A Comparison of Multilevel Logistic Regression Models with Parametric and Nonparametric Random Intercepts

Olga Lukočienė *, Jeroen K. Vermunt

Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

The sensitivity of multilevel logistic regression models for misspecification of the random effects distribution is studied. More specifically, it is investigated whether using a nonparametric specification of the random effects distribution reduces bias and increases efficiency when random effects are not normally distributed. For moderate intraclass correlations, this turns out to be true as long as the level-1 sample size is not too small. However, when the level-1 sample size is very small (say 3), the standard parametric approach outperforms the nonparametric approach, even when the random effects distribution is misspecified. For small intraclass correlations, the two approaches perform equally well.

Key words: Mixed models; Hierarchical models; Multilevel logistic regression analysis; Random intercept; EM algorithm; Nonparametric maximum likelihood

1 Introduction

In the biomedical, social and behavioral sciences, it is common to collect data with a nested, multilevel, or hierarchical structure. It is therefore not surprising that the last decades there has been an increase in the use of multilevel models in these fields (Hox, 2002; Skrondal and Rabe-Hesketh, 2004; Snijders and Bosker, 1999). Examples of nested data structures include persons nested within families, pupils nested within schools, patients nested within primary

* Corresponding address: Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Tel.: +31134662544; fax: +31134663002.

Email address: o.lukociene@uvt.nl (Olga Lukočienė).

care physicians, and repeated measurements nested within subjects. In more general terms, lower-level or level-1 observational units (persons, pupils, patients, or repeated measurements) are nested within higher-level or level-2 observational units (families, schools, primary care physicians, or subjects).

Specific for multilevel data sets is that observations are correlated; that is, level-1 units (pupils, time points) belonging to the same level-2 unit (schools, subjects) tend to be more alike than level-1 units from different level-2 units. Methods for dealing with correlated data are usually classified as marginal or conditional models (Lee and Nelder , 2004). In marginal models such as the GEE approach by Zeger, Liang, and Albert (1988), the correlation between observations is treated as a nuisance factor. In contrast, in conditional models, specification of the dependence structure is part of the model building. Random effects models – sometimes also referred to as subject-specific models – belong to the family of conditional models, since results are conditional on the level-2 units' unobserved random effects. A limitation of random effects models that may be problematic in particular types of applications is that these can only capture positive associations between nested observations. Alternative conditional models which can also yield negative associations are, for example, transition models in which a person's state at a particular time point is modeled conditional on the state at the previous time point.

In this research, we focus on conditional models which use random effects. Whereas initially random effects were introduced for linear regression models, currently they were also applied in combination with the more general class of generalized linear models, yielding what is often referred to as the family of generalized linear mixed models (GLMMs) (Stiratelli *et al.*, 1984; Breslow and Clayton, 1993) or hierarchical generalized linear models (HGLMs) (Lee and Nelder , 2004). Usually the unobserved random effects are assumed to come from a particular parametric distribution, typically multivariate normal (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993). But it is clear that parametric distributional assumptions about the random effects are unlikely to hold in practice (Aitkin, 1999). Various studies found that misspecification of the distribution of random effects results in a light loss of efficiency of the regression estimators (Neuhaus *et al.*, 1992; Heagerty and Kurland, 2001; Maas and Hox, 2004).

As an alternative to using a mixing distribution from a parametric family, one may use a nonparametric specification for the random effects distribution (Laird, 1978; Heckman and Singer, 1982). This involves using a discrete mixing distribution defined by a set of unknown locations and weights to approximate an underlying continuous mixing distribution with an unknown form. Maximum likelihood (ML) estimation of the resulting finite mixture or latent class model is straightforward using the Expectation-Maximization (EM) algorithm: since the likelihood is a finite mixture no (numerical) integration is

involved. By choosing the number of latent classes to maximize the likelihood, the nonparametric maximum likelihood (NPML) estimator is obtained (Laird, 1978; Heckman and Singer, 1984; Böhning, 2000). When used in the context of regression analysis, one obtains what is sometimes referred to as a latent class or mixture regression model (Leisch , 2004; Vermunt and van Dijk, 2001; Wedel and DeSarbo, 1994).

Latent class and random coefficients regression models have always been seen as rather different approaches for dealing with dependent observations. Recently, the connection between the two approaches was stressed and it was shown that latent class regression methods cannot only be used to identify latent classes with different regression coefficients, but may also yield the typical random-coefficient modelling output; that is, estimates for the fixed and random effects (Aitkin, 1999; Hartzel *et al.*, 2001; Rabe-Hesketh *et al.*, 2005; Vermunt and van Dijk, 2001). As pointed out by Aitkin (1999), an important advantage of such a nonparametric approach is that there is no need to introduce possibly inappropriate and unverifiable assumptions about the distribution of the random effects. But this is certainly not enough to prefer this particular method, which is generally available in mixture modelling software such as GLLAMM (Skrondal and Rabe-Hesketh, 2004), Latent GOLD (Vermunt and Magidson, 2005), and Mplus (Múthen and Múthen, 1998).

Based on a limited scope simulation study for a random intercept ordinal logit model, Hartzel *et al.* (2001) concluded tentatively that the parametric approach yields more reliable estimates for both the fixed and random intercept terms, although it had some difficulties when the random effects distribution was extremely skewed. For the remaining fixed effects both approaches produce essentially unbiased estimates. They indicated that more research is needed to provide a final conclusion about the relative performance of the two methods. In contrast, based on another small simulation study for three types of GLMMs, Agresti *et al.* (2004) advised always to use a nonparametric instead of a parametric specification for the random effects distribution in order to prevent loss of efficiency. Though the simulation studies by Hartzel *et al.* (2001) and Agresti *et al.* (2004) seem to yield contradictory conclusions, closer inspection of their designs provides a possible explanation for the encountered differences. Hartzel *et al.* (2001) used small lower-level sample sizes (4 and 7) combined with a moderate higher-level sample size (100) and small values of the random effects variances. Agresti *et al.* (2004) used moderate to large lower-level sample sizes (10, 20, and 100) combined with small higher-level sample sizes (10 and 30) and moderate to large random effects variances. Our hypothesis is that the differences in conclusions are the result of these differences in simulation set up, and that lower-level and higher-level sample sizes and random effects variances should be more systematically varied to provide a complete answer.

This paper provides such a more systematic comparison of the two random effects approaches for the two-level random intercept logistic regression model. More specifically, the two research questions that are addressed are:

(1) Should the nonparametric model be preferred in situations in which underlying assumptions of the parametric model do not hold?
(2) Does it harm using a nonparametric model – say for practical reasons – when the assumptions of the parametric model hold?

A simulation study was conducted in which a broad range of data sets were generated in order to cover all typical populations in biomedical, social, and behavioral science research. More specifically, we varied the true distribution of the random effects, the size of the intraclass correlation coefficient (ICC), and the level-1 and level-2 sample sizes. We are interested in whether these simulation design factors affect the answers to our two research questions.

The next section describes the models of interest. Section 3 discusses the set up of the simulation study. Results of the simulation study are presented in Section 4. In Section 5, we present an application of the parametric and nonparametric random effects logistic regression model to a real life data set. The last section provides the reader with a discussion along with conclusions and practical recommendations.

2 The two-level random-intercept model

This section introduces two-level generalized linear models with either a parametric or a nonparametric random intercept. Let y_{ij} denote the observed response of the level-1 unit i, $i = 1, \ldots, n_j$, belonging to level-2 unit j, $j = 1, \ldots, n$, $\boldsymbol{x}_{ij}$ the vector of explanatory variables, and u_j the unobservable common random effect for all level-1 units within level-2 unit j. The vector $\boldsymbol{x}_{ij}$ may contain different types of explanatory variables; that is, variables that vary between level-1 units, between level-2 units, or between both level-1 and level-2 units, as well as (cross-level) interaction terms. In a GLMM, the conditional mean of y_{ij}, $E[y_{ij}|\boldsymbol{x}_{ij}, u_j]$, denoted by μ_{ij} is related to the linear predictor as follows:

$$g(\mu_{ij}) = \boldsymbol{\beta}'\boldsymbol{x}_{ij} + u_j, \tag{1}$$

where $g(\cdot)$ is what is referred to as the link function. Note that this is the special case in which only the intercept is random.

The typical specification for the random intercept term u_j, $j = 1, \ldots, n$, is to assume that this is an independently and identically distributed normal

random variable with mean zero and variance σ_u^2; that is, $u_j \sim N(0, \sigma_u^2)$. An equivalent alternative is to treat the mean as a free parameter and fix the β for the intercept to 0. Consistent with this distributional assumption, parameters of GLMMs are usually estimated by ML, where construction of the likelihood function is simplified by the fact that y_{ij} can be assumed to be independent within level-2 units conditionally on the observed predictors and the unobserved random effects. ML estimation involves maximizing the following marginal likelihood function:

$$L(\boldsymbol{\beta}, \sigma_u^2) = \prod_{j=1}^{n} \int_u \left[\prod_{i=1}^{n_j} f(y_{ij}|\boldsymbol{x}_{ij}, u; \boldsymbol{\beta}) \right] f(u; \sigma_u^2) du, \tag{2}$$

where $f(y_{ij}|\boldsymbol{x}_{ij}, u; \boldsymbol{\beta})$ represents the error distribution at level-1 or, equivalently, the conditional density of y_{ij}. Note that the fixed effects $\boldsymbol{\beta}$ and the variance σ_u^2 are the unknown parameters to be estimated. Except for the situation in which a continuous response variable is modelled with an identity link function and a normal level-1 error distribution, maximization of the likelihood requires the optimization of a numerically integrated likelihood. For this numerical integration, one may use a technique called Gauss-Hermite quadrature, which uses an optimal discrete approximations of the normal distribution. The most common algorithms for maximizing the resulting numerically integrated marginal likelihood are the EM algorithm (Agresti *et al.*, 2000; Bock and Aitkin, 1981; Dempster *et al.*, 1977) and gradient methods, such as the Fisher scoring (Longford, 1987) and Newton-Raphson algorithm (Pan and Thompson, 2003; Rabe-Hesketh *et al.*, 2004). In our study we used numerical integration with 50 nodes. For maximization a combination of EM and Newton-Raphson was used, where the estimation process starts with EM iterations and switches to Newton-Raphson when the relative change in the parameter values is very small.

As was indicated in the introduction, usually nothing or very little it is known about the underlying distribution of the random effects. To prevent possible misspecification, it may therefore be attractive to assume the random effects u_j to come from an unspecified mixing distribution concentrated on a finite number of latent classes or mass points (Aitkin, 1999; Böhning, 2000; Heckman and Singer, 1984; Laird, 1978). Let K denote the number of latent classes, k a particular latent class, and u_k^* the unknown value of the random effect u_j when level-2 unit j belongs to latent class k, and let $\pi_k = P(u_j = u_k^*)$ represent the probability that a randomly selected level-2 unit belongs to latent class k. Using such a K-class discrete characterization of the random effects distribution yields the following marginal likelihood function:

$$L(\boldsymbol{\beta}, \boldsymbol{u}^*, \boldsymbol{\pi}) = \prod_{j=1}^{n} \sum_{k=1}^{K} \prod_{i=1}^{n_j} f(y_{ij}|\boldsymbol{x}_{ij}, u_j = u_k^*; \boldsymbol{\beta})\, \pi_k, \tag{3}$$

where $f(y_{ij}|\boldsymbol{x}_{ij}, u_j = u_k^*; \boldsymbol{\beta})$ is the class-specific conditional density function of y_{ij}. Note that $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$, and that moreover one identifying location constraint should be imposed on the u_k^* parameters, e.g., $\sum_{k=1}^{K} u_k^* \pi_k = 0$, which implies that the u_k^* are centered. The unknown parameters to be estimated are the fixed effects $\boldsymbol{\beta}$, $K-1$ free mass point locations u_k^*, and $K-1$ free mass point weights π_k. Even though the random effects variance itself is not a model parameter, it can easily be obtained as follows: $\sigma_{u^*}^2 = \sum_{k=1}^{K} (u_k^*)^2 \pi_k$.

Maximization of the marginal likelihood function in equation (3) for a specific K can, as in the parametric case, be achieved by means of the EM and/or Newton-Raphson algorithm. The use of multiple sets of starting values is usually required because of the risk of ending up in a local maximum.

In a standard finite mixture modelling context one estimates the model of interest for different values of K and stops increasing the number of classes when the model fit no longer improves according to the BIC, AIC or another criterion. However, to obtain the solution corresponding to the NPML estimate of the random effects distribution, we not only have to maximize (3) for specific values of K, but we simultaneously have to find the value of K – say K_{NPML} – that yields the largest marginal likelihood value. In other words, we have to find the saturation point at which increasing K no longer results in an increase of the likelihood function. A method to find K_{NPML} proposed by various authors involves introducing latent classes one by one using directional (Gateaux) derivatives (Böhning, 2000; Lindsay, 1983, 1995; Rabe-Hesketh *et al.*, 2003). A much simpler alternative approach is to estimate the model with a large number of latent classes, K_{MAX}. When $K_{MAX} > K_{NPML}$, the ML estimates for u_k^* will be equal for some latent classes and/or the estimate for π_k will be equal to zero for some latent classes (Böhning, 2000). In other words, classes may be merged (equal u_k^*) and/or removed (π_k equal to zero). To prevent local maxima this procedure should be repeated with several sets of starting values. Moreover, to guarantee that also the more difficult to find mass points located at $-\infty$ and $+\infty$ are encountered when needed in the NPML solution, it is a good idea to include latent classes located at $-\infty$ and $+\infty$ in each starting set (Hartzel *et al.*, 2001; Wood and Hinde, 1987). In the dichotomous response case we will deal with in the next sections, these correspond to success probabilities equal to 0 and 1, respectively.

3 Design of the simulation study

To keep the simulation study feasible, we will restrict ourselves to one particular type of GLMM, namely to the multilevel binary logistic regression model. The reason for this choice is that whereas binary outcome variables are very

commonly used in sociological, behavioral, and biomedical studies, most attention is typically paid to models for continuous responses. Moreover, it is well documented that binary data are more sensitive to specification issues in multilevel analysis than continuous variables: in linear regression analysis, fully ignoring a random intercept does not bias parameter estimates, which is not the case in logistic regression analysis (Agresti *et al.*, 2000).

The population model we use is a two-level random-intercept logistic regression model with one level-1 and one level-2 explanatory variable; that is,

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j. \tag{4}$$

We assume that x_{1ij} – the explanatory variable for level-1 unit i in level-2 unit j – takes on the values 0 and 1 with probability 0.5, and that x_{2j} – the explanatory variable for level-2 unit j – takes on the values 0 and 1 with probability 0.5 independently of x_{1ij}. For the fixed intercept β_0 and regression slopes β_1 and β_2, we used the same values across simulation replications. More specifically, we set their values to: $\beta_0 = -2$, $\beta_1 = \beta_2 = 2$. This yields large but not too extreme differences between the response probabilities for $u_j = 0$. More specifically, the corresponding response probabilities for the four possible combination of explanatory variables are

$$\begin{aligned}
&P(y = 1|x = 1, z = 1, u = 0) = e^2/(1 + e^2) = 0.88,\\
&P(y = 1|x = 1, z = 0, u = 0) = e^0/(1 + e^0) = 0.5,\\
&P(y = 1|x = 0, x = 1, u = 0) = e^0/(1 + e^0) = 0.5
\end{aligned}$$

and

$$P(y = 1|x = 0, z = 0, u = 0) = e^{-2}/(1 + e^{-2}) = 0.12.$$

So far we discussed only the elements that were not varied in the simulations study. The factors that were varied are the specification of the random effects distribution and the level-1 and level-2 sample sizes. We wish to assess how the parametric and nonparametric models perform under different true random effects distributions and whether the performance depends on the level-1 and level-2 sample sizes.

Let us first look at the various specifications we used for the random effects distribution. We not only varied the form of the distribution, but also its variance. For the latter, it is important to note that in a logit model the level-1 errors are assumed to come from a logistic distribution with mean 0 and variance $\pi^2/3 \approx 3.29$. The ICC is therefore equal to:

$$ICC = \sigma_u^2/(\sigma_u^2 + 3.29). \tag{5}$$

Hox and Maas (2001) found that the value of the ICC may affect the accuracy of the estimates, which is why we included this factor in the simulation design. We set the ICC equal to 0.1 and 0.3, which corresponds to small and moderate values. The random effects variance σ_u^2 is easily derived from the above formula: $\sigma_u^2 = 3.29 \cdot ICC \,/\, (1 - ICC)$.

Data sets were generated using six distributional forms for the random effects, three continuous distributions – exponential, normal, and uniform – and three two-class discrete mixing distributions with membership probabilities of 0.10, 0.25, and 0.50 for the first class. With these choices we have apart from the normal distribution, distributions that considerably deviate from normal in terms of skewness, kurtosis, and discontinuity.

The other two factors that were varied are the level-1 and level-2 sample sizes. More specifically, for the number of level-2 units we used $n = 30$, 100, and 1000 and for the number of level-1 units $n_j = 3$, 10, and 50. These sample sizes were chosen to be in agreement with the simulation studies of Kreft and de Leeuw (1998) and Maas and Hox (2004), and to cover the full range of small, moderate, and large sample sizes encountered in biomedical, behavioral, and social science research. For example, in family surveys and in panel studies the combination of $n = 1000$ and $n_j = 3$ is rather common. Moreover, according to Kreft and de Leeuw (1998), $n = 30$ is the minimum number of level-2 units required for a meaningful multilevel analysis with random effects models. In organizational surveys, it is common to have about as many as 50 level-1 units within each level-2 unit, mostly combined with 30 to 100 level-2 units.

Combining the 4 design factors – ICC value, distributional form, level-2 sample size, and level-1 sample size – yields a total of $2 \times 6 \times 3 \times 3 = 108$ conditions. We generated 1000 data sets for each of these conditions. For each simulated data set, the unknown model parameters were estimated using the parametric approach assuming that random effects come from a normal distribution and using the NPML approach.

4 Results of the simulation study

The aim of the simulation study was to determine the bias and relative efficiency of the parametric and nonparametric random effects approaches under different true random effects distributions and sample sizes. Let θ be one of the parameters of interest, in our case the fixed effects β_0, β_1, and β_2, and the standard deviation of the random effects distribution σ_u, which in the nonparametric case is computed from the nodes' locations and weights. The ML estimate of θ obtained in replication s, $s = 1, \ldots, 1000$, is denoted by $\widehat{\theta}_s$. Rather than using the standard definitions of bias and relative efficiency –

Table 1
Efficiency for the conditions $n = 1000$, $ICC = 0.3$, and $n_j = 50$ or $n_j = 10$

n_j	True distribution	Model	$\lvert\hat{\beta}_{0s} - \beta_0\rvert$	$\lvert\hat{\beta}_{1s} - \beta_1\rvert$	$\lvert\hat{\beta}_{2s} - \beta_2\rvert$	$\lvert\hat{\sigma}_s - \sigma\rvert$
50	Exponential	Normal	0.05	0.02	0.05	0.11
		Nonparametric	0.04	0.02	0.03	0.06
	Normal	Normal	0.04	0.02	0.06	0.02
		Nonparametric	0.04	0.02	0.06	0.02
	Uniform	Normal	0.04	0.02	0.06	0.02
		Nonparametric	0.04	0.02	0.05	0.02
	Discrete with $\pi_1 = 0.1$	Normal	0.09	0.02	0.06	0.16
		Nonparametric	0.03	0.02	0.02	0.04
	Discrete with $\pi_1 = 0.25$	Normal	0.07	0.02	0.06	0.02
		Nonparametric	0.03	0.02	0.02	0.02
	Discrete with $\pi_1 = 0.5$	Normal	0.04	0.02	0.07	0.05
		Nonparametric	0.03	0.02	0.02	0.01
10	Exponential	Normal	0.08	0.04	0.06	0.11
		Nonparametric	0.06	0.04	0.05	0.10
	Normal	Normal	0.05	0.04	0.06	0.04
		Nonparametric	0.05	0.04	0.06	0.03
	Uniform	Normal	0.05	0.04	0.07	0.06
		Nonparametric	0.05	0.04	0.07	0.03
	Discrete with $\pi_1 = 0.1$	Normal	0.13	0.04	0.05	0.23
		Nonparametric	0.05	0.04	0.04	0.06
	Discrete with $\pi_1 = 0.25$	Normal	0.07	0.04	0.06	0.04
		Nonparametric	0.05	0.04	0.05	0.03
	Discrete with $\pi_1 = 0.5$	Normal	0.06	0.04	0.08	0.11
		Nonparametric	0.05	0.04	0.06	0.02

$E(\widehat{\theta}_s - \theta)$ and $E\left[(\widehat{\theta}_s - \theta)^2\right]$ – we used a more robust definition to prevent that the results are affected by a very small number of replications with boundary estimates. More specifically, when using the NPML estimator, especially in the conditions with large number of level-2 units and small number of level-1 units, there is a (small) positive probability that one of the latent classes is located at $-\infty$ or $+\infty$. When such boundary estimates may occur $E(\widehat{\theta}_s - \theta)$ and $E\left[(\widehat{\theta}_s - \theta)^2\right]$ do not exist. This not only to applies to σ_u, but also to β_0, β_1, and β_2. To prevent this problem from occurring we define bias as the median of $(\widehat{\theta}_s - \theta)$ and relative efficiency as the median of $|\widehat{\theta}_s - \theta|$. For similar approaches, see Agresti *et al.* (2004); Galindo-Garre *et al.* (2004).

Table 1 reports results on relative efficiency for a level-2 sample size of 1000, level-1 sample sizes of 50 and 10, and $ICC = 0.3$. It can be observed, that the assumption of a normally distributed random intercept can give a moderate loss of efficiency compared to the NPML estimator when the true distribution of random intercept is continuous but not normal. On the other hand, when the true random intercept is normal, a nonparametric approach does not yield any loss of efficiency. In all situations with a discrete true distribution, we find a considerable loss of efficiency when a misspecified parametric model is used. Though details are not provided here, very similar results were obtained for the same level-1 and ICC conditions – thus with 50 and 10 level-1 units and $ICC = 0.3$ – but with the smaller numbers of 100 and 30 level-2 units.

Table 2
Efficiency for the conditions $n_j = 3$, $ICC = 0.3$, and $n = 1000$, $n = 100$, or $n = 30$

n	True distribution	Model	$\lvert\hat{\beta}_{0s} - \beta_0\rvert$	$\lvert\hat{\beta}_{1s} - \beta_1\rvert$	$\lvert\hat{\beta}_{2s} - \beta_2\rvert$	$\lvert\hat{\sigma}_s - \sigma\rvert$
1000	Exponential	Normal	0.10	0.08	0.09	0.12
		Nonparametric	0.11	0.09	0.09	0.20
	Normal	Normal	0.07	0.08	0.10	0.07
		Nonparametric	0.08	0.08	0.10	0.11
	Uniform	Normal	0.08	0.08	0.10	0.08
		Nonparametric	0.09	0.08	0.10	0.08
	Discrete with $\pi_1 = 0.1$	Normal	0.11	0.07	0.08	0.17
		Nonparametric	0.11	0.08	0.07	0.21
	Discrete with $\pi_1 = 0.25$	Normal	0.08	0.08	0.09	0.07
		Nonparametric	0.09	0.08	0.09	0.07
	Discrete with $\pi_1 = 0.5$	Normal	0.08	0.08	0.10	0.06
		Nonparametric	0.09	0.08	0.10	0.10
100	Exponential	Normal	0.24	0.24	0.28	0.24
		Nonparametric	0.30	0.26	0.29	0.33
	Normal	Normal	0.26	0.25	0.29	0.21
		Nonparametric	0.28	0.27	0.28	0.26
	Uniform	Normal	0.24	0.25	0.29	0.21
		Nonparametric	0.29	0.27	0.31	0.24
	Discrete with $\pi_1 = 0.1$	Normal	0.25	0.23	0.27	0.33
		Nonparametric	0.29	0.24	0.26	0.54
	Discrete with $\pi_1 = 0.25$	Normal	0.25	0.25	0.30	0.22
		Nonparametric	0.29	0.25	0.30	0.25
	Discrete with $\pi_1 = 0.5$	Normal	0.25	0.25	0.31	0.22
		Nonparametric	0.28	0.26	0.32	0.22
30	Exponential	Normal	0.46	0.45	0.52	0.45
		Nonparametric	0.58	0.50	0.61	0.98
	Normal	Normal	0.46	0.47	0.54	0.42
		Nonparametric	0.59	0.54	0.64	0.68
	Uniform	Normal	0.46	0.51	0.54	0.41
		Nonparametric	0.63	0.58	0.65	0.80
	Discrete with p $(u_{01}) = 0.1$	Normal	0.45	0.47	0.54	0.57
		Nonparametric	0.58	0.53	0.59	1.19
	Discrete with p $(u_{01}) = 0.25$	Normal	0.46	0.49	0.55	0.44
		Nonparametric	0.56	0.54	0.64	0.73
	Discrete with p $(u_{01}) = 0.5$	Normal	0.46	0.48	0.58	0.42
		Nonparametric	0.61	0.54	0.65	0.59

There is no need to present all the details on the results for $ICC = 0.1$, the condition corresponding to a small level-2 variance, since these can easily be summarized. Irrespective of the level-1 and level-2 sample sizes and the form of the true random effects distribution, the parametric and nonparametric estimates are equally efficient. This holds even if the distribution of random effects is misspecified.

The efficiency estimates obtained with the smallest level-1 unit sample size ($n_j = 3$) and the largest ICC ($ICC = 0.3$) are reported in Table 2. Under these conditions, the parametric approach clearly outperforms the nonparametric approach. The former is more efficient irrespective of whether the true underlying distribution is misspecified or not. Even for the discrete true distributions, the parametric approach is the preferred one in terms of efficiency. The differences become larger as the level-2 sample size decreases and are

Table 3
Bias for the conditions $n = 1000$, $ICC = 0.3$, and $n_j = 50$ or $n_j = 10$

n_j	True distribution	Model	$\hat{\beta}_{0s} - \beta_0$	$\hat{\beta}_{1s} - \beta_1$	$\hat{\beta}_{2s} - \beta_2$	$\hat{\sigma}_s - \sigma$
50	Exponential	Normal	-0.04	0.00	0.01	-0.11*
		Nonparametric	-0.02	0.00	0.00	-0.05
	Normal	Normal	0.00	0.00	0.00	-0.01
		Nonparametric	0.00	0.00	0.00	0.00
	Uniform	Normal	-0.01	0.00	0.01	0.02
		Nonparametric	0.00	0.00	0.00	0.00
	Discrete with $p(u_1) = 0.5$	Normal	0.09	0.01	0.04	-0.16*
		Nonparametric	0.00	0.00	0.00	0.00
	Discrete with $p(u_{01}) = 0.25$	Normal	0.06	0.00	0.02	0.02
		Nonparametric	0.00	0.00	0.00	0.00
	Discrete with $p(u_{01}) = 0.1$	Normal	-0.02	0.00	0.05	0.05
		Nonparametric	0.00	0.00	0.00	0.00
10	Exponential	Normal	-0.07	0.01	0.00	-0.11*
		Nonparametric	-0.04	0.01	0.01	-0.09*
	Normal	Normal	0.00	0.00	0.01	0.00
		Nonparametric	0.00	0.00	0.01	-0.01
	Uniform	Normal	-0.01	0.00	0.02	0.06
		Nonparametric	0.00	0.01	0.01	0.01
	Discrete with $p(u_{01}) = 0.5$	Normal	0.13*	0.00	0.02	-0.23*
		Nonparametric	0.01	0.01	0.01	-0.01
	Discrete with $p(u_{01}) = 0.25$	Normal	0.06	0.00	0.00	0.02
		Nonparametric	-0.01	0.01	0.01	0.01
	Discrete with $p(u_{01}) = 0.1$	Normal	-0.03	0.01	0.05	0.11*
		Nonparametric	-0.02	0.01	0.02	0.01

* Cases with medians absolute value over 5%.

larger for σ_u than for the β parameters.

The second evaluation criterion of interest is the bias in the parameter estimates. As was indicated above, we quantified bias as the median of the difference between estimated and true parameter value across simulation replications. Table 3 provides the estimated biases of the parameter estimates for a level-2 unit sample size of 1000, level-1 unit sample sizes of 50 or 10, and $ICC = 0.3$, and Table 4 for the 3 conditions with level-1 sample size of 3 and $ICC = 0.3$. Reported biases are marked by a "*" in these two tables when they are larger than 5% of true parameter value.

The conclusions that can be derived from Tables 3 and 4 are similar to those from Tables 1 and 2. Table 3 shows that the bias of the NPML estimator is negligible for all true distributions. The parametric estimator yields biased estimates for σ_u and β_0 when the true distribution is discrete or when the true distribution is continuous but asymmetric (exponential distribution). Although not reported here, very similar results were obtained when the number of level-2 units is decreased to 100 and 30 (and level-1 and ICC settings kept constant).

As was also the case for efficiency, in the $ICC = 0.1$ conditions, the parametric and nonparametric approach perform equally well in terms of bias (results are not listed here). All biases are negligible, irrespectively of whether the random

Table 4
Bias for the conditions $n_j = 3$, $ICC = 0.3$, and $n = 1000$, $n = 100$, or $n = 30$

n	True distribution	Model	$\hat{\beta}_{0s} - \beta_0$	$\hat{\beta}_{1s} - \beta_1$	$\hat{\beta}_{2s} - \beta_2$	$\hat{\sigma}_s - \sigma$
1000	Exponential	Normal	-0.08	0.01	0.01	-0.11*
		Nonparametric	-0.08	0.02	0.02	-0.18*
	Normal	Normal	-0.01	0.01	0.01	0.00
		Nonparametric	-0.02	0.01	0.01	-0.07
	Uniform	Normal	-0.01	0.01	0.02	0.06
		Nonparametric	-0.01	0.01	0.03	0.03
	Discrete with $p(u_{01}) = 0.5$	Normal	0.02	0.01	0.01	-0.05
		Nonparametric	0.11*	0.02	0.03	-0.21*
	Discrete with $p(u_{01}) = 0.25$	Normal	0.07	0.00	0.01	0.01
		Nonparametric	0.00	0.02	0.03	-0.01
	Discrete with $p(u_{01}) = 0.1$	Normal	-0.02	0.02	0.02	0.01
		Nonparametric	-0.03	0.03	0.03	0.10*
100	Exponential	Normal	-0.07	0.02	0.00	-0.14*
		Nonparametric	-0.11*	0.06	0.06	-0.17*
	Normal	Normal	-0.02	0.03	0.03	-0.04
		Nonparametric	-0.06	0.07	0.10*	-0.06
	Uniform	Normal	-0.02	0.01	0.02	0.03
		Nonparametric	-0.07	0.05	0.10*	0.02
	Discrete with p $(u_{01}) = 0.5$	Normal	0.06	0.00	0.02	-0.29*
		Nonparametric	0.13*	0.08	0.10*	-0.34*
	Discrete with p $(u_{01}) = 0.25$	Normal	0.09	0.01	0.01	-0.04
		Nonparametric	-0.02	0.06	0.09	0.01
	Discrete with p $(u_{01}) = 0.1$	Normal	-0.02	0.03	0.04	0.06
		Nonparametric	-0.09	0.09	0.12*	0.09*
30	Exponential	Normal	-0.11*	0.05	0.01	-0.19*
		Nonparametric	-0.21*	0.19*	0.19*	-0.20*
	Normal	Normal	-0.02	0.04	0.07	-0.10*
		Nonparametric	-0.19*	0.19*	0.31*	-0.11*
	Uniform	Normal	-0.02	0.03	0.04	0.01
		Nonparametric	-0.19*	0.20*	0.24*	0.07
	Discrete with p $(u_{01}) = 0.5$	Normal	0.00	0.07	0.04	-0.48*
		Nonparametric	0.11*	0.24*	0.26*	-0.61*
	Discrete with p $(u_{01}) = 0.25$	Normal	0.09	0.08	0.02	-0.16*
		Nonparametric	-0.09	0.22*	0.25*	-0.08*
	Discrete with p $(u_{01}) = 0.1$	Normal	-0.05	0.04	0.05	0.00
		Nonparametric	-0.24*	0.18*	0.28*	0.10*

* Cases with medians absolute value over 5%.

effect distribution is correctly specified or not.

The results reported in Table 4 show that with a small level-1 sample size and large ICC, the nonparametric approach performs worse than the parametric one even when in the latter the underlying random intercept distribution is misspecified.

5 Real data example

The use of logistic regression analysis with a random intercept is illustrated with a data set from the 1988 Bangladesh Fertility Survey (Huq and Cle-

Table 5
Parameter estimates and log-likelihood values for the logistic regression models estimated with the 1988 Bangladesh Fertility Survey data set

	No random intercept		Parametric		Nonparametric	
	Coef	SE	Coef	SE	Coef	SE
Intercept	-1.568	0.126	-1.690	0.148	-1.664	.
No children	0.000	.	0.000	.		
1 child	1.059	0.152	1.109	0.158	1.100	0.159
2 children	1.288	0.167	1.377	0.175	1.368	0.176
3 or more children	1.216	0.171	1.346	0.180	1.327	0.181
Age	-0.024	0.008	-0.027	0.008	-0.026	0.008
Urban	0.797	0.105	0.732	0.120	0.719	0.122
Intercept Std. Dev.			0.464	0.079	0.472	.
ICC			0.061		0.063	
Log-likelihood	-1228.365		-1206.674		-1204.523	

land, 1990). It contains information on 1934 women who live in 60 areas of Bangladesh. It is a two-level data set: women are the level-1 units which are nested within living areas, the level-2 units. The dependent variable is the use of contraceptives yes/no. Since this is a binary response variable, it is natural to use a logit link function. Level-1 predictors are the woman's number of living children measured in four categories (no children, 1 child, 2 children, 3 or more children), and the woman's age (centered around the mean). The single level-2 predictor is type of region of residence (urban or rural). Number of living children is used as a categorical predictor, which "no children" as the reference category.

Using the Latent GOLD 4.0 software (Vermunt and Magidson, 2005), we estimated a standard logistic regression model without random effects, as well as parametric and nonparametric random intercept models. For the nonparametric model we used the "zero-inflated" option to make sure that mass points at $-\infty$ and $+\infty$ are encountered. Table 5 reports the parameter estimates and the value of the log-likelihood function for the three estimated models. Comparison of the log-likelihood values indicates that the random intercept is needed. The nonparametric specification yields a slightly larger log-likelihood value than the parametric specification. The NPML solution contains 5 mass points which are located at -1.138, -2.342, -1.608, -1.867, and $-\infty$ with weight equal to 0.350, 0.255, 0.254, 0.1299, and 0.011, respectively.

The parameter estimates obtained with the parametric and nonparametric approach are very similar in this application. This confirms the results of our simulation study in which we found that the two approaches yield almost indistinguishable results for small ICC values (note that the ICC is about 0.06 in this application). It should be noted that we excluded the mass point located at $-\infty$ and with a very small weight in the computation of the mean and the standard deviation of the intercept for the NPML solution. Inclusion of this mass point yields a mean equal to $-\infty$ and a standard deviation equal

to ∞.

6 Conclusions and discussion

The two questions that we wished to answer using the simulation study were 1) whether the NPML estimator performs better in terms of bias and efficiency compared the parametric model when the latter is misspecified, and 2) whether the NPML estimator performs equally well in terms of bias and efficiency compared the parametric model when the latter is correctly specified. This was studied for small and large level-1 and level-2 sample sizes, for small and moderate ICC values, and for different types of random effects distributions. We are now able to answer these two questions for the two-level random intercept logistic regression model.

The simulation study showed that the results depend strongly on the level-1 sample size and on the ICC values. More specifically for the larger ICC value and moderate or large level-1 sample size, we found exactly what we expected: the NPML method performs better than the parametric method when assumptions of the latter are violated and equally well when they are not violated. In such cases we should thus always use a NPML estimator since we do not know whether the assumptions hold. For small ICC values, both approaches perform equally well, so either of the two can be used in such situations. Again, it does not harm using the NPML method when the assumptions of the parametric approach hold.

In one set of conditions the NPML method turned out to be problematic; that is, when the number of level-1 units is very small ($n_j = 3$) and the ICC is not very small ($ICC = 0.3$). In these conditions the parametric approach outperformed the NPML estimator even when the true underlying distribution of random intercept was far from normal. In other words, when the number of level-1 units is very small, it is better to use a parametric random effects model.

The results of our study are in agreement with the studies by Hartzel *et al.* (2001) and Agresti *et al.* (2004), which as mentioned in the introduction, yielded seemingly contradictory results. Similar to Hartzel *et al.* (2001), we found that with small level-1 sample sizes it may be better to use a parametric random effects model, even if this misspecifies the true random effects distribution. Moreover, similar to Agresti *et al.* (2004), we found that with moderate and large level-1 sample sizes and larger ICC values, using a nonparametric approach is preferred when the underlying assumptions of the parametric model do not hold and does not harm when they hold. In other words, the level-1 sample size and the ICC value are the critical factors.

One limitation of our study is that, as in the studies by Hartzel *et al.* (2001) and Agresti *et al.* (2004), we investigated the performance of the methods only in terms of bias and efficiency of the estimated fixed and random effects parameters. We are aware of the fact that sometimes prediction of the random effects may even be more important than estimation of the fixed and random effects parameters. Another simulation study would be needed to determine how well the various methods perform in terms of prediction.

Another limitation of our study is that it concerns logistic regression models with only a random intercept. It is not clear whether our findings can be generalized to models containing also random slopes; that is, from the univariate to the multivariate random effects case. Random slopes introduce several additional complications, both in the parametric and nonparametric approach. In future research, we will investigate whether the conclusions drawn here also apply to models with random slopes.

In our study we investigated two different specifications for the random effects distribution: the parametric approach with an underlying normal distribution and the nonparametric approach using an unspecified discrete mixing distribution. As a third alternative one may use a combination of these two: a finite mixture of normal distributions (Magder and Zeger, 1996; Verbeke and Molenberghs, 2000). Whereas such an approach may have particular advantages, such as that contrary to the nonparametric approach it yields non-discrete random effects, Agresti *et al.* (2004) obtained somewhat disappointing results with this approach in the context of a log linear model for an odds ratio. Nevertheless, we believe that this hybrid approach may be promising in other situations.

Another topic that we did not address in this article is the possibility to use a semi-parametric approach in which the number of mass points is not increased till the maximum of the log-likelihood is found, but in which instead a penalized log-likelihood is maximized (or minimized). A possibility may, for example, be to select the number of mass points minimizing the Bayesian Information Criterion (BIC).

References

Agresti, A., Booth, J.G., Hobert, J.P., Caffo, B., 2000. Random-effects modeling of categorical response data. *Sociological Methodology* **30**, 27–80.

Agresti, A., Caffo, B., Ohman-Strickland, P., 2004. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis* **47**, 639–653.

Aitkin, M. 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–128.

Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters. *Psychometrika* **46**, 443–459.

Böhning, D., 2000. *Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others.* London: Chapman & Hall.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**(1), 1–38.

Galindo-Garre, F., Vermunt, J.K., Bergsma, W., 2004. Bayesian posterior estimation of logit parameters with small samples. *Sociological Methods and Research*, **39**(33), 88–117.

Hartzel, J., Agresti, A., Caffo, B., 2001 . Multinomial logit random effects models. *Statistical Modelling* **1**(2), 81–102.

Heagerty, P.J., Kurland, B.F., 2001. Misspecified maximum likelihood estimates and generalized liner mixed models. *Biometrica* **88**, 973–985.

Heckman, J.J., Singer, B., 1982. *Population heterogeneity in demographic models.* In Multidimensional Mathematical Demography, edited by K. Land and A. Rogers. New York: Academic Press.

Heckman, J.J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica* **52**, 271–320.

Hox, J., 2002. *Multilevel analysis: techniques and applications.* Mahwah, NJ: Lawrence Erlbaum.

Hox, J.J., Maas, C.J.M., 2001. The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling* **8**, 157–174.

Huq, N.M., Cleland, J., 1990. Bangladesh Fertility Survey 1989 (Main Report). Dhaka: *National Institute of Population Research and Training*

Kreft, I.G.G., de Leeuw, J., 1998. *Introducing Multilevel Modeling.* Sage, Newbury Park, CA.

Laird, N., 1978. Nonparametric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association* **73**, 805–811.

Lee, Y., Nelder, J.A., 2004. Conditional and marginal models: Another Viewn. *Statistical Science* **19**, 219–228.

Leisch, F., 2004. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software* **11 (8)**.

Lindsay, B.G., 1983. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics* **11**, 86–94.

Lindsay, B.G., 1995. *Mixture Models: Theory, Geometry and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, vol.5. Hayward, CA: Institute of Mathematical statistics.

Longford, N.T., 1987. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817–827.

Maas, C.J.M., Hox, J.J., 2004. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis* **46**, 427–440.

Magder, L.S., Zeger, S.L., 1996. A smooth nonparametric estimate of mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association* **91**, 1141–1151.

Múthen, L.K., Múthen, B.O., 1998. *Mplus User's Guide.* Los Angele: Muthen & Muthen.

Neuhaus, J.M., Hauck, W.W., Kalbfleisch, J.D., 1992. The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika* **79**, 755–762.

Pan, J.X., Thompson, R., 2003. Gauss-Hermite quadrature approximation for estimationin in generalised linear mixed models. *Computational Statistics* **18**, 57–78.

Rabe-Hesketh, S., Pickles, A., Skrondal, A., 2003. Correcting for covariate measurement error in logistic regresson using nonparametric maximum likelihood estimation. *Statistical Modelling* **3**, 215–232.

Rabe-Hesketh, S., Pickles, A., Skrondal, A., 2004. Generalized multilevel structural equation modeling. *Psychometrica* **69**, 167–190.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**, 301–323.

Skrondal, A., Rabe-Hesketh, S., 2004. *Generalized latent variables modeling: multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Snijders, T.A.B., Bosker, R.J., 1999. *Multilevel analysis.* London: Sage Publications.

Stiratelli, R., Laird, N., Ware, J.H., 1984. Random-effects models for serial observations with binary responses. *Biometrics* **40**, 961–971.

Verbeke, G., Molenberghs, G., 2000. *Linear mixed models for longitudinal data.* Springer, Berlin.

Vermunt, J.K., Magidson, J., 2005. *Technical Guide to Latent GOLD: Basic and Advanced.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., van Dijk, L., 2001. A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter* **13**, 6–13.

Wedel, M., DeSarbo, W.S., 1994. *A review of recent developments in latent class regression models.* in Advanced Methods of Marketing Research, R.P. Bagozzi, ed. Cambridge: Blackwell Publishers, 352-388.

Wolfinger, R., O'Connell, M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.

Wood, A., Hinde, J., 1987. Binomial variance component models with a nonparametric assumption concerning random effects. *In Longitudinal Data Analysis: Surrey Conference on Sociological Theory and Method 4, 110128.*

Avebury, Aldershot.

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