J. R. Statist. Soc. A (2001)
164, *Part* 2, *pp.* 339–355

Improved estimation procedures for multilevel models with binary response: a case-study

Germán Rodríguez and Noreen Goldman

Princeton University, USA

[Received August 1997. Revised September 2000]

Summary. During recent years, analysts have been relying on approximate methods of inference to estimate multilevel models for binary or count data. In an earlier study of random-intercept models for binary outcomes we used simulated data to demonstrate that one such approximation, known as marginal quasi-likelihood, leads to a substantial attenuation bias in the estimates of both fixed and random effects whenever the random effects are non-trivial. In this paper, we fit three-level random-intercept models to actual data for two binary outcomes, to assess whether refined approximation procedures, namely penalized quasi-likelihood and second-order improvements to marginal and penalized quasi-likelihood, also underestimate the underlying parameters. The extent of the bias is assessed by two standards of comparison: exact maximum likelihood estimates, based on a Gauss–Hermite numerical quadrature procedure, and a set of Bayesian estimates, obtained from Gibbs sampling with diffuse priors. We also examine the effectiveness of a parametric bootstrap procedure for reducing the bias. The results indicate that second-order penalized quasi-likelihood estimates provide a considerable improvement over the other approximations, but all the methods of approximate inference result in a substantial underestimation of the fixed and random effects when the random effects are sizable. We also find that the parametric bootstrap method can eliminate the bias but is computationally very intensive.

Keywords: Gibbs sampling; Immunization; Marginal quasi-likelihood; Multilevel logit models; Parametric bootstrap; Penalized quasi-likelihood; Prenatal care

1. Introduction

There is a strong and growing interest among social scientists in understanding the influence of the social context, such as families and communities, on a wide range of behaviours. This trend has renewed interest in multilevel statistical models. At the same time, there is an increased recognition of the need to account for clustering in the complex sample designs that are used in virtually all social and health surveys. Statistical procedures that ignore clustering tend to underestimate the variance of the estimated coefficients and can lead to the mistaken identification of 'statistically significant' effects. In addition, in models such as logistic regression, where the relationship between the response and the predictors is not linear, ignoring clustering can result in large biases in the parameter estimates themselves, as we demonstrate later.

In the last decade the estimation of multilevel models has received considerably more attention than in the past, and various computer programs have been developed for fitting these models to hierarchically clustered data. In the case of multilevel models for normally distributed outcomes, maximum likelihood procedures are now readily available (see for example Rasbash and Woodhouse (1995)). However, in the case of binary responses or count

Address for correspondence: Germán Rodríguez, Office of Population Research, Wallace Hall, Princeton University, Princeton, NJ 08544-2091, USA.
E-mail: grodri@princeton.edu

 0964–1998/01/164339

data, maximum likelihood procedures have proved to be intractable for all except the simplest type of multilevel models (such as variance component models) because they involve irreducibly high dimensional integrals (Breslow and Clayton, 1993). As a consequence, most analysts have relied on approximate methods of inference, most notably marginal quasi-likelihood (MQL) (Goldstein, 1991) and penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993; Schall, 1991).

In earlier work we demonstrated that estimates derived by MQL for binary outcomes can be subject to substantial bias when the amount of clustering is 'sufficiently large to be interesting' (Rodríguez and Goldman, 1995). The magnitude of the bias was assessed by using simulated data designed to incorporate the hierarchical structure of actual or plausible data sets and estimating a three-level model with two versions of MQL, involving first-order and second-order approximations. In response to our study, Goldstein and Rasbash (1996) evaluated estimates based on PQL and introduced an improved approximation, known as second-order PQL, that 'largely eliminates the biases' in the situation that we had previously described. Key results from these studies are summarized later.

In this paper we use actual data from a survey in Guatemala to assess the performance of both the original and the more recently developed approximations. We focus on two binary outcomes that exhibit moderate and large amounts of clustering at the family and community levels:

(a) obtaining a complete set of immunizations for children who have received at least one immunization and
(b) using modern prenatal care for pregnancies where some form of care was used (Pebley *et al.*, 1996).

We fit three-level variance component models and compare estimates based on first- and second-order MQL and PQL with two standards: exact maximum likelihood estimates obtained by using a Gauss–Hermite numerical quadrature procedure and a set of Bayesian estimates obtained by using Gibbs sampling with diffuse priors (Zeger and Karim, 1991). We also use a parametric bootstrap procedure to reduce the bias of MQL and PQL estimates; this method was first proposed by Kuk (1995) and has been applied successfully in the context of two-level models by Goldstein (1996).

In Section 2 we describe the model and the various estimation procedures. In Section 3 we summarize key results based on our simulated data sets, whereas in Section 4 we present results based on the actual Guatemalan data. Finally we review and discuss our conclusions in Section 5. The simulated and actual data can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. The model and estimation procedures

In this paper we consider a simple multilevel extension of the ordinary logistic model: a random-intercept (also known as a variance component) model that incorporates random effects at two hierarchical levels other than the individual, in our case the family and the community. Let Y_{ijk} denote the response of the ith child (pregnancy) of the jth family (mother) in the kth community. We assume that, given random effects U_{jk} and U_k representing unobserved family and community characteristics respectively, the Y_{ijk} are independent Bernoulli random variables with (conditional) expectation π_{ijk}. We further assume that the logit of this probability satisfies

$$\text{logit}(\pi_{ijk}) = \beta_0 + \boldsymbol{\beta}_1 \mathbf{x}_{ijk} + \boldsymbol{\beta}_2 \mathbf{x}_{jk} + \boldsymbol{\beta}_3 \mathbf{x}_k + U_{jk} + U_k, \tag{1}$$

where $\mathbf{x}_{ijk}$, $\mathbf{x}_{jk}$ and $\mathbf{x}_k$ represent (vectors of) observed characteristics at the individual, family and community levels, with corresponding fixed effects $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$. For estimation purposes we further assume that the random effects are independent and normally distributed, with

$$\begin{aligned} U_{jk} &\sim N(0, \sigma_2^2), \\ U_k &\sim N(0, \sigma_3^2). \end{aligned} \tag{2}$$

This model is shown as a directed acyclical graph in Fig. 1, which follows the conventions in Spiegelhalter *et al.* (1996). Each variable or parameter in the model appears as a node, with boxes denoting known quantities and ovals denoting unknown quantities. Full arrows denote probabilistic dependences whereas broken arrows indicate deterministic relationships. The nested structure of children (or pregnancies) within families within communities is shown by using stacked sheets.

In the developments below we shall let $\mathbf{Y}$ denote the vector of responses and $[\mathbf{Y}]$ its distribution. Similarly, $\mathbf{U}$ is the vector of family and community random effects, $\boldsymbol{\beta}$ is the vector of fixed effects and, with a slight abuse of notation, $\boldsymbol{\sigma}^2$ is a vector containing the variance components, or variances of the random effects at the family and community levels. With this notation, model (1) can be written as a special case of the generalized linear mixed model, with

$$\text{logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U}, \tag{3}$$

where $\mathbf{X}$ is the model matrix for the fixed effects (containing the constant and the observed covariates) and $\mathbf{Z}$ is the model matrix for the random effects (containing 1s and 0s to select the appropriate random effects corresponding to each observation).

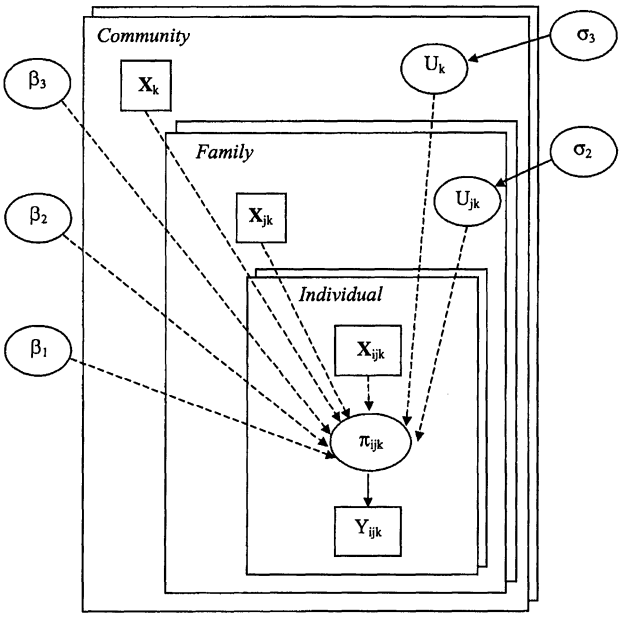

Fig. 1. Three-level logit model with family and community effects on an individual level binary outcome

2.1. Maximum likelihood

To fit the multilevel model by maximum likelihood we need to obtain the unconditional distribution $[\mathbf{Y}]$ of the response. To do this we multiply the conditional Bernoulli distribution $[\mathbf{Y}|\mathbf{U}]$ of the response given the random effects by the Gaussian density $[\mathbf{U}]$ of the random effects, to obtain the joint distribution

$$[\mathbf{Y}, \mathbf{U}] = [\mathbf{Y}|\mathbf{U}][\mathbf{U}],$$

and finally 'integrate out' the random effects to obtain $[\mathbf{Y}]$. It is the last step that requires numerical integration. We use Gauss–Hermite numerical quadrature, evaluating the conditional distribution $[\mathbf{Y}|\mathbf{U}]$ by using a 20×20 grid of values for U_{jk} and U_k. Once the integrals have been replaced by discrete sums, first and even second derivatives can be obtained in a laborious but straightforward process. Standard numerical procedures can then be applied. Often a mixture of steepest descent to improve initial estimates followed by Newton–Raphson or quasi-Newton iterations works well. This is the strategy that we used to obtain the estimates reported in Pebley *et al.* (1996). For further details see Longford (1993).

2.2. Bayesian estimation

Recent developments in Bayesian inference avoid the need for numerical integration by repeated sampling from the posterior distribution of the parameters using Gibbs sampling, a technique first used in the context of generalized linear models by Zeger and Karim (1991). To apply this framework we adopt a Bayesian view, treating the parameters (as well as the observations) as random variables. We augment our basic model by assigning prior (or hyperprior) distributions $[\boldsymbol{\beta}]$ to the fixed effects and $[\boldsymbol{\sigma}^2]$ to the variances of the random effects (strictly speaking, we work with the precisions $1/\boldsymbol{\sigma}^2$). To obtain Bayesian estimates that are roughly comparable with maximum likelihood we use non-informative or vague priors. Specifically, we assume that $\beta_i \sim N(0, 1/\tau)$ with precision $\tau = 0.0001$ (so the variance is 10000) and that $1/\sigma_i^2 \sim \Gamma(\epsilon, \epsilon)$ with $\epsilon = 0.1$ (so the mean is 1 and the variance is 10). Almost identical results were obtained by using a Pareto prior for $1/\sigma_i^2$, which is equivalent to a uniform prior for σ_i (see Spiegelhalter *et al.* (1996), page 38).

We then estimate the model by using the Gibbs sampler as implemented in the software package BUGS (Spiegelhalter *et al.*, 1996). A good introduction to Gibbs sampling may be found in Casella and George (1992). Briefly, the Gibbs sampler is a Markov chain Monte Carlo method for simulating observations from a joint distribution by sampling repeatedly from the so-called full conditional distributions. In our case we need to sample from the posterior distribution of the parameters given the data, say $[\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{U}|\mathbf{Y}]$. The Gibbs sampler tells us that we can sample instead from the three full-conditional distributions

$$[\boldsymbol{\beta}|\boldsymbol{\sigma}^2, \mathbf{U}, \mathbf{Y}], \quad [\boldsymbol{\sigma}^2|\boldsymbol{\beta}, \mathbf{U}, \mathbf{Y}] \quad \text{and} \quad [\mathbf{U}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{Y}]. \tag{4}$$

As shown in Zeger and Karim (1991), the first two distributions further simplify to $[\boldsymbol{\beta}|\mathbf{U}, \mathbf{Y}]$ (i.e. the posterior distribution of the fixed effects depends on the random effects $\mathbf{U}$ and the response $\mathbf{Y}$, but not on the variance components $\boldsymbol{\sigma}^2$) and $[\boldsymbol{\sigma}^2|\mathbf{U}]$ (i.e. given the random effects $\mathbf{U}$, the response $\mathbf{Y}$ and the fixed effects $\boldsymbol{\beta}$ have no further information about the variance components $\boldsymbol{\sigma}^2$). The Gibbs sampler draws observations from each of these distributions in turn. If $\boldsymbol{\beta}_k$, $\boldsymbol{\sigma}_k^2$ and $\mathbf{U}_k$ denote a sample (with $k = 0$ for starting values) then the Gibbs sampler generates a new sample drawing $\boldsymbol{\beta}_{k+1}$ from $[\boldsymbol{\beta}|\mathbf{U}_k, \mathbf{Y}]$, $\boldsymbol{\sigma}_{k+1}^2$ from $[\boldsymbol{\sigma}^2|\mathbf{U}_k]$ and finally $\mathbf{U}_{k+1}$ from $[\mathbf{U}|\boldsymbol{\beta}_{k+1}, \boldsymbol{\sigma}_{k+1}^2, \mathbf{Y}]$. Under reasonably general conditions the distribution of the samples converges to the desired joint posterior distribution as $k \to \infty$. Usually one discards a 'burn-

in' period which is sufficiently long to ensure that the chain has converged to its stationary distribution and uses the remaining observations to estimate features of the posterior distributions (such as the posterior means) by using standard statistical procedures. Note that successive observations are not independent.

2.3. Marginal quasi-likelihood

The approximate estimation procedures that are evaluated here can be motivated by considering a linearized form of the multilevel logit model. Note from the discussion leading to equation (1) that we can write

$$\mathbf{Y} = \boldsymbol{\pi} + \boldsymbol{\epsilon}, \qquad \text{with } \boldsymbol{\pi} = f(\mathbf{X}\boldsymbol{\beta} + \mathbf{ZU}), \tag{5}$$

where $\mathbf{Y}$ is the vector of individual responses, f is the inverse logit (or antilogit) transformation and $\boldsymbol{\epsilon}$ is a heteroscedastic error term with mean 0 and (conditional) variance given by a diagonal matrix with entries $\pi(1-\pi)$.

First-order MQL approximates $\boldsymbol{\pi}$ by using a first-order Taylor series expansion of $f(\cdot)$ around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\mathbf{U} = \mathbf{0}$, where $\boldsymbol{\beta}_0$ is the current estimate of the fixed effects. The expansion has the structure of a multilevel linear model that can be fitted by using standard algorithms (e.g. Goldstein (1995)), leading to an improved estimate of $\boldsymbol{\beta}$, which is then used as the new pivot. The procedure is iterated to convergence. Longford (1994) used a quadratic approximation to the log-likelihood function that is equivalent to first-order MQL and leads to exactly the same estimates (Rodríguez and Goldman, 1995).

Second-order MQL extends the Taylor series expansion by adding the second-order term corresponding to the random effects $\mathbf{U}$, but not second-order terms on the fixed effects $\boldsymbol{\beta}$, nor mixed or cross-product terms. This strategy leads again to an approximating multilevel linear model that can be fitted by using standard algorithms. See Goldstein (1991) and Rodríguez and Goldman (1995) for details.

2.4. Penalized quasi-likelihood

It should not be surprising, given the nature of the assumptions underlying MQL (in particular the expansion about $\mathbf{U} = \mathbf{0}$), that the approximation performs well when the random effects are small (i.e. their variances are close to 0) but may fail if the random effects are moderate or large. In Section 3 we provide evidence that MQL results in a substantial bias even for moderate amounts of clustering.

An alternative procedure that may work better under these circumstances is to use a non-zero pivot for the random effects in the Taylor series expansion. In particular, we may expand $f(\cdot)$ about $\mathbf{U} = \mathbf{U}_0$ where $\mathbf{U}_0$ is the empirical Bayes estimate (or predictor) of the random effects, defined as the mean of $[\mathbf{U}|\mathbf{Y}]$ estimated at the current parameter values. For any given value of $\mathbf{U}_0$ the resulting approximating model is again a linear multilevel model and can be estimated by using standard algorithms. The improved estimates of both the fixed and the random effects are then used to obtain a new approximating linear model, and the procedure is iterated to convergence.

Breslow and Clayton (1993) termed this procedure PQL because it is related to work by Green (1987) on semiparametric regression. The same procedure was derived from a Bayesian perspective by Laird (1978) and Stiratelli *et al.* (1984) and has been used by Schall (1991).

To improve further on first-order PQL, Goldstein and Rasbash (1996) proposed incorporating a second-order term on the random effects (but no second-order terms on the

fixed effects and no mixed terms). They referred to the resulting procedure as second-order penalized or predictive quasi-likelihood, or second-order PQL.

All four approximations discussed here, which will be abbreviated MQL-1, MQL-2, PQL-1 and PQL-2, have been implemented in the software package MLn (Rasbash and Woodhouse, 1995) and its successor MLwiN (Goldstein *et al.*, 1998). MQL-1 is also available in VARCL (Longford, 1994) and PQL-1 has been implemented in HLM (Bryk *et al.*, 1996).

2.5. The parametric bootstrap

Bootstrapping is a popular technique for assessing the bias and estimating the standard errors of parameter estimates in a wide variety of models. Kuk (1995) proposed an iterated bootstrap procedure that has been successfully applied in the context of a two-level logit model by Goldstein (1996). The basic idea is to generate a number of samples from the model evaluated at current parameter estimates. The model is then estimated for each of these samples and the estimates averaged across replications. The difference between the values used in the simulation and these averages provides an estimate of the bias of the approximate procedure, which can then be used to correct the parameter estimates. Because the bias depends on the values of the parameters, several iterations are required. This procedure has been implemented in the latest version of MLwiN (Goldstein *et al.*, 1998).

3. The simulation study

Our interest in the estimation of multilevel models for binary outcomes arose from a series of studies of health care utilization in Guatemala (Pebley and Goldman, 1992; Goldman and Pebley, 1994; Pebley *et al.*, 1996). Preliminary exploratory analyses had revealed large family and community level effects in the use of various forms of modern health care, even after controlling for observed covariates at the individual, family and community levels (Pebley and Goldman, 1992). Yet more formal multilevel analyses using the software packages VARCL (Longford, 1994) and ML3 (Prosser *et al.*, 1991), which produced identical estimates, suggested very small influences of the family and community on the use of health care.

To resolve the inconsistency between our exploratory and confirmatory analyses, we decided to validate the estimation procedures used in existing software by running them on simulated data for which the true parameter values were known. Although we designed 10 sets of simulations with varying hierarchical structures and degrees of clustering, we focus here on an analysis that used the same structure as the actual Guatemalan data on prenatal care, with 2449 births pertaining to 1558 mothers who were living in 161 communities. We created three composite variables capturing characteristics of the pregnancy, mother and community and set their coefficients to 1. We then added random effects to represent unobserved characteristics of the mother and community; these were sampled from standard normal distributions with mean 0 and variance 1. Finally, we simulated binary responses satisfying model (1). This strategy was used to generate 100 data sets that were then analysed using VARCL. A full description of the procedure is presented in Rodríguez and Goldman (1995).

The results shown in Table 1 reveal that first-order MQL is subject to a substantial bias in the estimation of both the fixed and the random effects. The attenuation in the estimates of the β-coefficients is of the order of 25%, whereas the underestimation of the family level random effect is a particularly severe 90%. The estimates based on second-order MQL

Table 1. Estimates for simulated data using the Guatemala structure†

Effects	*True value*	*Results from the following methods:*		
		MQL-1	*MQL-2*	*PQL-2*
Fixed effects				
Individual	1	0.74	0.85	0.96
Family	1	0.74	0.86	0.96
Community	1	0.77	0.91	0.96
Random effects (σ)				
Family	1	0.10	0.28	0.73
Community	1	0.73	0.76	0.93

†Source: the MQL estimates are from Rodríguez and Goldman (1995) and the PQL estimates are from Goldstein and Rasbash (1996).

represent an improvement, but remain downwardly biased, with an underestimation of 72% for the problematic family level effect.

The last column shows second-order PQL estimates, obtained by Goldstein and Rasbash (1996) from the first 25 of our 100 simulated data sets. These estimates represent a considerable improvement over the earlier approximations, especially for the family level effect. The fixed parameter estimates are now within 4% of their true values, and the community level standard deviation is within 7%, but there remains a 27% bias in the estimate of variation at the family level. Goldstein and Rasbash (1996) also obtained estimates from PQL-1 and PQL-2 (not shown in Table 1) based on 200 simulations of the same data structure that is used for Table 1. Their results for PQL-2 agree closely with the values shown in Table 1; their estimates for PQL-1 lie between those for MQL and PQL-2.

Commenting on the comparisons presented in Table 1, Goldstein and Rasbash (1996) noted that

> 'the example chosen is based on large underlying random parameter values'

and added that

> 'in the more common case where variances in a random intercept model do not exceed about 0.5 the first-order PQL model can be expected to perform well, and for smaller variances the first-order MQL model will often be adequate'.

The problem with this recommendation, however, is that in practice we are unlikely to know the true magnitude of these effects. In fact, a naïve application of MQL would have led us to conclude that the family effect was less than 0.5, and therefore that MQL estimates were adequate. (A better strategy will be proposed in Section 5.)

4. Estimates from actual data

The simulated data used to generate Table 1 had indeed been designed to represent substantial degrees of clustering, with standard deviations of 1 for the family and community random effects, corresponding to intrafamily and intracommunity correlations of 0.38 and 0.19 respectively. However, our subsequent application of exact maximum likelihood procedures to data derived from the 1987 National Survey of Maternal and Child Health in Guatemala (Ministerio de Salud Pública y Asistencia Social and Instituto de Nutrición de Centro América y Panamá, 1989) revealed even higher degrees of clustering of health behaviours.

Based on six distinct (binary) outcome variables related to the use of health care during pregnancy or immunization of young children, the estimated standard deviations ranged from 2.3 to 7.4 at the family level, and from 1.0 to 4.6 at the community level, corresponding to intrafamily correlations between 0.66 and 0.95 and intracommunity correlations between 0.11 and 0.51 (Pebley *et al.*, 1996).

Although surprising to many analysts, these results are consistent with the exploratory analyses by Pebley and Goldman (1992) mentioned earlier and raise the question of the extent to which even the improved approximations can be trusted under these conditions. To examine this question we consider two outcomes that represent the range of clustering of health behaviour observed in the Guatemala study:

(a) obtaining a complete set of (eight) immunizations among children who receive at least one immunization and
(b) using modern prenatal care (i.e. doctors or nurses) among women who use some form of prenatal care.

4.1. Complete immunization

For this analysis we focus on 2159 living children (ages 1–4 years) who had received some immunization and analyse whether they had received a full set of immunizations, as a function of individual, family and community characteristics. These 2159 children come from 1595 families living in 161 communities (resulting in average numbers of children per family and community of 1.4 and 13.4 respectively). This outcome exhibits the smallest degree of clustering at both the family and the community levels of the six outcomes analysed in Pebley *et al.* (1996).

Table 2 shows parameter estimates obtained by using ordinary logistic regression analysis, the four approximate procedures discussed earlier, maximum likelihood and Gibbs sampling, as well as bootstrapped PQL estimates. To keep Table 2 manageable we do not report the standard errors. However, t-ratios for the maximum likelihood estimates may be found in Pebley *et al.* (1996) and indicate that the effects of child's age, husband's education, rural residence and proportion indigenous are statistically significant. The other variables were retained for comparability with the original study, which included them because they had been hypothesized to affect immunization and had been shown to be associated with other health-related outcomes.

Unlike in the analysis of the simulated data described earlier, there is no 'truth' against which we can assess alternative estimates derived from actual data. For this paper we use two standards for comparison: maximum likelihood estimates obtained via Gauss–Hermite quadrature and Bayesian estimates based on Gibbs sampling using non-informative priors. The fact that these two methods lead to similar estimates provides an informal validation of their use as a bench-mark.

Broadly speaking, the results of first-order MQL are very similar to those of ordinary logistic regression, with the sole exception of the coefficients of ethnicity, which change from positive (but not significant) to negligible. Second-order MQL produces a modest improvement; the fact that all coefficients increase in absolute magnitude suggests less attenuation than with MQL-1. First-order PQL produces results that are very similar to those of second-order MQL. Second-order PQL, however, shows less attenuation than the other three approximate methods. We started each of these procedures from the estimates obtained in the last step of the previous method. They all converged fairly quickly except for second-order PQL, which took 64 iterations.

Table 2. Estimates for the multilevel model of complete immunization among children receiving any immunization†

Effects	*Results for the following methods:*							
	Logit	*MQL-1*	*MQL-2*	*PQL-1*	*PQL-2*	*PQL-B*	*Maximum likelihood*	*Gibbs*
Fixed effects								
Individual								
Child age ⩾ 2 years‡	0.95	0.93	1.11	0.98	1.44	1.80	1.72	1.84
Mother age ⩾ 25 years	−0.08	−0.08	−0.10	−0.09	−0.16	−0.19	−0.21	−0.26
Birth order 2–3	−0.08	−0.09	−0.11	−0.10	−0.19	−0.15	−0.26	−0.29
Birth order 4–6	0.09	0.13	0.15	0.13	0.17	0.27	0.18	0.21
Birth order ⩾ 7	0.15	0.19	0.23	0.20	0.33	0.39	0.43	0.50
Family								
Indigenous, no Spanish	0.28	−0.04	−0.05	−0.05	−0.13	−0.06	−0.18	−0.22
Indigenous Spanish	0.22	0.01	0.01	0.00	−0.05	0.03	−0.08	−0.11
Mother's education primary	0.25	0.21	0.25	0.22	0.34	0.42	0.43	0.48
Mother's education secondary or better	0.30	0.22	0.27	0.23	0.34	0.46	0.42	0.46
Husband's education primary‡	0.29	0.28	0.34	0.30	0.44	0.57	0.54	0.59
Husband's education secondary or better	0.21	0.25	0.31	0.27	0.41	0.47	0.51	0.55
Husband's education missing	0.03	0.02	0.02	0.02	0.01	0.07	−0.01	0.00
Mother ever worked	0.25	0.19	0.24	0.20	0.31	0.37	0.39	0.42
Community								
Rural‡	−0.50	−0.47	−0.57	−0.50	−0.73	−0.93	−0.89	−0.96
Proportion indigenous, 1981‡	−0.78	−0.64	−0.78	−0.67	−0.95	−1.21	−1.15	−1.22
Random effects								
Standard deviations σ								
Family	—	0.63	0.72	0.73	1.75	2.69	2.32	2.60
Community	—	0.53	0.55	0.56	0.84	1.06	1.02	1.13
Intraclass correlations ρ								
Family	—	0.17	0.20	0.20	0.53	0.72	−0.66	0.71
Community	—	0.07	0.07	0.07	0.10	0.10	0.11	0.11

†The reference categories are child aged 1 year, mother's age less than 25 years, birth order 1, Ladino, mother no education, husband no education, mother never worked and urban residence.
‡Fixed effects significant at the 5% level according to the maximum likelihood analysis.

The column labelled maximum likelihood reports the estimates obtained by Gauss–Hermite quadrature. We note in passing that numerical integration requires working in the scale of the likelihood rather than the log-likelihood, as is customary, and that special care must be taken to avoid a substantial loss of precision in calculating probabilities that can differ by several orders of magnitude. We used the optimization routines built into S-PLUS, using functions written in C to evaluate the log-likelihood and its first and second derivatives. In practice we found the best results by using analytic first derivatives and numerical second derivatives, but we used analytic second derivatives in evaluating asymptotic standard errors after convergence. We also discovered that we needed to use 20 quadrature points at each level to attain an acceptable precision in our calculations. Note that the maximum likelihood estimates are as large as (in absolute magnitude) or larger than the PQL-2 estimates.

The final column reports estimates obtained by using the Gibbs sampler. We started the sampler from the maximum likelihood estimates which, given the diffuse nature of the priors, should be reasonably close to the posterior mode. Expecting quick convergence we used a burn-in run of 200 samples and then ran the sampler for 1000 iterations. A preliminary analysis of the results, however, revealed very slow mixing and poor convergence—

particularly for the parameters representing the variance components — so we restarted the sampler using the last seed and ran it for an additional 4000 iterations. (This is a computer-intensive procedure, taking 5 h on a Sun Sparcstation 20 computer under light load conditions.) A battery of diagnostic tests indicated that the longer run was adequate for our purposes. In particular, we used Geweke's (1992) procedure to test for convergence of each chain. We also ran the `gibbsit` software of Raftery and Lewis (1992, 1996) to check that we had enough iterations to estimate each posterior cumulative distribution function evaluated at the 95% credible limits within 0.0125 with probability 0.95. We also calculated the efficiency of the Markov chains for estimating the posterior mean of each parameter by using the method of Roberts (1996), ensuring that we had the equivalent of at least 100 independently and identically distributed observations in the worst case where the efficiency of the chain was estimated to be as low as 2%.

Fig. 2 provides a flavour of the type of output that was obtained from the Gibbs sampler. We have selected three fixed effects, one each from the individual, family and community levels, and the two random effects. Fig. 2(a) shows traces of the sampled values for a chain of 5000 iterations, after discarding the burn-in of 200. Fig. 2(b) shows the posterior densities, estimated from the 5000 samples by using a kernel smoother with bandwidth equal to 25% of the data. The first pair of plots corresponds to mother's age less than 25 years. This variable was selected because it is very well behaved. The trace plot shows the type of homogeneous mixing that is desirable. The second variable is the indicator of mother's primary education and is reasonably well behaved. The third pair of plots corresponds to the proportion indigenous in 1981 and was selected because it is one of the worst-behaved fixed effects. The fourth pair of plots corresponds to the standard deviation of the family random effect and shows an anomaly known as 'slow mixing', where the sampler appears to drift from highs to lows, rather than quickly covering the sample space. We believe that this problem is due in part to the small number of children per family, and it might be ameliorated by using a more informative prior. The standard deviation of the community random effect is much better behaved, although it tends to exhibit some peaks. Although we encountered slow mixing, the efficiency calculations described above indicate that the chain is sufficiently long to provide useful estimates for the comparisons that are presented in this paper.

The results reported in Table 2 are the empirical means of the last 5000 iterations. (Note that, although the Bayesian model is formulated in terms of the precision of the random effects, we monitored and plotted the standard deviation, calculated as the reciprocal of the square root of the precision.) The results of Gibbs sampling are in general agreement with the results from the maximum likelihood analysis, confirming the fact that we have substantial clustering at both the family and the community levels. In fact, the Bayesian analysis suggests that the standard deviations of the random effects could be even larger than estimated by maximum likelihood.

The general pattern of results indicates that the approximate procedures are subject to a substantial bias, with PQL being generally less biased than MQL. Consider, for example, the effect of husband's primary education, which according to the maximum likelihood analysis is significant at the 5% level. The estimated odds ratio (exponentiated coefficient) contrasting primary with no education increases from 1.33 in the ordinary logit regression to 1.56 using PQL-2, the best available approximation, to 1.72 according to maximum likelihood and 1.80 based on the Gibbs sampler. Similar remarks apply to the variance components. The estimate of the intrafamily correlation is 0.17 by the simpler approximation (MQL-1) and 0.53 by the best available approximation, when in fact the true value appears to be around 0.66–0.71. The intracommunity correlation is relatively modest and, not unexpectedly, seems to be estimated

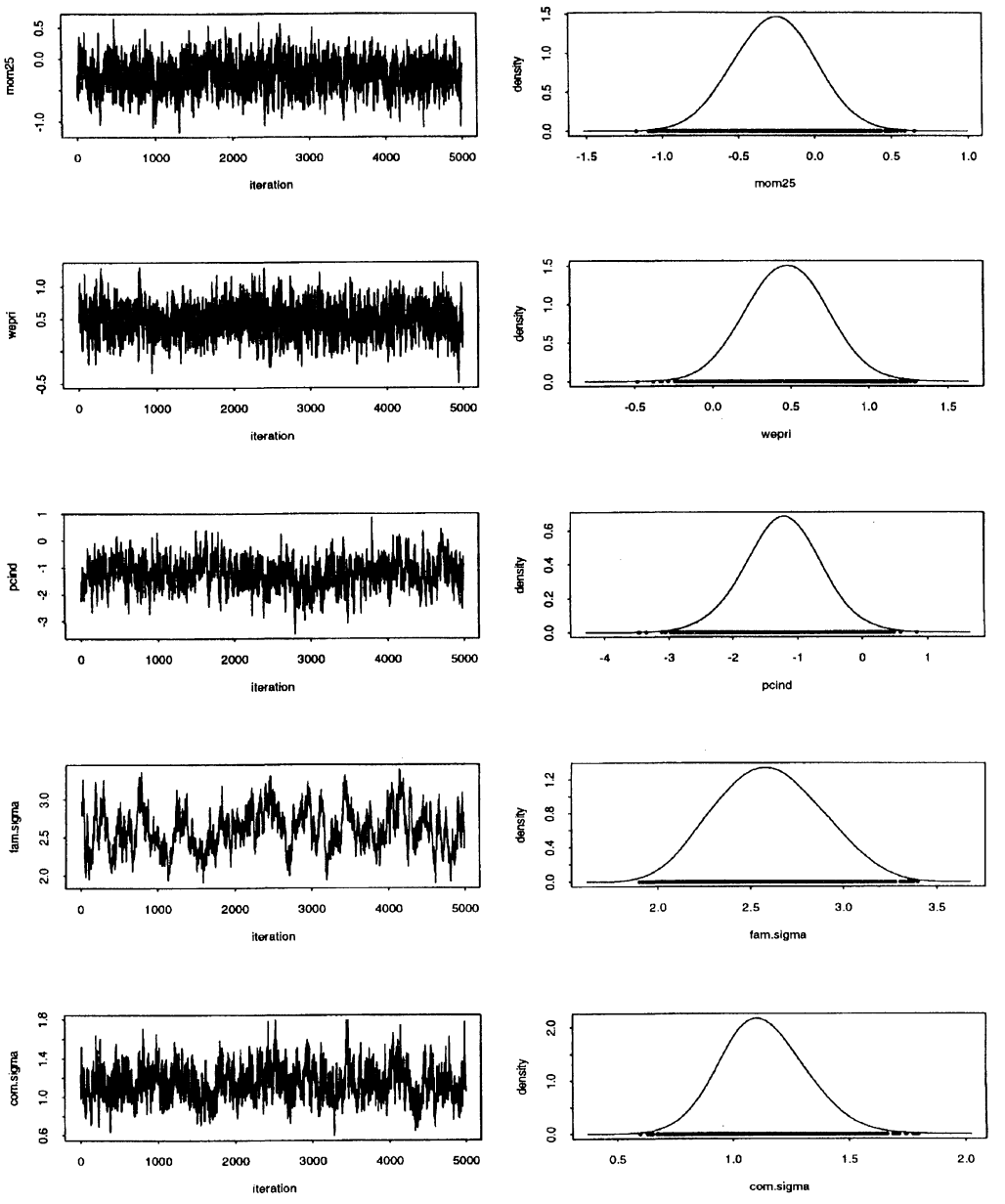

Fig. 2. Trace and density plots for Gibbs samples from the multilevel model of immunization

more consistently than the intrafamily correlation, with estimates ranging from 0.07 for MQL-1 to 0.10 for PQL-2 and 0.11 for maximum likelihood or Gibbs.

We attempted to correct the bias in the approximation procedures by using the parametric bootstrap method described by Kuk (1995) and Goldstein (1996) and implemented in MLwiN. We bootstrapped both the MQL-1 and the PQL-1 estimates, generating replicates of 100 samples each. We were willing to increase the size of each replicate to 400 samples (as done in Goldstein (1996)) if the results showed excessive noise, but this proved unnecessary.

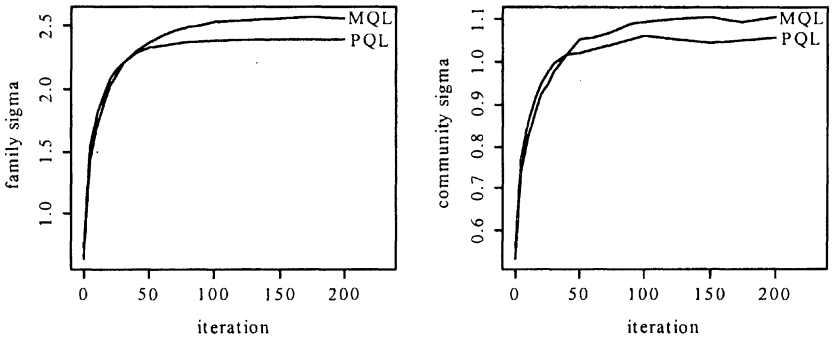

Fig. 3. Trajectories of the bootstrap estimates for the immunization model

We monitored the standard deviations of the family and community random effects, which increased monotonically from one replication to the next before settling down after about 150 replications, as shown in Fig. 3.

This is an extremely computer-intensive process, with iterations proceeding at a rate of no more than 10 replicates per hour on a computer with two 450-MHz Pentium II Xeon processors (although MLwiN uses only one processor). On the positive side, we found that monitoring the convergence was much simpler for the bootstrap procedure than it was for Gibbs sampling. The column labelled PQL-B in Table 2 shows the bootstrapped PQL-1 estimates, which were in somewhat closer agreement with the maximum likelihood estimates than the bootstrapped MQL-1 estimates. The general pattern of results indicates that the iterated bootstrap, although slow to converge, has successfully corrected the bias.

4.2. Modern prenatal care

The second outcome variable that is analysed here concerns the use of modern prenatal care among women using some form of prenatal care. This analysis is based on 2449 births (born in the 5-year period before the survey), pertaining to 1558 mothers who were living in 161 communities (resulting in average numbers of pregnancies per family and community of 1.6 and 15.2 respectively). This hierarchical structure is identical with that used in the simulated data set described earlier (although, unbeknown to the authors at the time, the latter incorporated smaller random effects). This outcome exhibits the largest amount of clustering at the family level ($\rho = 0.95$) of the six analyses presented in Pebley *et al.* (1996). It also exhibits a fair amount of clustering at the community level ($\rho = 0.19$), but comparable or higher values are found for other behaviours.

As in the immunization model, the explanatory variables for prenatal care include covariates at all three levels, although some of the variables differ from those in the previous model. According to the maximum likelihood analysis, the following effects are statistically significant: child's age, mother's age, ethnicity, mother's education, husband's education, husband's occupation, presence of a modern toilet, proportion indigenous and distance to the nearest clinic.

Table 3 shows the prenatal care estimates based on an ordinary logit model, the four approximations under consideration, maximum likelihood and Gibbs sampling, as well as a set of bootstrapped PQL-1 estimates. The first-order MQL estimates of the random effects exhibit a very substantial downward bias, which is only minimally reduced by first-order PQL.

Table 3. Estimates for the multilevel model of modern prenatal care among women using some form of prenatal care†

Effects	*Results for the following methods:*							
	Logit	*MQL-1*	*MQL-2*	*PQL-1*	*PQL-2*	*PQL-B*	*Maximum likelihood*	*Gibbs*
Fixed effects								
Individual								
Child aged 3–4 years‡	−0.20	−0.17	−0.25	−0.22	−0.44	−0.81	−1.04	−1.33
Mother aged ⩾ 25 years‡	0.32	0.31	0.38	0.36	0.58	1.35	1.08	1.26
Birth order 2–3	−0.10	−0.10	−0.16	−0.13	−0.20	−0.49	−0.75	−1.00
Birth order 4–6	−0.23	−0.23	−0.32	−0.26	−0.31	−0.97	−0.56	−0.49
Birth order ⩾ 7	−0.19	−0.28	−0.45	−0.30	−0.45	−1.08	−1.08	−1.21
Family								
Indigenous, no Spanish‡	−0.84	−0.97	−1.02	−1.22	−2.18	−4.63	−5.60	−7.54
Indigenous Spanish‡	−0.57	−0.56	−0.93	−0.67	−1.00	−2.54	−2.62	−4.00
Mother's education primary‡	0.31	0.35	0.59	0.42	0.65	1.64	1.89	2.62
Mother's education secondary or better‡	1.01	0.90	1.06	0.98	1.93	3.81	3.61	5.68
Husband's education primary	0.18	0.22	0.32	0.25	0.30	0.95	0.96	1.11
Husband's education secondary or better‡	0.68	0.69	0.85	0.82	1.59	3.07	4.37	4.85
Husband's education missing	0.00	0.06	0.07	0.06	0.01	0.16	0.13	0.02
Husband professional, sales, clerk	−0.32	−0.40	−0.49	−0.47	−0.64	−0.60	−0.62	−0.56
Husband agricultural self-employed	−0.54	−0.52	−0.66	−0.62	−0.86	−1.75	−1.77	−2.64
Husband agricultural employee‡	−0.70	−0.27	−0.33	−0.29	−0.25	−2.34	−2.67	−3.77
Husband skilled service	−0.37	−0.15	−0.19	−0.18	−0.05	−1.05	−0.80	−1.12
Modern toilet in household‡	0.47	0.37	0.57	0.41	0.94	1.72	2.01	2.69
Television not watched daily	0.32	0.27	0.48	0.31	0.53	1.16	1.35	2.03
Television watched daily	0.47	0.33	0.41	0.39	0.67	1.55	1.51	2.05
Community								
Proportion indigenous, 1981‡	−0.90	−0.97	−1.61	−1.12	−2.05	−4.48	−5.01	−6.61
Distance to nearest clinic‡	−0.01	−0.01	−0.01	−0.01	−0.02	−0.05	−0.05	−0.07
Random effects								
Standard deviations σ								
Family	—	1.01	1.74	1.25	2.75	6.66	7.40	10.24
Community	—	0.79	1.23	0.86	1.71	3.48	3.74	5.40
Intraclass correlations ρ								
Family	—	0.33	0.58	0.41	0.76	0.95	0.95	0.98
Community	—	0.13	0.19	0.13	0.21	0.20	0.19	0.21

†The reference categories are child aged 0–2 years, mother's age less than 25 years, birth order 1, Ladino, mother no education, husband no education, husband not working or unskilled occupation, no modern toilet in the household and no television in the household.
‡Fixed effects are significant at the 5% level according to the maximum likelihood analysis.

The second-order improvements to these approximations proved to be numerically unstable and both MQL-2 and PQL-2 failed to converge for the prenatal care data. The MQL-2 estimates shown in Table 3 correspond to about a dozen iterations starting from the MQL-1 estimates and were still fluctuating when the procedure aborted with a protection fault. The PQL-2 procedure appeared to oscillate between two solutions with similar likelihood values and was stopped after 250 iterations. The alternating estimates of the fixed effects were similar, except for husband's and wife's secondary education, the availability of a toilet and proportion indigenous. The estimated standard deviations of the random effects alternated between (2.75, 1.71) and (3:01, 1.49). The estimates shown in Table 3 correspond to the first set, which has the larger fixed effects, but similar conclusions would result from the

second set. The results suggest that second-order improvements reduce the bias in the estimates, particularly in the case of PQL, but the standard deviations of the random effects are still substantially underestimated, particularly at the family level.

The Gibbs sampler experienced some difficulty in converging with the prenatal care data, particularly for variance components and for fixed parameters involving relatively small groups in the higher socioeconomic strata. The application of the battery of diagnostic procedures described earlier leads us to conclude that a run of 5200 iterations was inadequate. Calculations based on the `gibbsit` software suggested that we would need 20000 iterations to achieve the same precision as in the analysis of immunization, whereas estimates based on Roberts's (1996) method indicated that the efficiency of the chain could be as low as 1% for some parameters. Instead of running just one very long chain, we heeded the advice of Gelman and Rubin (1992) and ran three additional chains with different starting values. Specifically, we started two chains at the maximum likelihood estimates of the fixed effects while setting the precisions of the random effects to values that were close to the first and third quartiles of their posterior distributions as estimated in the first run. For the third run we set the precisions close to the medians and varied the estimates of the fixed effects, setting them alternately to the first and third quartiles of the first run. We were relieved to note a substantial overlap in the posterior distributions obtained from all four runs. A calculation of the Gelman and Rubin (1992) criteria as implemented in the S function `itsim` provides no evidence against the notion that, although inefficient, the chains have covered the target distribution. The values reported under Gibbs in Table 3 are pooled estimates based on the last 5000 observations from each of the four chains. According to our efficiency calculations, these estimates are at least as precise as the estimates in Table 2.

With only one exception, the maximum likelihood and Gibbs estimates are each larger in absolute magnitude than any of the approximate estimates, for both the fixed effects and the standard deviations of the random effects. This result confirms the finding from Table 2 that non-trivial attenuation bias is associated with all the approximations. These two benchmarks are less consistent with each other than was the case for immunization, however, with the Bayesian estimates generally indicating stronger effects than the maximum likelihood estimates.

Our attempts to correct the bias in the approximation procedures by using the parametric bootstrap implemented in MLwiN were somewhat less successful than for the immunization model. Convergence of the iterated bootstrap for both MQL-1 and PQL-1 was excruciatingly slow, and both runs failed with overflow errors after more than 400 iterations, while the community and family standard deviations were still somewhat short of the maximum likelihood estimates. The column labelled PQL-B in Table 3 reports the bootstrapped PQL-1 estimates after 400 replicate sets and shows that most of the bias had been corrected at that point.

5. Discussion

In this analysis we used data related to health care in Guatemala to evaluate the performance of several approximate estimation procedures for three-level models of binary outcomes. The study confirms and extends our earlier results, which were based on simulations designed to replicate actual data sets (Rodríguez and Goldman, 1995). We find that MQL-1 estimates differ little from those in ordinary logit models, and that MQL-2 and PQL-1 offer only slight improvements. PQL-2 provides the best approximation, but the estimates of the fixed and random parameters continue to be attenuated in comparison with either maximum likelihood

or Gibbs sampling. Not surprisingly, the attenuation bias is much greater in the model of modern prenatal care, which incorporates considerably larger family and community effects than does the immunization model. The bias can be virtually eliminated by bootstrapping either MQL-1 or PQL-1 estimates, but this procedure proved computationally more intensive than Markov chain Monte Carlo estimation and failed to converge for the prenatal care model.

Some analysts find our results concerning the bias unexpected, believing that clustering can affect the standard errors but not the parameter estimates themselves. This is indeed the case for so-called marginal or population average models (Zeger *et al.*, 1988), but not for the conditional or unit-specific models considered here, where we model individual responses given family and community characteristics. The distinction is immaterial in linear models, but it is highly relevant in generalized linear models using a non-linear link function such as the logit or log-link (Goldstein and Rasbash, 1996). Further discussion of marginal models may be found in Diggle *et al.* (1994).

Another atypical feature of our results is the sheer magnitude of the estimated random effects, which are indeed larger than those found in other studies using different outcomes, such as child mortality (see, for example, Guo and Rodríguez (1992)). It is important to keep in mind, however, that the behavioural outcomes that were considered here are largely under the control of the family (e.g. whether or not to seek care from a biomedical practitioner during pregnancy). It is plausible that families typically behave consistently in deciding how to treat successive pregnancies or successive children. The relatively large community effects are consistent with previous hypotheses related to the importance of the availability of health services, the basic infrastructure and the ethnic and socioeconomic composition of the community in explaining the use of biomedical forms of health care.

A related concern stems from the fact that we have fairly small clusters at the family level, with 1.4–1.6 children per family. The properties of maximum likelihood estimators based on such small clusters are not known and can only be established by using Monte Carlo simulation. Our finding that estimates from the Gibbs sampler are of about the same magnitude as (or higher than) those from maximum likelihood provides an informal validation of the maximum likelihood estimates and furnishes us with two standards against which to assess the approximate procedures.

Given the large intrafamily correlations that we found, particularly for prenatal care, one may wonder whether we should study the outcome at the level of the family rather than the child or pregnancy, using binomial models with a random effect at the cluster level and an additional parameter to allow for underdispersion or overdispersion relative to the binomial variance. However, such models would assume that children or pregnancies are exchangeable: within a family the probability of immunization must be the same for all children, and the probability of prenatal care must be the same for all pregnancies. In particular, this means that we could not have child or pregnancy level covariates. An important objective of our original study was to assess the effect of child or pregnancy covariates on immunization and prenatal care. Moreover, in both analyses we found significant effects at the individual level; for example children aged 2–4 years are substantially more likely to be fully immunized than children under age 2 years, everything else being equal. An analysis at level 2 would have missed these effects.

In the light of the size of the variances of the random effects that underlie our two examples, our evaluation of MQL and PQL approximations subjects the procedures to rather stringent tests. Although neither of the examples explored in this paper embodies small random effects at both the family and the community levels, our earlier simulations suggest

that all the approximations are likely to work reasonably well in this case. Unfortunately, the analyst rarely has any information *a priori* on the degree of clustering in the data. As a result, it seems hazardous to rely on these approximations.

A strategy suggested by Goldstein and Rasbash (1996) is to compare MQL-1 and PQL-1 estimates and to accept them if they are similar. Unfortunately, a small difference is not necessarily indicative of a lack of bias. In Table 2, for example, the standard deviation of the community effect changed from 0.53 for MQL-1 to 0.56 for PQL-1, a small change that provides no indication that the maximum likelihood estimate is in fact 1.02, almost double the value. A better strategy is to calculate all four approximate procedures, where a small range would be a more reliable indicator of a small bias. Better still, we recommend computing PQL-1 estimates and running five iterations of the bootstrap. If the trajectories of the estimates remain flat, the PQL-1 estimates can be deemed satisfactory; otherwise the bootstrap should be continued to convergence or one should shift to Bayesian estimates by using Gibbs sampling.

These findings also suggest that further research is needed to provide the analyst with adequate tools for the efficient estimation of multilevel models for binary or count data. Maximum likelihood estimation needs to be validated for small clusters and implemented more widely. The numerical quadrature procedures that were used here can be applied to random-intercept models, but computer-intensive procedures such as Monte Carlo integration are probably unavoidable for more complex random-coefficient models. The recent work by McCulloch (1997) on Monte Carlo variants of the EM and Newton–Raphson algorithms for generalized linear mixed models provides promising avenues for further work. Although Bayesian methods are not yet generally accepted in the social science research community, the use of vague or uninformative priors has certainly made them more attractive.

Finally, for most analysts the lack of generally available software can be a crucial deterrent to the adoption of new estimation techniques. Maximum likelihood estimation of random-effect models for binary responses can be accomplished by using numerical quadrature. This technique was first used in the package Egret (Statistics and Epidemiology Research Corporation, 1995) for two-level random-intercept logit or probit models and has been implemented for more general models in the new package aML (Lillard and Panis, 2000). However, Bayesian estimation using the Gibbs sampler is available for a wide variety of multilevel models—including both random-intercept and random-coefficient models—in the package BUGS (Spiegelhalter *et al.*, 1996), and convergence diagnostics can be calculated by using a set of S-PLUS functions (Best *et al.*, 1996). The current version of MLwiN (Goldstein *et al.*, 1998) provides implementations of the Metropolis algorithm as well as the parametric bootstrap procedure. These methods, however, are computationally intensive and therefore not suitable for exploratory work; the approximations studied here, and also available in MLwiN, may be invaluable in this regard.

Acknowledgements

We gratefully acknowledge support for this research from National Institute of Child Health and Human Development grants R01 HD31327 and R01 HD35277. We also thank the reviewers for their helpful comments.

References

Best, N., Cowles, M. K. and Vines, K. (1996) *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*. Cambridge: Medical Research Council Biostatistics Unit.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear models. *J. Am. Statist. Ass.*, **88**, 9–25.

Bryk, A., Raudenbush, S. and Congdon, R. (1996) *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*. Chicago: Scientific Software International.

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *Am. Statistn*, **46**, 167–174.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford: Clarendon.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **4**, 457–511.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Clarendon.

Goldman, N. and Pebley, A. R. (1994) Childhood immunization and pregnancy-related services in Guatemala. *Hlth Transtn Rev.*, **4**, 29–44.

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

——— (1995) *Multilevel Statistical Models*. London: Arnold.

——— (1996) Consistent estimators for multilevel generalized linear models using an iterated bootstrap. *Multilev. Modllng Newslett.*, **8**, 3–6.

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *J. R. Statist. Soc.* A, **159**, 505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M. (1998) *A User's Guide to MLwiN*. London: Institute of Education.

Green, P. J. (1987) Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.*, **55**, 245–259.

Guo, G. and Rodríguez, G. (1992) Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *J. Am. Statist. Ass.*, **87**, 969–976.

Kuk, A. Y. C. (1995) Asymptotically unbiased estimation in generalized linear models with random effects. *J. R. Statist. Soc.* B, **57**, 395–407.

Laird, N. M. (1978) Empirical Bayes methods for two-way contingency tables. *Biometrika*, **65**, 581–590.

Lillard, L. A. and Panis, W. A. C. (2000) *aML: Multilevel Multiprocess Statistical Software, Release 1.0*. Los Angeles: EconWare.

Longford, N. T. (1993) *Random Coefficient Models*. Oxford: Clarendon.

——— (1994) Logistic regression with random coefficients. *J. Comput. Statist. Data Anal.*, **17**, 1–15.

McCulloch, C. E. (1997) Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.*, **92**, 162–170.

Ministerio de Salud Pública y Asistencia Social and Instituto de Nutrición de Centro América y Panamá (1989) *Encuesta Nacional de Salud Materno Infantil 1987*. Guatemala City: Ministerio de Salud Pública y Asistencia Social.

Pebley, A. R. and Goldman, N. (1992) Family, community, ethnic identity and the use of formal health care services in Guatemala. *Working Paper 92–12*. Office of Population Research, Princeton University, Princeton.

Pebley, A. R., Goldman, N. and Rodríguez, G. (1996) Prenatal and delivery care and childhood immunization in Guatemala: do family and community matter? *Demography*, **33**, 231–247.

Prosser, R., Rasbash, J. and Goldstein, H. I. (1991) *ML3: Software for Three-level Analysis*. London: Institute of Education.

Raftery, A. E. and Lewis, S. M. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Clarendon.

——— (1996) Implementing MCMC. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.

Rasbash, J. and Woodhouse, G. (1995) *MLn Command Reference*. London: Institute of Education.

Roberts, G. O. (1996) Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.

Rodríguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *J. R. Statist. Soc.* A, **158**, 73–89.

Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **40**, 719–727.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996) BUGS: Bayesian inference using Gibbs sampling. Cambridge: Medical Research Council Biostatistics Unit.

Statistics and Epidemiology Research Corporation (1995) *Egret*. Seattle: Statistics and Epidemiology Research Corporation.

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

Zeger, S. L. and Karim, R. M. (1991) Generalized linear models with random effects: a Gibbs sampler approach. *J. Am. Statist. Ass.*, **86**, 79–86.

Zeger, S. L., Liang, K. and Albert, P. S. (1988) Models for longitudinal data: a generalised estimating equation approach. *Biometrics*, **44**, 1049–1060.