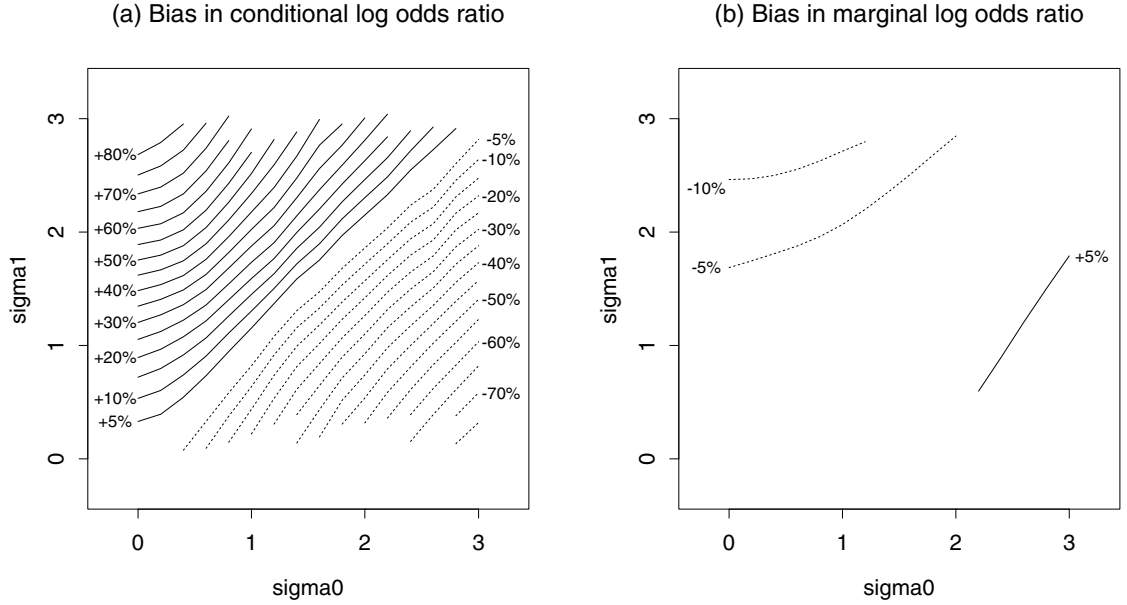


FIG. 1. *Asymptotic relative bias in the maximum likelihood estimate of the treatment contrast when $\sigma_0 = \sigma_1$ is incorrectly assumed.*

ered. There was little asymptotic bias in $\widehat{\beta_1}^C$, the estimated coefficient for the covariate that varies within cluster, $X_{1,ij}$. These results suggest that if scientific interest is in conditionally specified contrasts across clusters (level-2), then the dependence model must be correctly specified otherwise severe bias may result.

We also considered the analogous marginally specified latent variable model where

$$\text{logit } E\big(Y_{ij} \mid X_{1,ij}, X_{2,i}\big) = \beta_0^M + \beta_1^M X_{1,ij} + \beta_2^M X_{2,i}$$

with $\boldsymbol{\beta}^M = (-2.0, 1.0, 0.5)$ to illustrate bias as a function of the variance components. Figure 1b shows the bias in the level-2 covariate estimate $\widehat{\beta_2}^M$. The asymptotic relative bias ranges from $+10\%$ to $-15\%$, indicating that the MLE for the marginal contrast is much less sensitive to variance component specification.

These asymptotic bias calculations illustrate the following points: (1) conditional contrasts for within-cluster covariates may not be sensitive to correct specification of $[b_{2,i} | \mathbf{X}_i]$; (2) conditional contrasts for between-cluster covariates can be highly sensitive to assumptions regarding $[b_{2,i} | \mathbf{X}_i]$ with potentially severe bias; and (3) marginally specified coefficient ML estimates may be biased due to variance component misspecification but the magnitude of the bias is generally small.

In this section we have considered just one form of model misspecification. Other types of model violation are also important to explore, including mean misspecification through omitting covariates, and other potential violations of the random effects model such as assuming a random intercept when there is serial correlation in the response.

## 5. EXAMPLES

In this section we apply the new marginal multilevel model to the data sets described in Sections 1.1 and 1.2.

### 5.1 Two-Level Data: Weil Teratology Experiment

Weil (1970) presents data where the 21-day survival of pups from the litters of 16 exposed and 16 unexposed rats is compared. Let $Y_{ij}$ denote the survival of pup $j$, $j = 1, 2, \ldots, N_i$ (level-1), born to animal $i$, $i = 1, 2, \ldots, m$ (level-2). The single covariate of interest is a level-2 binary indicator of the treatment assignment of the mother. The raw data are published in Liang and Hanfelt (1994).

In each group the proportion that survive can be summarized by weighted averages of the proportion surviving for each mother, $\hat{p} = \Sigma_i w_i \hat{p}_i / \Sigma_i w_i$, where $\hat{p}_i = \Sigma_j Y_{ij} / N_i$. Using $w_i = 1$ yields simple averages of the individual proportions: $\hat{p}(X = 0) = 0.893$ and $\hat{p}(X = 1) = 0.746$ and a difference in log odds of $\text{logit}(0.746) - \text{logit}(0.893) = -1.05$. Using $w_i = N_i$ yields an estimator that simply divides the total number of surviving pups by the total number born, $\hat{p}(X = 0) = 0.899$ and $\hat{p}(X = 1) = 0.772$ with the treatment log odds contrast, $-0.961$. We can summarize the litter-to-litter variation in each treatment group by calculating the standard deviation of the empirical logits, $\log(Y_i + 1/2) - \log(N_i - Y_i + 1/2)$, where $Y_i = \Sigma_j Y_{ij}$. In the control group this standard deviation is 0.895 while in the treatment group it is 1.614. These data summaries suggest that the probability of pup survival is reduced for exposed mothers and that there is more between-litter variation in the exposed group. We can use a marginalized two-level logistic model to obtain model-based estimates of these components of variation and to perform confirmatory tests of the observed trends.

5.1.1 *Marginally specified models*. Consider the marginal linear logistic mean model

$$\text{logit } E\big(Y_{ij} \mid X_{2,i}\big) = \beta_0^M + \beta_1^M X_{2,i},$$

where $X_{2,i} = 1$ if mother was exposed to the chemical agent and 0 otherwise. Heterogeneity is introduced via the conditional logistic normal model

$$\text{logit } E\big(Y_{ij} \mid X_{2,i}, b_{2,i}\big) = \Delta(X_{2,i}) + b_{2,i},$$

where $[b_{2,i} \mid X_{2,i}]$ is assumed to be Gaussian with standard deviation $\sigma_2(X_{2,i})$ that may depend on the level-2 covariate, mother's exposure status. We make the standard assumption that conditional on $b_{2,i}$ the response variables $Y_{ij}$ and $Y_{ik}$ are independent Bernoulli random variables.

Table 1 shows the fitted marginal parameter estimates for a model that assumes a common variance component $\sigma_2(X = 0) = \sigma_2(X = 1)$ (referred to as Model 2), for a model that allows separate heterogeneity parameters (Model 3) and for a null model (Model 1). A likelihood ratio test of $H_0$: $\sigma_2(X = 0) = \sigma_2(X = 1)$ yields a change in deviance of $2 \times (118.20 - 116.33) = 3.74$ and a $p$-value of 0.053. Allowing the within-group heterogeneity to differ for the treatment groups yields the estimates $\hat{\sigma}_2(X = 0) = 0.451$ and $\hat{\sigma}_2(X = 01) = 0.451 + 1.362 = 1.816$. Thus these data suggest a minor amount of litter-to-litter variation in the control group but substantial variation within the treated group. The separate variance components model (Model 3) results in the point estimate $\hat{\beta}_1^M = -1.069$ with a 95% confidence interval of $(-2.00, -0.14)$. Using a common variance component (Model 2) results in a 20% smaller point estimate $\hat{\beta}_1^M = -0.867$ with a

TABLE 1
*Logistic-normal model estimates for teratology data presented in Weil* (1970); *a single binary covariate is used, where* $X_i = 1$ *for animals in the treatment group and* $X_i = 0$ *for animals in the control group*

| | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| **Coefficient** | **Estimate** | **s.e.** | **Estimate** | **s.e.** | **Estimate** | **s.e.** |
| *Marginal mean* ($\beta^M$) | | | | | | |
| Intercept | 1.540 | 0.263 | 2.031 | 0.395 | 2.175 | 0.286 |
| Treatment | | | −0.867 | 0.507 | −1.069 | 0.476 |
| *Level* 2 *heterogeneity* ($\sigma_2$) | | | | | | |
| Intercept | 1.476 | 0.345 | 1.345 | 0.332 | 0.451 | 0.572 |
| Treatment | | | | | 1.362 | 0.777 |
| log $L$ | −119.63 | | −118.20 | | −116.33 | |

| | Model 1* | | Model 2* | | Model 3* | |
|---|---|---|---|---|---|---|
| **Coefficient** | **Estimate** | **s.e.** | **Estimate** | **s.e.** | **Estimate** | **s.e.** |
| *Conditional mean* ($\beta^C$) | | | | | | |
| Intercept | 2.103 | 0.355 | 2.624 | 0.483 | 2.254 | 0.340 |
| Treatment | | | −1.080 | 0.626 | −0.565 | 0.669 |
| *Level* 2 *heterogeneity* ($\sigma_2$) | | | | | | |
| Intercept | 1.476 | 0.345 | 1.345 | 0.332 | 0.451 | 0.572 |
| Treatment | | | | | 1.362 | 0.777 |
| log $L$ | −119.63 | | −118.20 | | −116.33 | |

95% confidence of $(-1.96, 0.13)$. We see that the specification of the covariance model has some impact on the estimated treatment effect as reflected in the marginal mean.

5.1.2 *Conditionally specified models*. Alternatively, we may consider use of the traditional conditional logistic-normal parameterization

$$\text{logit } E\big(Y_{ij} \mid X_{2,i}, b_i\big) = \beta_0^C + \beta_1^C X_{2,i} + b_{2,i},$$

where $[b_{2,i} \mid X_{2,i}] \sim N[0, \sigma_2^2(X_{2,i})]$. The key difference in this approach is that the mean parameter $\boldsymbol{\beta}^C$ contrasts differences in the measured covariates for fixed values of the unobserved random effect $b_{2,i}$. When there is level-1 variation in covariates this conditional contrast becomes a pure within-cluster contrast. However, in this example the treatment covariate is a level-2 variable and as such there is no direct observable contrast between $X_{2,i} = 1$ and $X_{2,i'} = 0$, where $b_{2,i} = b_{2,i'}$. Such a contrast, although not directly observed, is estimated by the coefficient $\hat{\beta}_1^C$ and can justifiably be considered an extrapolation of the data.

Table 1 shows the fitted conditional parameter estimates for both the common heterogeneity (Model 2*) and the separate heterogeneity (Model 3*) variance components models. In this simple scenario with a single binary covariate there is an exact correspondence between the marginal and the conditional mean specifications. This is seen by the fact that the maximized log-likelihoods and variance component estimates are identical. However, using the conditionally specified mean in Model 3* yields a treatment contrast of $\hat{\beta}_1^C = -0.565$ with a 95% confidence interval of $(-1.88, 0.75)$. If we assume that $\sigma_2(X = 0) = \sigma_2(X = 1)$, then the conditional contrast nearly doubles with the point estimate $\hat{\beta}_1^C = -1.08$ and 95% confidence interval $(-2.31, 0.15)$ (Model 2). These regression models illustrate that the conditionally specified mean estimates and inferences are considerably more sensitive to the variance component assumptions, particularly for covariates that only vary between, and not within, clusters.

5.1.3 *Profile likelihood*. One advantage of likelihood-based inference is that interval estimation and hypothesis testing need not rely solely on the asymptotic normality of $\hat{\boldsymbol{\theta}}$. Likelihood or profile likelihood functions can be used to display the evidence in the data regarding key parameters. Figure 2 shows the profile likelihood functions for the marginal and conditional treatment contrast parameters. In the marginally specified model the profile likelihood function corresponds to $\log L\{\beta_1^M; \hat{\beta}_0^M(\beta_1^M), \hat{\boldsymbol{\alpha}}(\beta_1^M)\}$, where $\boldsymbol{\alpha}$ represents the variance components. Using the profile likelihood based on the assumption of separate variance components yields 95% likelihood ratio (LR) confidence inter-

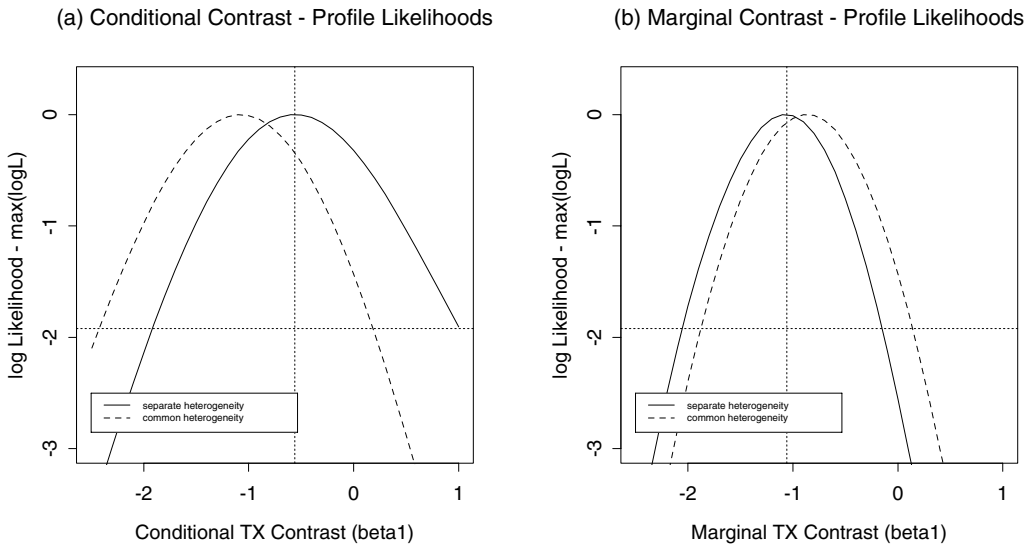


FIG. 2. *Profile likelihood curves for the treatment contrast parameter* (a) *using conditionally specified models and* (b) *using marginally specified models*: (*solid lines*) *profile likelihoods under the variance components model that assumes* $\sigma_0 \neq \sigma_1$; (*dashed lines*) *profile likelihoods under the model that assumes* $\sigma_0 = \sigma_1$; (*vertical dashed lines*) ***MLE*** *for the mean contrast under the separate variance components model*; (*horizontal dashed lines*) *at* $-1/2$ *times the* 95*th percentile of a* $\chi^2(1)$, *showing the profile likelihood confidence interval limits*.

vals $\beta_1^M \in (-2.05, -0.15)$ and $\beta_1^C \in (-1.92, 1.01)$. The LR confidence interval for the marginal parameter is only slightly larger than the Wald-based interval while for the conditional parameter the LR interval is skewed to the right, with the upper limit at 1.01 compared to 0.75 when constructed using the Wald statistic.

5.1.4 *Components of variation*: *graphical displays*. Since the logistic-normal standard deviation estimate is on the scale of the line predictor, we can directly compare the magnitude of $\hat{\sigma}(X_{2,i})$ to $\hat{\beta}_1^M$. To display these components of variation we plot the estimated conditional log odds $\hat{\Delta}(X_{2,i}) + \tilde{b}_{2,i}$ using approximate empirical Bayes estimates of the latent variables, as discussed in Section 3.3. In Figure 3 the marginal group contrast is displayed as the difference between the solid lines, and the within-group heterogeneity is seen by the variation among the individual empirical Bayes estimates. This graphical representation of both data and

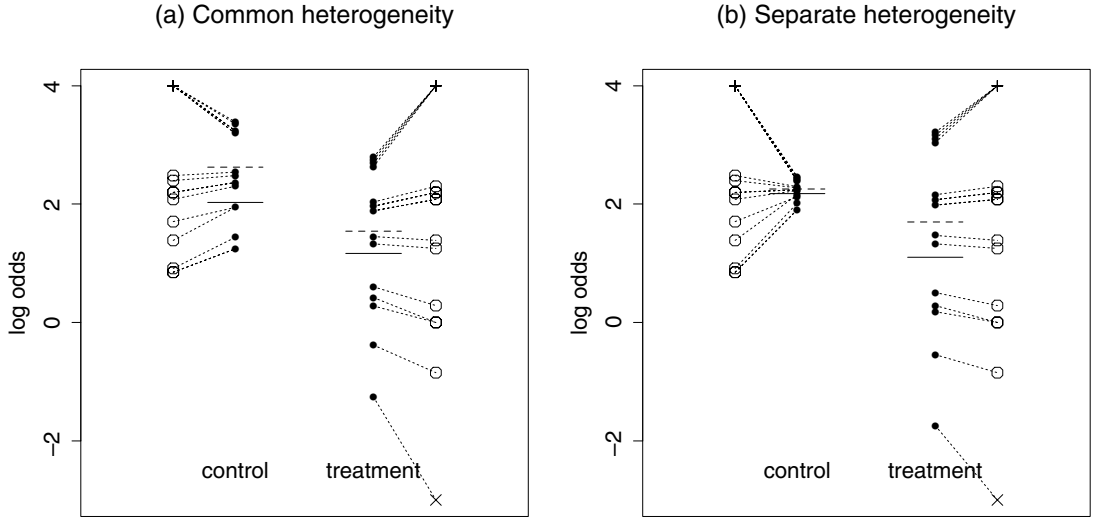


FIG. 3. *Teratology data*: *estimated conditional log odds* $\hat{\Delta}(\mathbf{X}_i) + \tilde{b}_i$ *using empirical Bayes estimates of individual effects are represented by* (●); *empirical logits* $\text{logit}(Y_i/N_i)$ *are represented by the open circles* (○). *Values of* $\text{logit}(Y_i/N_i)$ *that are infinite are represented with either* + *or* ×. *The solid horizontal lines show the marginal log odds* $\eta(\mathbf{X}_i)$, *and the dashed horizontal lines show the conditional log odds* $\hat{\Delta}(\mathbf{X}_i)$.

model components facilitates the relative comparison of the controlled (treatment assignment) and uncontrolled (individual animal variation) effects. The display of empirical Bayes estimates is also useful as a model diagnostic since it suggests that the degree of heterogeneity may be larger in the treatment group. One caveat is that the distribution of estimated random effects is known to underrepresent the true variability (Shen and Louis, 1998).

5.1.5 *Comparison with beta-binomial and quasi-likelihood approaches*. Liang and Hanfelt (1994) used both the beta-binomial model and quasilikelihood to analyze these data. The beta-binomial model with a single overdispersion parameter gives $\hat{\beta}_1^M = -0.665$ with a confidence interval (CI) of $(-1.57, 0.24)$. Allowing separate dispersion parameters results in $\hat{\beta}_1^M = -1.129$ with 95% CI $(-2.04, -0.22)$ showing that regression inferences from the beta-binomial model can also be very sensitive to the covariance assumptions. Using quasilikelihood methods Liang and Hanfelt (1994) obtain the point estimate $\hat{\beta}_1^M = -0.961$ using a scale variance model (with scale parameter $\phi_0 = 1.46$ in the unexposed group and $\phi_1 = 4.74$ in the exposed group), and $\hat{\beta}_1^M = -1.070$ using a beta-binomial variance function (with an intralitter correlation of $\hat{\rho}_0 = 0.05$ in the unexposed group and $\hat{\rho}_1 = 0.46$ in the exposed group). One advantage to our model formulation is that the heterogeneity parameters are on the scale of the linear predictor, thus facilitating interpretation of the magnitude of overdispersion. Finally, although we focus on likelihood inference in this manuscript, Heagerty (1999) discusses how quasilikelihood (estimating functions) can also be used to estimate the multilevel model parameters.

### 5.2 Three-Level Data: British Social Survey

Wiggins, Ashworth, O'Muircheartaigh and Galbraith (1990) describe data obtained through the *British Social Attitudes Panel Survey* conducted from 1983 through 1986. Subjects were asked whether they thought the law should allow an abortion in each of seven scenarios. Approximately 30% of the subjects replied "yes" to all seven cases, indicating that there should be no legal or governmental regulation of abortion. Our analysis of these data are for the derived response 1 = "the law should permit abortion for each of the seven situations," 0 = "there exist some situations where the law should not allow abortion." Logistic regression analysis of this response considers how the proportion of subjects whose opinion is summarized as "no legal restriction" versus "possible legal restriction" varies as a function of covariates.

The data we analyze are based on 264 participants that responded at each of the 4 study waves (annually from 1983 through 1986) and have complete covariate information. Our analysis is primarily intended to illustrate the statistical methodology, but it also complements earlier analyses (Wiggins et al., 1983). Details regarding the design of the survey can be found in McGrath and Waterton (1986). Since the data were sampled by polling district, the resulting data structure has nested clusters, with observations (level-1, $n_1 = 1{,}056$) nested within individuals (level-2, $n_2 = 264$) nested within districts (level-3, $n_3 = 54$).

Let $Y_{ijk}$ denote the response for district $i$, subject $j$ and year $k$ with $k = 1, 2, 3, 4$. Let $\mathbf{X}_{ijk} = (\mathbf{X}_{1,ijk}, \mathbf{X}_{2,ij}, \mathbf{X}_{3,i})$ denote the complete covariate vector for observation $Y_{ijk}$. In this example, the level-1 covariates are the indicator variables for time, $X_{1,ijk} = 1$ if year $= k$, for $k \geq 2$, while the level-2 covariates represent demographic characteristics of the subjects including social class (middle, upper working, lower working), gender (male, female) and religion (Protestant, Catholic, other, none). We also use a derived level-3 covariate, $X_{3,i}$, which is the percentage of the subjects within the district that are Protestant. Such a variable allows us to illustrate the decomposition of religion contrasts into both within-cluster and between-cluster comparisons (Neuhaus and Kalbfleisch, 1998). This variable is potentially of substantive interest since it measures the religious context or environment for the individuals in contrast to their own religious affiliation.

5.2.1 *Marginally specified models*. We first consider the following regression model for the marginal log odds $\eta(\mathbf{X}_{ijk})$:

$$\begin{aligned}\eta(\mathbf{X}_{ijk}) &= \text{logit } E\big(Y_{ijk}|\mathbf{X}_{ijk}\big) \\ &= \beta_0^M + \boldsymbol{\beta}_1^M \mathbf{X}_{1,ijk} + \boldsymbol{\beta}_2^M \mathbf{X}_{2,ij} + \beta_3^M X_{3,i}.\end{aligned}$$

The multilevel model is completed with the second conditional assumption

$$\begin{aligned}&\text{logit } E\big(Y_{ijk}\,|\mathbf{X}_{ijk}, z_{2,ij}, z_{3,i}\big) \\ &\quad = \Delta(\mathbf{X}_{ijk}) + \sigma_2(\mathbf{X}_{2,ij}) z_{2,ij} + \sigma_3 z_{3,i}\end{aligned}$$

for $z_{2,ij} \sim N(0,1)$ and $z_{3,i} \sim N(0,1)$ mutually independent. We assume throughout that $\sigma_3(\mathbf{X}_{3,i})$ does not depend on $X_{3,i}$ and is therefore simply a scalar. We consider several models that allow $\sigma_2(\mathbf{X}_{2,ij})$ to depend on each of the demographic variables.

Table 2 shows the results for several marginalized multilevel models. The "Independence" model is an ordinary logistic model which ignores the data clustering and for which the point estimate $\hat{\boldsymbol{\beta}}^M$ is consistent for $\boldsymbol{\beta}^M$ (Liang and Zeger, 1986) but the standard errors may be grossly incorrect. Model 1 fits a simple multilevel model with a scalar $\sigma_2$ and $\sigma_3$. We see that its point estimate $\hat{\boldsymbol{\beta}}^M$ is quite comparable to the ordinary logistic regression estimate. However, the resulting standard errors are substantially different. For variables that vary within cluster, such as the indicators for year, the logistic-normal model standard errors are smaller by approximately 25%. For variables that vary between clusters, the positive correlation results in standard errors that are larger by 40–60%. The resulting correction to the standard errors has a major impact on inference regarding the regression contrasts. This is not surprising given that the level-2 heterogeneity parameter is estimated as $\hat{\sigma}_2 = 2.14$, which represents substantial random individual-to-individual variation. The heterogeneity estimate $\hat{\sigma}_3 = 0.82$ measures the unexplained district-to-district variation.

Model 1 is based on the assumption that $[b_{2,ij}|\mathbf{X}_{2,ij}]$ does not depend on the covariates $\mathbf{X}_{2,ij}$. In Models 2 through 4, we relax this strong assumption. Use of these models serves to characterize person-to-person variation as a function of person-level characteristics and allows an assessment of the sensitivity of the mean parameter estimates to the simple assumption used in Model 1. Comparing Models 1 and 3 with a likelihood ratio test, we find moderate evidence that there is greater variation among women than among men, with a change in deviance of 3.12 ($p$-value = 0.077). Model 3 estimates the variation among males as $\hat{\sigma}_2(X_{2,ij4} = 0) = 1.69$ and among females as $\hat{\sigma}_2(X_{2,ij4} = 1) = 1.69 + 0.87 = 2.56$. One interpretation of these estimates is that women appear to respond more similarly over time than do men. Choice of the level-2 heterogeneity model does impact point estimates and standard errors of the marginal mean regression parameters. However, the fluctuation in the point estimates is small, on the order of $\pm 10\%$ as seen in Table 2.

The mean regression models decompose religion contrasts into within-cluster contrasts and be-

TABLE 2
*Marginal mean models for British Social Survey data*

| | Independence | | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Coefficient** | **Estimate** | **(s.e.)** | **Estimate** | **(s.e.)** | **Estimate** | **(s.e.)** | **Estimate** | **(s.e.)** | **Estimate** | **(s.e.)** |
| *Marginal mean* ($\beta^M$) | | | | | | | | | | |
| Intercept | −0.792 | (0.287) | −0.763 | (0.393) | −0.753 | (0.393) | −0.751 | (0.390) | −0.740 | (0.395) |
| Year: 2 | −0.433 | (0.200) | −0.446 | (0.153) | −0.417 | (0.156) | −0.438 | (0.154) | −0.453 | (0.152) |
| Year: 3 | 0.038 | (0.191) | 0.025 | (0.144) | 0.031 | (0.145) | 0.040 | (0.145) | 0.014 | (0.144) |
| Year: 4 | 0.181 | (0.189) | 0.165 | (0.143) | 0.172 | (0.144) | 0.147 | (0.144) | 0.155 | (0.142) |
| Class: upper working | −0.328 | (0.191) | −0.348 | (0.216) | −0.335 | (0.215) | −0.326 | (0.216) | −0.370 | (0.215) |
| Class: lower working | −0.431 | (0.167) | −0.267 | (0.208) | −0.269 | (0.208) | −0.272 | (0.206) | −0.300 | (0.208) |
| Gender | −0.279 | (0.140) | −0.349 | (0.205) | −0.364 | (0.206) | −0.315 | (0.205) | −0.320 | (0.205) |
| Religion: catholic | −0.421 | (0.341) | −0.384 | (0.480) | −0.416 | (0.476) | −0.406 | (0.477) | −0.389 | (0.471) |
| Religion: other | −0.601 | (0.250) | −0.634 | (0.360) | −0.657 | (0.366) | −0.700 | (0.365) | −0.712 | (0.343) |
| Religion: none | 0.718 | (0.179) | 0.707 | (0.256) | 0.653 | (0.260) | 0.678 | (0.253) | 0.704 | (0.258) |
| % protestant | 0.858 | (0.298) | 0.799 | (0.479) | 0.806 | (0.472) | 0.768 | (0.475) | 0.796 | (0.483) |
| *Level 2 heterogeneity* ($\sigma_2$) | | | | | | | | | | |
| Intercept | | | 2.140 | (0.238) | 2.274 | (0.316) | 1.689 | (0.338) | 2.433 | (0.404) |
| Class: upper working | | | | | 0.342 | (0.513) | | | | |
| Class: lower working | | | | | −0.460 | (0.599) | | | | |
| Gender | | | | | | | 0.871 | (0.464) | | |
| Religion: catholic | | | | | | | | | −0.581 | (0.946) |
| Religion: other | | | | | | | | | −1.143 | (0.661) |
| Religion: none | | | | | | | | | −0.301 | (0.589) |
| *Level 3 heterogeneity* ($\sigma_3$) | | | | | | | | | | |
| Intercept | | | 0.818 | (0.295) | 0.724 | (0.308) | 0.788 | (0.287) | 0.847 | (0.281) |
| log $L$ | −622.57 | | −531.92 | | −531.04 | | −530.36 | | −530.55 | |

tween-cluster contrasts (Neuhaus and Kalbfleisch, 1998). The variable %Protestant is a district-level covariate measuring the proportion of the sampled subjects that are Protestant. The coefficient estimate for this variable is 0.768 (Model 3) indicating increasing odds of response among subjects that live in districts with a higher proportion of Protestants. The categorical religion contrasts (Catholic, other, none, with the reference Protestant) are then interpreted as comparing response rates among subjects whose religious affiliation differ but who live within similar districts (the percentage of Protestants are equal). Nonsignificantly lower rates are observed among Catholics and other, while significantly higher rates are observed among those that report no religion.

5.2.2 *Conditionally specified models*. The conditionally specified multilevel model is given by

$$\begin{aligned}\operatorname{logit} E\left(Y_{ijk} \mid \mathbf{X}_{ijk}, b_{2,ij}, b_{3,i}\right) &= \beta_0^C + \mathbf{X}_{1,ijk}\boldsymbol{\beta}_1^C + \mathbf{X}_{2,ij}\boldsymbol{\beta}_2^C \\ &\quad + X_{3,i}\beta_3^C + b_{2,ij} + b_{3,i},\end{aligned}$$

where we assume $b_{2,ij} = \sigma_2(\mathbf{X}_{2,ij})z_{2,ij}$, $z_{2,ij} \sim N(0,1)$, and $b_{3,i} = \sigma_3 z_{3,i}$, $z_{3,i} \sim N(0,1)$, mutually independent. Table 3 presents model estimates for conditionally specified models with different assumptions regarding $\sigma_2(\mathbf{X}_{2,ij})$. Comparison of Model 1* to the corresponding marginal Model 1 in Table 2 shows the well-known relationship that $|\hat{\beta}_j^C| > |\hat{\beta}_j^M|$ for logistic-normal models (Zeger, Liang and Albert, 1988; Neuhaus and Jewell, 1993). In this example we see differences on the order of 50–85%. However, the key distinction is in the interpretation of these parameters. For example, the conditional gender contrast $\hat{\beta}_{2,3}^C = -0.600$ compares the log odds of a woman favoring no abortion restrictions compared to a man who otherwise has identical covariates, both $\mathbf{X}_{ijk}$ and $b_{ijk}$. Since the $b_{ijk}$ are unobserved and since the same person cannot be both a man and a woman, we cannot empirically control for $b_{ijk}$. In this sense, a conditional mean contrast for gender represents an extrapolation of the model to this unobservable scenario. It is possible to observe the difference in log odds for women compared to men that are otherwise equivalent with respect to measured covariates $\mathbf{X}_{ijk}$. An estimate of this is given by the corresponding marginal contrast, such as $\hat{\beta}_{2,3} = -0.349$ in Model 1.

Allowing dependence of $\sigma_2$ on covariates has a substantial impact on point estimates of $\boldsymbol{\beta}^C$ as seen

TABLE 3
*Conditional mean models for British Social Survey data*

| | Model 1* | | Model 2* | | Model 3* | | Model 4* | |
|---|---|---|---|---|---|---|---|---|
| **Coefficient** | **Estimate** | **(s.e.)** | **Estimate** | **(s.e.)** | **Estimate** | **(s.e.)** | **Estimate** | **(s.e.)** |
| *Conditional mean* ($\beta^C$) | | | | | | | | |
| Intercept | −1.388 | (0.685) | −1.279 | (0.693) | −1.001 | (0.665) | −1.452 | (0.679) |
| Year: 2 | −0.761 | (0.266) | −0.770 | (0.267) | −0.758 | (0.266) | −0.760 | (0.267) |
| Year: 3 | 0.060 | (0.252) | 0.060 | (0.252) | 0.062 | (0.252) | 0.060 | (0.252) |
| Year: 4 | 0.300 | (0.251) | 0.299 | (0.250) | 0.303 | (0.250) | 0.303 | (0.250) |
| Class: upper working | −0.623 | (0.378) | −0.587 | (0.427) | −0.667 | (0.374) | −0.708 | (0.374) |
| Class: lower working | −0.499 | (0.361) | −0.310 | (0.405) | −0.513 | (0.356) | −0.658 | (0.371) |
| Gender | −0.600 | (0.358) | −0.738 | (0.354) | −0.876 | (0.389) | −0.477 | (0.356) |
| Religion: catholic | −0.609 | (0.803) | −0.653 | (0.730) | −0.725 | (0.782) | −0.376 | (0.950) |
| Religion: other | −1.049 | (0.604) | −1348 | (0.616) | −1.319 | (0.615) | −0.487 | (0.586) |
| Religion: none | 1.263 | (0.452) | 0.861 | (0.469) | 1.019 | (0.445) | 1.384 | (0.481) |
| % protestant | 1.458 | (0.837) | 1.541 | (0.778) | 1.129 | (0.809) | 1.456 | (0.821) |
| *Level* 2 *heterogeneity* ($\sigma_2$) | | | | | | | | |
| Intercept | 2.138 | (0.236) | 2.776 | (0.592) | 1.642 | (0.294) | 2.450 | (0.408) |
| Class: upper working | | | −0.005 | (0.769) | | | | |
| Class: lower working | | | −1.135 | (0.662) | | | | |
| Gender | | | | | 0.994 | (0.494) | | |
| Religion: catholic | | | | | | | −0.498 | (1.055) |
| Religion: other | | | | | | | −1.222 | (0.668) |
| Religion: none | | | | | | | −0.348 | (0.608) |
| *Level* 3 *heterogeneity* ($\sigma_3$) | | | | | | | | |
| Intercept | 0.816 | (0.295) | 0.619 | (0.320) | 0.790 | (0.285) | 0.835 | (0.282) |
| log $L$ | −531.83 | | −529.42 | | −529.71 | | −530.35 | |

by comparing Models 1* through 4* in Table 3. Using likelihood ratio tests to compare Model 3* to Model 1* results in a change of deviance of 4.24, with $p$-value 0.039. Adopting Model 3*, where heterogeneity depends on gender, results in a 46% change in the estimate of the conditional gender contrast, from $-0.600$ to $-0.876$; the $t$-statistics for testing $\beta^C = 0$ also change dramatically from $-1.68$ for Model 1* to $-2.25$ in Model 3*. The values of the other level-2 and -3 regression coefficients in the conditional model vary by as much as a factor of 2 as the assumptions about the random effects vary.

The level-1 contrast for year is a pure within-cluster contrast and appears quite insensitive to the latent variable assumptions. These conditional contrasts are also estimable without making any assumptions regarding the distribution of $b_{2,ij} + b_{3,i}$ using conditional likelihood methods to eliminate the level-2 and level-3 random effects (Conaway, 1989). A conditional likelihood approach is not possible for level-2 and level-3 covariates since these do not vary within the lowest level of clustering and are therefore totally eliminated from the conditional likelihood.

## 6. DISCUSSION

This paper presents an alternate formulation of the popular multilevel model in which the marginal rather than the conditional mean given latent variables is modeled as a function of covariates. Our marginalization of the multilevel model is analogous to that of Fitzmaurice and Laird (1993) and Azzalini (1994) for conditional models given observed variables.

The regression coefficients in marginal multilevel models represent contrasts in the expected response given observed covariates, averaged over unobserved latent variables. As such, these parameters are directly estimable from the data and are reasonably insensitive to misspecification of the assumptions about the latent variable distribution as we have illustrated through bias calculations and two examples. Regression coefficients from the traditional conditional formulation of the multilevel model represent contrasts between the expected response holding both the observed covariates and the unobserved latent variables fixed. While such contrasts can have desirable causal interpretations and are therefore of substantive interest, they may not be directly observable and hence can be model-based extrapolations of the data. As is the case with all extrapolations, they can be highly sensitive to the choice of model as again illustrated by the bias calculations and in the two logistic regression examples.

Often the motivation for including random effects in a multilevel regression model is simply to account for correlation among clustered observations. In discussing the role of statistical models, David Cox (1990) comments:

> It is important to distinguish the parts of the model that define aspects of subject matter interest, the primary aspects, and the secondary aspects that indicate efficient methods of estimation and assessment of precision (page 171).
>
> Especially in empirical models, it is desirable that parameters (e.g., contrasts, regression coefficients and the like) have an interpretation largely independent of the secondary features of the models used (page 173).

Therefore, if the primary objective of analysis is to make inference regarding the mean response as a function of multilevel covariates then a marginalized model may be preferred. Alternatively, if the main scientific interest is in the variance components $\boldsymbol{\alpha}$, then the conditionally specified model may be preferable.

The marginal multilevel model parameterizes the mean regression model the same way that has been done in marginalized log-linear models (Fitzmaurice and Laird, 1993). It has the advantage, however, that the parameterization of $\text{cov}(\mathbf{Y}_i)$ and higher-order moments does not depend on the dimension of $\mathbf{Y}_i$ so that models can be used effectively with data sets such as our two examples in which the number of responses varies across clusters.

The marginal multilevel model has a distinct advantage over other likelihood-based marginal model approaches (e.g., Molenberghs and Lesaffre, 1994; Lang and Agresti, 1994; Heagerty and Zeger, 1996) in that fairly simple latent variable assumption with a parsimonious number of parameters can lead to rich classes of models for the association among observations from the same cluster. Alternative approaches, such as using pairwise log-odds ratios and higher-order contrasts, often require a large number of parameters to account for association patterns.

We estimate the marginal mean regression models discussed here using full maximum likelihood but they can also be estimated using estimating equations (GEE). Hence, this marginal formulation of the multilevel model allows us to separate two distinct issues: (1) whether a marginal or conditional (given latent variables) regression is appropriate to address a particular scientific question; and (2) whether full likelihood or estimating function estimation is favored. Like Liang and Zeger (1986), we have shown through examples that the

mean regression parameters in the marginal multilevel model are reasonably insensitive to the assumed form of the covariance structure. Further work on the relative merits of estimation using a marginal multilevel model likelihood versus an estimating equation approach is warranted.

## APPENDIX

### Calculation of $\Delta(\mathbf{X}_{ijk})$ and Derivatives

Numerical evaluation of the integral in equation (6) can be accomplished with excellent accuracy for a wide range of parameter values using either a general numerical integration method such as Gauss–Hermite quadrature (Abramowitz and Stegun, 1972) or a specialized method such as least maximal approximants (LMA) (Monahan and Stefanski, 1992).

Given $[\eta(\mathbf{X}_{ijk}), \boldsymbol{\alpha}]$, we use Newton–Raphson to solve for the implied conditional parameter $\Delta(\mathbf{X}_{ijk})$. For this we require

$$A_{ijk} = \frac{\partial}{\partial \Delta(\mathbf{X}_{ijk})} h\big[\eta(\mathbf{X}_{ijk})\big] = \int h'\big[\Delta(\mathbf{X}_{ijk}) + b_{ijk}\big]\, dF_{\boldsymbol{\alpha}}(b_{ijk}),$$

which we also obtain numerically.

For maximum likelihood estimation of marginalized models we require derivatives of the deconvolution solution, $\Delta(\mathbf{X}_{ijk})$, with respect to $\eta(\mathbf{X}_{ijk})$ and $\boldsymbol{\alpha}$. Use of the chain rule then yields derivatives with respect to the $\boldsymbol{\beta}^M$. Necessary derivatives can be obtained via implicit differentiation of the convolution equation. Consider the case of Gaussian random effects, $b_{ijk} = \sigma(\mathbf{X}_{ijk})z$, where $z \sim N(0, 1)$. Define

$$B_{ijk} = \int h'_{ijk}\, z\phi(z)\, dz,$$

$$C_{ijk} = \int h''_{ijk}\, \phi(z)\, dz,$$

$$D_{ijk} = \int h''_{ijk}\, z\phi(z)\, dz,$$

$$E_{ijk} = \int h''_{ijk}\, z^2\phi(z)\, dz,$$

where $h_{ijk} = h[\Delta(\mathbf{X}_{ijk}) + \sigma(\mathbf{X}_{ijk})z]$, $h'_{ijk} = \partial h(x)/\partial x$ evaluated at $\Delta(\mathbf{X}_{ijk}) + \sigma(\mathbf{X}_{ijk})z$ and $h''_{ijk} = \partial^2 h(x)/\partial x^2$ evaluated at $\Delta(\mathbf{X}_{ijk}) + \sigma(\mathbf{X}_{ijk})z$.

Using these expressions we can write the required derivatives as

$$\frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \eta(\mathbf{X}_{ijk})} = \frac{\mu'_{ijk}}{A_{ijk}},$$

$$\frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \sigma(\mathbf{X}_{ijk})} = -\frac{B_{ijk}}{A_{ijk}},$$

$$\frac{\partial^2 \Delta(\mathbf{X}_{ijk})}{\partial \eta(\mathbf{X}_{ijk})^2} = \left\{ \mu''_{ijk} - \left[\frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \eta(\mathbf{X}_{ijk})}\right]^2 C_{ijk} \right\} \Big/ A_{ijk},$$

$$\frac{\partial^2 \Delta(\mathbf{X}_{ijk})}{\partial \eta(\mathbf{X}_{ijk})\, \partial \sigma(\mathbf{X}_{ijk})} = -\left[ \frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \eta(\mathbf{X}_{ijk})} \frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \sigma(\mathbf{X}_{ijk})} C_{ijk} + \frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \eta(\mathbf{X}_{ijk})} D_{ijk} \right] \Big/ A_{ijk},$$

$$\frac{\partial^2 \Delta(\mathbf{X}_{ijk})}{\partial \sigma(\mathbf{X}_{ijk})^2} = -\left\{ \left[\frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \sigma(\mathbf{X}_{ijk})}\right]^2 C_{ijk} + 2\frac{\partial \Delta(\mathbf{X}_{ijk})}{\partial \sigma(\mathbf{X}_{ijk})} D_{ijk} + E_{ijk} \right\} \Big/ A_{ijk},$$

where $\mu'_{ijk}$ and $\mu''_{ijk}$ are the derivatives of $h(x)$ evaluated at $\eta(\mathbf{X}_{ijk})$.

## ACKNOWLEDGMENTS

We are grateful to the Multilevel Models Project, Institute of Education, University of London for making the British Social Survey data publicly available on their Web site. Funding for this research was supported by NIH Grants R01 MH56639, P30 MH38725 and P01 CA76466-01.

## REFERENCES

ABRAMOWITZ, K. M. and STEGUN, I. A. (1972). *Handbook of Mathematical Functions*. Dover, New York.

AERTS, M. and CLAESKENS, G. (1997). Local polynomial estimation in multiparameter likelihood models. *J. Amer. Statist. Assoc.* **92** 1536–1545.

AZZALINI, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* **81** 767–775.

BISHOP, Y., FEINBERG, S. and HOLLAND, P. (1975). *Discrete Multivariate Analysis*. MIT Press.

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. Roy. Statist. Soc. Ser. B* **61** 265–285.

BRESLOW, N. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

CAREY, V. C., ZEGER, S. L. and DIGGLE, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80** 517–526.

CONAWAY, M. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *J. Amer. Statist. Assoc.* **84** 53–62.

COX, D. R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5** 169–174.

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

DEMING, W. E. and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11** 427–444.

DIGGLE, P. J. (1988). An approach to the analysis of repeated measures. *Biometrics* **44** 959–971.

DIGGLE, P. J., LIANG, K.-Y. and ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford Univ. Press.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

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.

GILKS, W., RICHARDSON, S. and SPIEGELHALTER, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

GLONEK, G. F. V. and MCCULLAGH, P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B* **57** 533–546.

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31** 1208–1212.

GOLDSTEIN, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika* **78** 45–51.

GOLDSTEIN, H. (1995a). *Multilevel Statistical Models*. Arnold, London.

GOURIEROUX, C., MONFORT, A. and TROGNON, A (1984). Pseudo-maximum likelihood methods: theory. *Econometrica* **52** 681–700.

GRAUBARD, B. I. and KORN, E. L. (1994). Regression analysis with clustered data. *Statistics in Medicine* **13** 509–522.

HEAGERTY, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55** 688–698.

HEAGERTY, P. J. and ZEGER, S. L. (1996). Marginal regression models for clustered ordinal measurements. *J. Amer. Statist. Assoc.* **91** 1024–1036.

HEDEKER, D. and GIBBONS, R. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50** 933–944.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970.

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

LIANG, K.-Y. and HANFELT, J. (1994). On the use of the quasi-likelihood method in teratological experiments. *Biometrics* **50** 872–880.

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

LIPSITZ, S., LAIRD, N. and HARRINGTON, D. (1991). Generalized estimating equations for correlated binary data: using odds ratios as a measure of association. *Biometrika* **78** 153–160.

MACDONALD, I. L. and ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**, 162–170.

MCGRATH, K. and WATERTON, J. (1986). British social attitudes, 1983–1986, panel survey. Technical report, London Social and Community Planning Research.

MOLENBERGHS, G. and LESAFFRE, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.* **89** 633–644.

MONAHAN, J. F. and STEFANSKI, L. A. (1992). Normal scale mixture approximations to $F^*(x)$ and computation of the logistic-normal integral. In *Handbook of the Logistic Distribution* (N. Balakrishnan, ed.) 529–540. Dekker, New York.

NEUHAUS, J. M., HAUCK, W. W. and KALBFLEISCH, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79** 755–762.

NEUHAUS, J. M. and JEWELL, N. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* **80** 807–815.

NEUHAUS, J. M. and KALBFLEISCH, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54** 638–645.

NEUHAUS, J. M., KALBFLEISCH, J. D. and HAUCK, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Internat. Statist. Rev.* **59** 25–35.

PENDERGAST, J. F., GANGE, S. J., NEWTON, M. A., LINDSTROM, M. J., PALTA, M. and FISHER, M. R. (1996). A survey of methods for analyzing clustered binary response data. *Internat. Statist. Rev.* **64** 89–118.

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

ROTNITZKY, A. and WYPIJ, D. (1994). A note on the bias of estimators with missing data. *Biometrics* **50** 1163–1170.

SHEN, W. and LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. Roy. Statist. Soc. Ser. B* **60** 455–471.

STIRATELLI, R., LAIRD, N. and WARE, J. (1984). Random effects models for serial observations with binary responses. *Biometrics* **40** 961–970.

WEDDERBURN, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61** 439–447.

WEIL, C. S. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology* **8** 177–182.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25.

WIGGENS, R. D., ASHWORTH, K., O'MUIRCHEARTAIGH, C. A. and GALBRAITH, J. I. (1990). Multilevel analysis of attitudes to abortion. *The Statistician* **40** 225–234.

WOLFINGER, R. and O'CONNELL, M. (1993). Generalized linear mixed models: a pseudolikelihood approach. *J. Statist. Comput. Simulation* **48** 233–243.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

ZEGER, S. L., LIANG, K.-Y. and ALBERT, P. A. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44** 1049–1060.

# Comment

**Emmanuel Lesaffre and Bart Spiessens**

The authors have combined the best of two worlds. Their approach allows a marginal interpretation of the parameters combined with a computational flexibility. On top of this, the approach also permits the estimation of subject-specific parameters. Therefore, this paper is an important contribution to the development of statistical models which allow an easy interpretation of the so-called fixed-effects parameters and at the same time not suffering from too many practical restrictions on the dimensions of the problem.

It is recognized that the focus of the paper was on establishing a model which allows maximum interpretability and practical flexibility. However, the authors give the impression that it does not matter much whether linearity in the covariates is assumed on the marginal or the conditional level. Is not checking for goodness-of-fit an essential part in applied statistics? In their examples no mention is made on which level, conditional or marginal, it is best to assume linearity in the covariates. In this respect, we are also interested to know what the functions $\Delta(X_{ijk})$ look like when linearity is assumed on the marginal level.

Further, the authors give the impression that they are offering a free lunch. With computing capabilities that rapidly increase, it may not seem to be too important that procedures should be fast in practice. The multivariate Dale model (Molenberghs and Lesaffre, 1994) is a natural extension of the logistic regression model in the case of repeated ordinal data. Yet, despite its elegance, problems of dimensions higher than 5 cannot be tackled in a reasonable amount of time at this moment. We wonder whether similar problems occur for the marginalized conditional approach when more than two random effects need to be assumed. Further, we would welcome some details on how stable their computational procedure is when solving the integral equations. Even for so-called robust procedures, such as the Gauss–Hermite method, one needs to be cautious in simple situations. Indeed, we recently came across a logistic random-effects model which, when applied to a dermatological clinical trial, (De Backer et al., 1998) yielded a highly significant treatment effect at baseline although the trial was randomized. A nonsignificant treatment effect was obtained with a marginal model. This analysis was done with MIXOR (Hedeker and Gibbons, 1996) and with an own-written GAUSS program using 10 quadrature points. The nonsignificant result disappeared only when the number of quadrature points approached 30, much more than what is generally recommended. Furthermore, the treatment effect kept on changing when increasing the number of quadrature points further.

Furthermore, it would be useful to warn the user that the random-effects presentation of association between measurements is just a paradigm. At the end of the day, the user is just maximizing a marginalized model, which may or may not correspond to an underlying hierarchical model. This is already true in the most simple case of a linear mixed model (see Lesaffre, Verbeke and Kenward, 2000).

We would like to conclude by congratulating the authors on a very nice and elegant paper that will certainly have an important impact on the way we will analyze future studies with correlated data.

*Emmanuel Lesaffre is Professor; Bart Spiessens is research assistant, Faculty of Medicine, Biostatistical Centre K. U. Leuven, U. Z. St. Rafael, Kapucijnenvoer 35, B-3000 Leuven, Belgium.*

# Comment

**John M. Neuhaus**

This paper provides data analysts with effective new likelihood-based methods for analyzing clustered and longitudinal data and reemphasizes the importance of using statistical methods that measure covariate effects of scientific interest. Heagerty and Zeger provide valuable contributions to fitting marginal models to clustered and longitudinal data. However, I believe that the authors's approach for estimating conditional model parameters does not outperform existing approaches. The authors's approach is best suited to provide likelihood-based marginal or population-averaged analysis of clustered and longitudinal data as an alternative to the estimating equations-based estimators and methods that approaches such as GEE1 and GEE2 provide. The authors's approach for marginal model parameters enjoys all the advantages of likelihood-based methods. For example, one can test hypotheses using likelihood ratio procedures, construct likelihood-based confidence intervals and validly apply the approach when data are missing at random.

It would be interesting to compare the authors's likelihood approach for marginal models to Goldstein's (1995a) first-order marginal quasilikelihood (MQL) approach for multilevel models. In particular, it would be interesting to compare the estimates from MQL fits to the teratology and British Social Survey data to the authors's marginal model estimates. The MQL approach approximates the mixed model integral in (6) using a Taylor series expansion about no random effects (i.e., $b = 0$). As Neuhaus and Segal (1997) point out, the first-order MQL approach estimates the parameters of marginal models fitted to clustered and longitudinal data rather than those of conditional models. The MQL approach involves an approximate likelihood that one can use to carry out any likelihood-based procedure for a flexible class of dependence structures based on random effects. Widely available commercial software fits the MQL approach.

Since the conditionally specified model also determines the marginal mean, one can use deconvolution methods to estimate the parameters of the conditional model. However, this will involve more work than simply fitting the conditional model. Methods to fit single random effect conditional models extend easily to cases where the random effects distribution depends on covariates and to a few nested random effects. In general, I do not find the objective of finding a single model that can provide estimates of both population-averaged and conditional covariate effects particularly compelling. One should simply fit a model, or set of models, that measures covariate effects of scientific interest.

The authors's analyses of the two data sets complement the investigation of Neuhaus, Hauck and Kalbfleisch (1992) of the effects of mixing distribution misspecification and point out additional problems with the conditional model parameters associated with cluster-level covariates. When the variability of the random effects depends on a covariate, one would expect that inference about that covariate will depend on the correct specification of the random effects model. Indeed, different assumptions about dependence of random effects on covariates lead to noticeably different parameter estimates of both the marginal and conditional models. The conditional model estimates are more sensitive to assumptions about random effects dependence than are marginal estimates but conditional treatment effects in Table 1 and conditional level-2 covariate effects in Table 3 are not of scientific interest. As the authors point out, conditional models measure change in the expected value of the response associated with a unit increase in a covariate among observations with the same random effect (e.g., within clusters). Such changes never occurred in the teratology data set and it makes little sense to fit a conditional model in this case to produce an extrapolation to what the treatment effect might have been had the same experimental unit received both treatments. Conditional model estimates of the effects of cluster-level covariates may be sensitive to misspecification of the dependence of random effects on cluster-level covariates but this is a nonissue since one should not report conditional estimates for such covariates in the first place. A similar argument applies to the level-2 covariates in the British Social Survey data.

*John M. Neuhaus is Professor, Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94143-0560* (*e-mail: john@biostat.ucsf.edu*).

While the authors present two data sets that exhibit random effects distributions that differ by the levels of a covariate, further investigation into the amount and kind of data one would need to detect such differences seems worthwhile. Semiparametric mixture models (e.g., Lindsay, 1995) yield nonparametric estimates of the mixing distribution and provide a measure of the amount of information a data set contains about this distribution. Fitting such models to binary data typically results in nonparametric distribution estimates with very few points of support, indicating that the data contain little information to distinguish between competing mixing distributions. Allowing the mixing distribution to depend on covariates is analogous to allowing interaction between covariates and random effects. Since data requirements to detect significant interaction are often large, one would expect that one would need very large data sets to detect dependence of random effects distributions on covariates.

The most striking feature of Table 3 is the insensitivity of the conditional estimates corresponding to the year effects to different assumptions about random effects. The year effects are purely within-person covariates and are exactly the kinds of effects conditional models are designed to estimate. Conditional models are best suited to estimating the association of within-cluster changes in a covariate with within-cluster changes in the outcome. Such associations are often of central scientific interest in studies with covariates that vary within clusters and/or over time. Purely within-cluster covariates are orthogonal to all covariates that are constant within clusters, including random effects. Thus, conditional models for purely within-person covariates separate models for covariate effects from models for response dependence, as does the authors's approach for marginal models. A conditional likelihood analysis of the year effects would yield estimates and standard errors very similar to those in Table 3 and would not involve the specification of a random effects distribution. When sufficient statistics for the random effects exist, as with generalized linear mixed models with a canonical link function, the conditional likelihood approach provides effective, intuitively appealing estimates of the effects of within-cluster changes in covariates. With noncanonical link functions one can effect an analogous analysis by decomposing covariates into within- and between-cluster components and fitting a mixed effects model with such covariates to changes in the outcome, as in Neuhaus and Kalbfleisch (1998).

In summary, the authors provide an effective, new likelihood-based approach for fitting marginal models to clustered and longitudinal data. However, existing mixed effects models and conditional likelihood approaches provide more straightforward, computationally efficient estimates of the associations of within-cluster covariate changes with changes in the outcome than the authors's conditional approach.

# Comment

**Stephen W. Raudenbush**

This article is a major contribution to statistical methods for multilevel data and, in particular, for generalized linear models with nested random effects. It provides a single, likelihood-based approach to inference about conditional regression models (which assess associations between $X$ and $Y$ holding constant the random effects) and marginal models (which assess associations between $X$ and $Y$ based on averaging over the random effects). Until now, the two kinds of regression models have been artificially separated by limitations in available methods of statistical estimation. In particular, maximum likelihood (ML) was available for conditional but not marginal regression models.

Moreover, likelihood-based methods also produce-empirical Bayes estimates of random effects. These estimates are useful in many applications including small-area estimation (cf. Morris, 1983; Tsutakawa, 1988). The fact that these have been available only within the conditional model tended to reinforce the notion that the conditional model provides a richer summary of evidence than does the marginal model.

*Stephen W. Raudenbush is Professor, School of Education, University of Michigan, 610 East University, Ann Arbor, Michigan 48109.*

The current article enables the scientist to simultaneously estimate the marginal and conditional mean structure, providing also empirical Bayes summaries for random effects. Of course, the transformed marginal and conditional mean cannot both be linear in the covariates. But the approach leaves that choice to the analyst. The availability of likelihood inference for both models places this choice on a more principled footing.

A key question is this: Should the choice between the marginal and conditional linear regression models be based on purely statistical grounds, such as robustness? Or, should that choice be prescribed entirely by the substantive aims of the study? The authors show that "regression parameters in conditionally specified models are more sensitive to random effects assumptions than their counterparts in the marginal formulation." Because these assumptions are hard to check, the argument is that the marginal model is often or even typically preferable to the conditional model. Thus, it appears that the choice between the two linear models can often be made on statistical grounds.

There is a logical problem, however. If conditional regression results vary as a function of assumptions, at least some of them must also differ from marginal results regarding the apparent association between $X$ and $Y$. That the marginal answer is robust does not make it a better answer unless the scientific question truly requires a marginal inference. Thus, some systematic approach is needed to classify the inferential goals in multilevel analyses to guide choice between conditional and marginal inference for regression coefficients.

To develop such a classification, I draw upon the seminal work of Lindley and Smith (1972), whose hierarchical construction of the multilevel model pinpoints possible targets of inference in a way that my colleagues and I have found extremely informative (cf. Raudenbush, 1988; Bryk and Raudenbush, 1992).

## LEVEL 1

The first level of the model describes the association between the observed data $Y_i$, $i = 1, \ldots, n$, and unobservable random quantities $\theta_i = \theta(b_i)$, referred to in Bayesian language as parameters (cf. Lindley and Smith, 1972), but known elsewhere as latent variables or random coefficients (cf. Longford, 1993). Often, $\theta_i = \Delta_i + Z_i b_i$, in the language of the current article, where $b_i$ is a random effect and $Z_i$ is the random effects design matrix. We thus have $f(Y_i \mid \theta_i)$, $i = 1, \ldots, n$, as the "level-1" model. For example, $Y_i$ might be a vector of repeated measurements on subject $i$, while the elements of $\theta_i$ are individual growth parameters (Laird and Ware, 1982). Alternatively $\theta_i$ might be an effect size estimated by $Y_i$ in each of $n$ experiments (cf. Raudenbush and Bryk, 1985; Morris and Normand, 1992). Other possibilities include the case where $Y_i$ is regressed on $X_i$ in each of $n$ schools and $\theta_i$ is a vector of school-specific regression coefficients. In some cases, $Y_i$ are the observed data and $\theta_i$ are the "complete" data for subject $i$ (cf. Goldstein, 1995b, Chapter 4). Or $Y_i$ might be the fallible data and $\theta_i$ are the "true" values of the data (cf. Raudenbush and Sampson, 1999a) so that $f(Y_i \mid \theta_i)$ is a measurement model (Fuller, 1987).

## LEVEL 2

The second level of the model specifies, in Bayesian terms, an exchangeable prior distribution for $\theta_i$. We thus have $p(\theta_i \mid \beta, D)$, where $\beta$ are regression coefficients and $D$ is the variance–covariance matrix of $b_i$. This second-level distribution might also be called the distribution of the latent variables or the random coefficients. Here $p(\theta_i \mid \beta, D)$ specifies the distribution of growth parameters across a population of repeatedly observed subjects, the distribution of effect sizes across replicated experiments, the distribution of school-specific regression coefficients across a population of schools or the distribution of the complete data (or "true scores") across survey respondents.

## MARGINALIZATION

The two levels of the model then lead to the marginalization

$$q(Y \mid \beta, D) = \int \prod_{i=1}^{n} f(Y_i \mid \theta_i) p(\theta_i \mid \beta, D)\, d\theta. \quad (1)$$

Prior to publication of the current article, $\beta = \beta^C$ involved the "conditional" linear regression parameters for the relevant nonlinear link function. The current article creates the option of parameterizing $\beta = \beta^M$, that is, the marginal linear regression model.

## TARGETS OF INFERENCE

The first target of inference is $q(Y \mid \beta, D)$, which governs the association between $X$ and $Y$. Based on the current article, likelihood inference is now available for conditional and marginal parameterizations of $\beta$. These focus on the $X$–$Y$ relationship. This target of inference represents the scenario in the introduction to the article. The subject-matter consideration is whether one wants to hold constant the random effect in assessing the $X$–$Y$ asso-

ciation. If the appeal of each of these options seems similar, one might opt for marginal inference on robustness considerations, although this decision does not really avoid the logical problem mentioned above.

The second target of inference is $f(Y_i \mid \theta_i)$, for $\pi_i$, $i = 1, \ldots, n$. One seeks an inference about the growth parameters of a particular subject, the effect size of a particular study or the regression coefficients in a particular school. Empirical Bayes methods, based on $p(\theta \mid Y, \beta = \hat{\beta}, D = \hat{D})$ provide a reasonable approach for estimating this distribution if the number of clusters is large. Such inference had been available within the framework of the conditional regression interpretation. The current article allows such inference for both marginal and conditional parameterizations. The choice of marginal versus conditional version of $\beta$ here should presumably not have an appreciable effect on inference.

The third target of inference is the distribution of the latent variables themselves, that is, $p(\theta_i \mid \beta, D)$. This would clearly be the target in a missing data model. In that case one attempts to estimate the parameters of the complete-data distribution from the incomplete data, based on the assumption the data are missing at random (Little and Rubin, 1987). Another case involves a measurement model, standard in education, wherein $Y_i$ is a vector of binary responses, each indicating a correct or incorrect response to a question on a test, and $\theta_i$ is the ability of examinee $i$. Here $\beta$ describes the associations between covariates, $X$, and student ability, while $D$ might describe the dispersion of ability in the population of students (cf. Bock, 1989). Note that $\beta$ and $D$ contain information about the regression of one element of $\theta_i$ conditional on other elements (cf. Lillard and Farmer, 1998; Raudenbush and Sampson, 1999b). It would appear that, for this third target of inference, the conditional model must be the model of choice. After all, it is the distribution of the latent data that is of interest, not the distribution of $Y$.

In sum, for target (1) there is a choice and it may make a difference; for (2) there is a choice that should not make a difference; for (3) the conditional model is the only choice that fits conceptually.

As the authors point out, there is a serious potential problem with conditional inference in the case of these nonlinear link functions. Results may be sensitive to distributional assumptions that are hard to check. When marginal inference is not a viable choice conceptually, the remaining strategy appears to be sensitivity analysis. Inferences about $\beta^C$ and $D$ must be checked against alternative plausible assumptions about distributional family and the structure of variation at each level (cf. Seltzer, 1993).

# Rejoinder

**Patrick J. Heagerty and Scott L. Zeger**

We agree with Stephen Raudenbush that marginal and conditional multilevel regression models have been "artificially separated" by association with different methods of estimation. Marginal models are commonly fitted using estimating equations while conditional models are typically associated with likelihood and Bayesian estimation methods. In this article we have attempted to carefully separate the form of the regression model from procedures used to estimate model parameters. Our focus on the use of a marginally specified mean model with a likelihood construction traditionally used for conditionally specified models allows an unconfounded focus on the difference between the marginal and conditional regression models. Likelihood estimation in nonlinear hierarchical models remains challenging in many situations and our use of numerical quadrature is intended to illustrate the feasibility of our approach rather than be the definitive computational method of estimation. We are grateful to Emmanuel Lesaffre, Bart Spiessens, John Neuhaus and Stephen Raudenbush for their thoughtful commentary which helps to illuminate important issues relating to the choice of a regression model and issues pertaining to estimation and robustness. First we comment on statistical models and then discuss parameter estimation.

## MODELS

Raudenbush addresses the fundamental question of when to consider use of a marginal versus conditional multilevel regression model and provides a

clear and comprehensive classification driven by the scientific aims of analysis. We completely agree that the substantive goals should dictate the choice of regression model. Our demonstration of the sensitivity of certain parameter estimates is intended to emphasize that conditional regression estimates for between-cluster covariates suffer both in their interpretation and in their estimation; they are essentially extrapolations. In fact, Neuhaus goes so far as to say that "one should not report conditional estimates for such covariates in the first place." Additional discussions of the appropriate domain of application for marginal and conditional models have been given in Graubard and Korn (1994) and Neuhaus, Kalbfleisch and Hauck (1991). More recently, related issues of collapsibility and causal inference have been discussed in Greenland, Robins and Pearl (1999). We hope this article, and the thoughtful discussion, helps clarify the distinction between the marginal and the conditional multilevel regression formulation so that data analysts can make an informed choice that satisfies their specific scientific objectives.

Lesaffre and Spiessens (L&S) remind us that assessing model fit is also an important part of any data analysis. Whether a variable can be modeled linearly on either the marginal or conditional scale can be assessed empirically, and deviations from linearity can be accommodated by generalizing the regression form. We make no assertions as to where it is "best" to assume linearity since this decision should be guided by both substantive consideration and empirical evaluation. Although it is true that linearly may not hold for both marginal and conditional linear predictors (depending on the link function), our experience is that the difference in the structure of the regression model is frequently small. In Table R1 we present the saturated two-level conditional regression structure $\Delta(\mathbf{X}_{ij}) = \beta_0^C + \beta_1^C X_{1,ij} + \beta_2^C X_{2,i} + \beta_3^C X_{1,ij} X_{2,i}$ for different values of $\sigma$ that induces an additive marginal logistic model $\eta(\mathbf{X}_{ij}) = \beta_0^M + \beta_1^M X_{1,ij} + \beta_2^M X_{2,I}$ with binary level-1 covariate $X_{1,ij} = (0, 1)$ and binary level-2 covariate $X_{2,i} = 0$ or 1. We find that the conditional model is approximately additive unless $\sigma$ is large. Although the marginal and conditional covariate structures are similar, the magnitude of conditional regression coefficients does depend on the value of the heterogeneity parameter. Additional complexity results when the heterogeneity parameters depend on covariates.

TABLE R1
*Conditional mean parameters $\boldsymbol{\beta}^C$ as a function of $\sigma$ for a fixed marginal $\boldsymbol{\beta}^M = (-1.00, 0.50, 0.25, 0.00)$*

| $\sigma$ | $\beta_0^C$ | $\beta_1^C$ | $\beta_2^C$ | $\beta_3^C$ |
|---|---|---|---|---|
| 0.0 | −1.00 | 0.50 | 0.25 | 0.00 |
| 0.5 | −1.06 | 0.53 | 0.26 | 0.00 |
| 1.0 | −1.20 | 0.60 | 0.30 | 0.00 |
| 2.0 | −1.63 | 0.80 | 0.40 | 0.01 |
| 3.0 | −2.12 | 1.02 | 0.50 | 0.04 |

TABLE R2
*MQL and QEE parameter estimates for the teratology data*

| | MQL | | QEE | |
|---|---|---|---|---|
| | Estimate | s.e. | Estimate | s.e. |
| *Marginal mean* $\boldsymbol{\beta}^M$ | | | | |
| Intercept | 2.174 | 0.293 | 2.176 | 0.284 |
| Treatment | −1.058 | 0.486 | −1.072 | 0.470 |
| *Level 2 heterogeneity* $\sigma_2$ | | | | |
| Intercept | 0.511 | 0.500 | 0.428 | 0.380 |
| Treatment | 0.827 | 0.600 | 1.464 | 0.722 |

Finally, Neuhaus criticizes the concept of a single model that attempts to do all things. We make no claim that our model is universally applicable. Different tasks require different tools.

## ESTIMATION

Neuhaus inquires about approximate maximum likelihood procedures that may offer computationally simple parameter estimation, such as marginal quasilikelihood (MQL). In Table R2 we have used two approximate methods to fit marginalized GLMMs to the teratology data analyzed in Section 5.1. The first estimation method is MQL as described in Breslow and Clayton (1993). The second method is based on quadratic estimating equations (QEE) and is described in Heagerty (1999). Both methods provide marginal regression estimates that are similar to the MLE presented in Table 1 (Model 3). However, MQL estimates of heterogeneity parameters appear negatively biased (for the treatment group); MQL estimates of variance components are generally not consistent and the negative bias observed in our teratology example is in accord with simulation results reported in Breslow and Clayton (1993). The QEE method yields consistent estimates of mean parameters and variance components since estimates result as the solution to a pair of unbiased estimating equations. More work is warranted in the development and evaluation of consistent approximate ML methods for multilevel categorical data.

Lesaffre and Spiessens discuss an example in which a large number of quadrature points was required to accurately evaluate the marginal likelihood function. Simple Gaussian quadrature that evaluates the likelihood with a small number of

nodes centered at the mean of the random effects has been shown to be potentially inaccurate (Pinheiro and Bates, 1995). Either a larger number of quadrature points can be used or the quadrature can be modified to center at the conditional mode of the random effects. In our examples we used $K = 20$ and $K = 50$ quadrature points. Quadrature methods become computationally impractical with a moderate or high dimensional random effects distribution. However, choice between the marginalized and the conditional regression models is not impacted by the dimension of the random effects since the transformation from the marginal linear predictor $\eta(\mathbf{X}_i)$ to the conditional linear predictor $\Delta(\mathbf{X}_i)$ only requires solution of an integral equation in one dimension (over the marginal distribution of $b_{ij}$) regardless of the dimensionality of $\mathbf{b}_i = \text{vec}(b_{ij})$. Improving algorithms for likelihood estimation of GLMMs remains an active research area and advances in fitting conditionally specified models can be expected to transfer to marginally specified models.

Finally, we agree with Neuhaus that conditional logistic regression offers a simple and attractive method of estimation for within-cluster covariates. Two main limitations are that no summaries of between-cluster systematic or random variation are obtained, and the basic model may be too simplistic (random intercepts only) for certain settings such as with longitudinal data.

## SUMMARY

Regression analysis of multilevel categorical data requires selection of a mean model, a dependence model and a method of estimation. In this manuscript we have decoupled the choice of mean and dependence models, permitting greater flexibility in the choice of statistical approaches.

## ADDITIONAL REFERENCES


DE BACKER, M., DE VROEY, C., LESAFFRE, E., SCHEYS, I. and DE KEYSER, P. (1998). Twelve weeks of continuous onychomycosis caused by dermatophytes: a double blind comparative trial of terbafine 250 mg/day versus itraconazole 200 mg/day. *J. Amer. Acad. Dermatology* **38** 5, 3, S57–S63.

HEDEKER, D., and GIBBONS, R. D. (1996). MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* **49** 157–176.

LESAFFRE, E., VERBEKE, G. and KENWARD, M. (2000). On the two-stage interpretation of the linear mixed model. Unpublished manuscript (in preparation).

LINDSAY, B. (1995). *Mixture Models: Theory, Geometry and Applications* IMS, Hayward, CA.

NEUHAUS, J. M. and SEGAL, M. R. (1997). An assessment of approximate maximum likelihood estimators in generalized linear mixed models. In *Modelling Longitudinal and Spatially Correlated Data* (T. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. Warren and R. D. Wolfinger, eds.) 11–22. Springer, New York.

BOCK, R. D. (1989). *Multilevel Analysis of Educational Data*. Academic Press, New York.

BRYK, A. and RAUDENBUSH, S. (1992). *Hierarchical Linear Models for Social and Behavioral Research: Applications and Data Analysis Methods*. Sage, Newbury Park, CA.

FULLER, W. A. (1987). *Measurement Error Models*. Wiley, New York.

GOLDSTEIN, H. (1995b). *Multilevel Statistical Models*, 2nd ed. Wiley, New York.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrika* **65** 581–590.

LILLARD, L. A. and FARMER, M. M. (1998). Functional limitations, disability and perceived health of the oldest old: an examination of health status in AHEAD. Paper presented to the Survey Research Center, Univ. Michigan.

LINDLEY, D. and SMITH, A. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

LITTLE, R. and RUBIN, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

LONGFORD, N. (1993). *Random Coefficient Models*. Clarendon, Oxford.

MORRIS, C. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65.

MORRIS, C. and NORMAND, S. (1992). Hierarchical models for combining information and for meta-analysis. In *Bayesian Statistics 4* (J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith, eds.) 321–344. Oxford Univ. Press.

RAUDENBUSH, S. (1988). Educational applications of hierarchical linear models: a review. *J. Educational Statist.* **13**(20) 85–116.

RAUDENBUSH, S. and BRYK, A. (1985). Empirical Bayes meta-analysis. *J. Educational Statist.* **10** 75–98.

RAUDENBUSH, S. W. and SAMPSON, R. (1999a). Ecometrics: toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology* **29** 1–41.

RAUDENBUSH, S. W. and SAMPSON, R. (1999b). Assessing direct and indirect associations in multilevel designs with latent variables. *Sociological Methods and Research* **28** 123–153.

SELTZER, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach. *J. Educ. Statist.* **18** 207–235.

TSUTAKAWA, R. (1988). Mixed model for studying geographic variability in mortality rates. *J. Amer. Statist. Assoc.* **83** 37–42.

PINHEIRO, J. C. and BATES, D. M. (1995). Approximations to the log-likelihood function in the non-linear mixed-effects model. *J. Comput. Graph. Statist.* **4** 12–35.

GREENLAND, S., ROBINS, J. M. and PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.* **14** 29–46.