Biometrics 53, 1527–1537
December 1997

Random Effects Probit and Logistic Regression Models for Three-Level Data

Robert D. Gibbons and Donald Hedeker

Biometric Laboratory, University of Illinois at Chicago,
912 S. Wood Street, Chicago, Illinois 60612, U.S.A.

Summary

In analysis of binary data from clustered and longitudinal studies, random effect models have been recently developed to accommodate two-level problems such as subjects nested within clusters or repeated classifications within subjects. Unfortunately, these models cannot be applied to three-level problems that occur frequently in practice. For example, multicenter longitudinal clinical trials involve repeated assessments within individuals and individuals are nested within study centers. This combination of clustered and longitudinal data represents the classic three-level problem in biometry. Similarly, in prevention studies, various educational programs designed to minimize risk taking behavior (e.g., smoking prevention and cessation) may be compared where randomization to various design conditions is at the level of the school and the intervention is performed at the level of the classroom. Previous statistical approaches to the three-level problem for binary response data have either ignored one level of nesting, treated it as a fixed effect, or used first- and second-order Taylor series expansions of the logarithm of the conditional likelihood to linearize these models and estimate model parameters using more conventional procedures for measurement data. Recent studies indicate that these approximate solutions exhibit considerable bias and provide little advantage over use of traditional logistic regression analysis ignoring the hierarchical structure. In this paper, we generalize earlier results for two-level random effects probit and logistic regression models to the three-level case. Parameter estimation is based on full-information maximum marginal likelihood estimation (MMLE) using numerical quadrature to approximate the multiple random effects. The model is illustrated using data from 135 classrooms from 28 schools on the effects of two smoking cessation interventions.

1. Introduction

When studying medical interventions, for example, the effectiveness of a particular health care service, we must not only consider factors that influence the response process within an individual, but also consider the social structure, context, and environment in which that individual is embedded. For example, in a multicenter clinical trial, subjects within a given site are randomly assigned to treatments and prospectively studied over time. Here, data exist at three levels: measurement occasion, subject, and site. There will be correlation between the repeated experiences of an individual subject as well as correlation between the experiences of subjects within a site since they will be treated within a common therapeutic environment by the same therapists. Although it is likely that the association within individuals is stronger than the association between individuals within the same site, both components of variability are important and, as we will show, to ignore either possibility leads to invalid tests of hypotheses, inconsistent estimates of uncertainty, and misleading inferences and conclusions regarding overall significance of the intervention of interest.

A similar problem occurs in cross-sectional studies in which subjects are clustered at two levels. For example, in prevention studies, various educational programs designed to minimize risk taking behavior (e.g., smoking prevention and cessation) may be compared where randomization to various design conditions is at the level of the school and the intervention is performed at the level of the

Key words: Clustered data; Logistic regression; Longitudinal data; Multilevel data, Probit regression; Random effect models.

classroom. In this case, each subject is characterized by a single response, but subjects are nested within classrooms and classrooms are nested within schools.

A large and growing literature exists for two-level random effect regression models originally introduced by Laird and Ware (1982) and Dempster, Rubin, and Tsutakawa (1981). These models are appropriate for normally distributed measurements and parameter estimation is achieved using a mixture of empirical Bayes and maximum marginal likelihood estimation (MMLE). For linear models in the exponential family, closed form solutions to the likelihood equations exist, making computational aspects of the solution reasonably straightforward using conventional methods. In contrast, however, for nonlinear models, for example probit or logistic response functions for binary or ordinal response data, likelihood equations do not have a simple closed-form solution and the distribution of the random effects must be numerically evaluated. Several workers have developed strategies of varying computational complexity for the two-level case. For example, Gilmour, Anderson, and Rae (1985) applied the principle of generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983) to mixed models for binary response data. Gibbons and Bock (1987) developed a random effects probit model for assessing trend in correlated proportions, and Stiratelli, Laird, and Ware (1984) developed a random effects logit model for a similar application. Gibbons et al. (1994) and Gibbons and Hedeker (1994) further generalized the random effects probit model for application to multiple time-varying and time-invariant covariates and alternate response functions and prior distributions. Using quasi-likelihood methods in which no distributional form is assumed for the outcome measure, Liang and Zeger (1986) and Zeger and Liang (1986) have shown that consistent estimates of regression parameters and variance estimates can be obtained regardless of time dependence. Koch et al. (1977) and Goldstein (1991) have illustrated how random effects can be incorporated into log-linear models. Generalizations of the logistic regression model in which the values of all regression coefficients vary randomly over individuals have been proposed by Wong and Mason (1985) and Conaway (1989). More recently, Hedeker and Gibbons (1994) have extended the model of Gibbons and Bock (1987) to the case of ordinal response measures. These models are applicable to both longitudinal and clustered problems.

In contrast, very little work has been done on generalizing models for discrete response data to the case of three-level data structures. Approximate procedures by Longford (1988, 1994) and Goldstein (1991) have used first- and second-order Taylor series expansions of the logarithm of the conditional likelihood to linearize these models, which provides analytical solution of the required integration. Rodriguez and Goldman (1995) have shown that the two approaches are identical; however, the approximate solutions exhibited considerable bias and provided little advantage over use of traditional logistic regression analysis ignoring the hierarchical structure. Qaqish and Liang (1992) have applied the method of generalized estimating equations (Liang and Zeger, 1986) to models for correlated binary data with multiple levels of nesting in which marginal probabilities and odds ratios are allowed to have general regression structures.

The purpose of this paper is to extend the two-level model of Gibbons et al. (1994) to the three-level case, preserving the nonlinearity of the model and showing how the required integrals can be numerically evaluated to any practical degree of accuracy.

2. The Three-Level Probit Model

To express the three-level model in a general way, it is useful to use the following matrix representation. Stacking the unobservable latent response vectors of each subject within a cluster ($\boldsymbol{y}_{ij}$), the three-level model for the resulting N_i response vector for the ith three-level unit (classroom, clinic, etc.), $i = 1, 2, \ldots, N$, can be written as follows:

$$\underset{N_i \times 1}{\underset{\boldsymbol{y}_i}{\begin{bmatrix} \boldsymbol{y}_{i1} \\ \boldsymbol{y}_{i2} \\ \boldsymbol{y}_{i3} \\ \cdots \\ \cdots \\ \boldsymbol{y}_{in_i} \end{bmatrix}}} = \underset{N_i \times ((n_i \times r)+1)}{\underset{\boldsymbol{X}_i}{\begin{bmatrix} \mathbf{1}_{i1} & \boldsymbol{X}_{i1} & 0 & 0 & \ldots & 0 \\ \mathbf{1}_{i2} & 0 & \boldsymbol{X}_{i2} & 0 & \ldots & 0 \\ \mathbf{1}_{i3} & 0 & 0 & \boldsymbol{X}_{i3} & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{1}_{in_i} & 0 & 0 & 0 & \ldots & \boldsymbol{X}_{in_i} \end{bmatrix}}} \underset{((n_i \times r)+1) \times 1}{\underset{\boldsymbol{\beta}_i^*}{\begin{bmatrix} \beta_{0i} \\ \boldsymbol{\beta}_{i1} \\ \boldsymbol{\beta}_{i2} \\ \boldsymbol{\beta}_{i3} \\ \cdots \\ \boldsymbol{\beta}_{in_i} \end{bmatrix}}}$$

$$+\underset{N_i\times(p+1)}{\underset{\boldsymbol{W}_i}{\begin{bmatrix}\mathbf{1}_{i\,1} & \boldsymbol{W}_{i\,1}\\ \mathbf{1}_{i\,2} & \boldsymbol{W}_{i\,2}\\ \mathbf{1}_{i\,3} & \boldsymbol{W}_{i\,3}\\ \cdots & \cdots\\ \cdots & \cdots\\ \mathbf{1}_{i\,n_i} & \boldsymbol{W}_{i\,n_i}\end{bmatrix}}}\underset{(p+1)\times 1}{\underset{\boldsymbol{\alpha}}{\begin{bmatrix}\alpha_0\\ \alpha_1\\ \cdots\\ \alpha_p\end{bmatrix}}}+\underset{N_i\times 1}{\underset{\boldsymbol{e}_i}{\begin{bmatrix}\boldsymbol{e}_{i\,1}\\ \boldsymbol{e}_{i\,2}\\ \boldsymbol{e}_{i\,3}\\ \cdots\\ \cdots\\ \boldsymbol{e}_{i\,n_i}\end{bmatrix}}}, \qquad (1)$$

where the independent components are distributed as $\beta_{0\,i} \sim \mathrm{N}(0, \sigma^2_{(3)})$, $\boldsymbol{\beta}_{ij} \sim \mathrm{N}(0, \boldsymbol{\Sigma}_{(2)})$, and $\boldsymbol{e}_i \sim \mathrm{N}(0, \sigma^2_\varepsilon \boldsymbol{I}_{N_i})$. Notice there are n_i subjects within cluster i and N_i total observations within cluster i (the sum of all repeated observations for all subjects within the cluster). The number of random subject-level effects is r and the number of fixed covariates in the model (excluding the intercept) is p. In the case of binary responses, each person has an $n_{ij} \times 1$ vector $\boldsymbol{y}_{ij}$ of underlying response strengths, an $n_{ij} \times r$ design matrix $\boldsymbol{X}_{ij}$ for their r random effects $\boldsymbol{\beta}_{ij}$, and an $n_{ij} \times p$ matrix of covariates $\boldsymbol{W}_{ij}$. The covariate matrix usually includes the random effect design matrix so that the overall intercept, linear term, etc., is estimated and thus the random effects represent deviations from these overall terms.

A characteristic of the probit model is the assumption that there is an unobservable latent variable (y_{ijk}) related to the actual binary response through a threshold concept (Bock, 1975). We assume the underlying latent variable y_{ijk} is continuous and that the binary response $\mathrm{y}_{ijk} = 1$ occurs when y_{ijk} exceeds a threshold γ (i.e., $P(\mathrm{y}_{ijk} = 1) = P(y_{ijk} > \gamma)$). In terms of the latent response strength for subject j in cluster i on occasion k (y_{ijk}), we can rewrite (1) as

$$y_{ijk} = \beta_{0i} + \boldsymbol{x}'_{ijk}\boldsymbol{\beta}_{ij} + \boldsymbol{w}'_{ijk}\boldsymbol{\alpha} + \varepsilon_{ijk}. \qquad (2)$$

With the above mixed regression model for the latent variable y_{ijk}, the probability that $y_{ijk} = 1$ (a positive response occurs), conditional on $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}$, is given by

$$\begin{aligned} P(y_{ijk} = 1 \mid \boldsymbol{\beta}^*, \boldsymbol{\alpha}) &= (2\pi\sigma^2_\varepsilon)^{-1/2} \int_\gamma^\infty \exp\left[-\frac{1}{2\sigma^2_\varepsilon}(y_{ijk} - \beta_{0i} - \boldsymbol{x}'_{ijk}\boldsymbol{\beta}_{ij} - \boldsymbol{w}'_{ijk}\boldsymbol{\alpha})^2\right] dy \\ &= \Phi[-(\gamma - z_{ijk})/\sigma_\varepsilon], \end{aligned} \qquad (3)$$

where $z_{ijk} = \beta_{0i} + \boldsymbol{x}'_{ijk}\boldsymbol{\beta}_{ij} + \boldsymbol{w}'_{ijk}\boldsymbol{\alpha}$ and $\Phi(\cdot)$ represents the cumulative standard normal density function. Without loss of generality, the origin and unit of z may be chosen arbitrarily. For convenience, let $\gamma = 0$ and, to insure identifiability, let $\sigma_\varepsilon = 1$.

Let $\mathbf{y}_i$ be the vector pattern of binary responses from cluster i for the n_i individuals examined at the n_{ij} timepoints. Assuming independence of the responses conditional on the random effects, the probability of any pattern $\mathbf{y}_i$, given $\boldsymbol{\beta}^*_i$ and $\boldsymbol{\alpha}$, is equal to the product of the probabilities of the individual responses (both between and within individuals in cluster i), i.e.,

$$\ell(\mathbf{y}_i \mid \boldsymbol{\beta}^*_i, \boldsymbol{\alpha}) = \prod_{j=1}^{n_i} \prod_{k=1}^{n_{ij}} \left[\Phi(z_{ijk})\right]^{\mathrm{y}_{ijk}} \left[1 - \Phi(z_{ijk})\right]^{1-\mathrm{y}_{ijk}}. \qquad (4)$$

Then the marginal probability of $\mathbf{y}_i$ is expressed as the following integral of the likelihood, $\ell(\cdot)$, weighted by the prior density $g(\cdot)$:

$$h(\mathbf{y}_i) = \int_{\boldsymbol{\beta}^*} \ell(\mathbf{y}_i \mid \boldsymbol{\beta}^*_i, \boldsymbol{\alpha}) g(\boldsymbol{\beta}^*) d\boldsymbol{\beta}^*, \qquad (5)$$

where $g(\boldsymbol{\beta}^*)$ represents the distribution of $\boldsymbol{\beta}^*$ in the population.

Orthogonalization of the Model Parameters. For numerical solution of the likelihood equations, Gibbons and Bock (1987) orthogonally transform the response model using the Cholesky decomposition of $\boldsymbol{\Sigma}_{\beta^*}$ (Bock, 1975). Specifically, let $\boldsymbol{\beta}^* = \boldsymbol{T}^*\boldsymbol{\theta}^*$, where $\boldsymbol{T}^*\boldsymbol{T}^{*'} = \boldsymbol{\Sigma}_{\beta^*}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_{\beta^*}$. Then $\boldsymbol{\theta}^* = \boldsymbol{T}^{*-1}\boldsymbol{\beta}^*$, and so $\mathcal{E}(\boldsymbol{\theta}^*) = \mathbf{0}$ and $\mathcal{V}(\boldsymbol{\theta}^*) = \boldsymbol{T}^{*-1}\boldsymbol{\Sigma}_{\beta^*}(\boldsymbol{T}^{*-1})' = \boldsymbol{I}$. The reparameterized model is then

$$z_{ijk} = \sigma_{(3)}\theta_{0i} + \boldsymbol{x}'_{ijk}\boldsymbol{T}\boldsymbol{\theta}_{ij} + \boldsymbol{w}'_{ijk}\boldsymbol{\alpha}, \qquad (6)$$

where θ_{0i} and $\boldsymbol{\theta}_{ij}$ are the standardized random effects for cluster i and individual j in cluster i, respectively. Notice that since only a single random cluster effect is assumed, $\sigma_{(3)}$ is a scalar while $\boldsymbol{T}$ is the Cholesky (i.e., square root) factor of the $r \times r$ matrix $\boldsymbol{\Sigma}_{(2)}$. The marginal probability then becomes

$$h(\mathbf{y}_i) = \int_{\boldsymbol{\theta}^*} \ell(\mathbf{y}_i \mid \boldsymbol{\theta}^*, \boldsymbol{\alpha}) g(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*, \tag{7}$$

where $g(\boldsymbol{\theta}^*)$ is the multivariate standard normal density.

The major problem with this representation of the marginal probability is that the dimensionality of $\boldsymbol{\theta}^*$ is $(n_i \times r) + 1$ and numerical integration of equation (7) would be exceedingly slow and computationally intractable if $(n_i \times r) + 1$ is greater than 10. Note, however, that conditional on the cluster-effect $\theta_{(3)}$, the responses from the n_i subjects in cluster i are independent; therefore, the marginal probability can be rewritten as

$$h(\mathbf{y}_i) = \int_{\theta_{(3)}} \left\{ \prod_{j=1}^{n_i} \int_{\boldsymbol{\theta}_{(2)}} \left(\prod_{k=1}^{n_{ij}} \left[\Phi(z_{ijk})\right]^{1-\mathrm{y}_{ijk}} \left[1 - \Phi(z_{ijk})\right]^{\mathrm{y}_{ijk}} \right) g(\boldsymbol{\theta}_{(2)}) d\boldsymbol{\theta}_{(2)} \right\} g(\theta_{(3)}) d\theta_{(3)}, \tag{8}$$

where $\boldsymbol{\theta}_{(2)}$ are the r subject-level random effects. Here the integration is of dimensionality $r + 1$ and is tractable as long as the number of level two random effects is no greater than three or four. In longitudinal studies, we typically have one or two random effects at level two (e.g., a random intercept and/or trend for each individual) and one random effect at level three (e.g., a random cluster effect).

Estimation. The estimation of the covariate coefficients $\boldsymbol{\alpha}$ and the population parameters in $\boldsymbol{T}$ requires differentiation of the log likelihood function with respect to these parameters. The log likelihood for the patterns from the N clusters can be written as

$$\log L = \sum_{i}^{N} \log h(\mathbf{y}_i). \tag{9}$$

Let $\boldsymbol{\eta}$ represent an arbitrary parameter vector; then for $\boldsymbol{\alpha}$, and the unique elements of the Cholesky factor $\boldsymbol{T}$, we get

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\eta}} &= \sum_{i=1}^{N} h^{-1}(\mathbf{y}_i) \int_{\theta_{(3)}} E_{i\theta_{(3)}} \\ &\quad \times \left\{ \sum_{j=1}^{n_i} (e_{ij\theta_{(3)}})^{-1} \int_{\boldsymbol{\theta}_{(2)}} \sum_{k=1}^{n_{ij}} \left(\frac{y_{ijk} - \Phi(z_{ijk})}{\Phi(z_{ijk})[1 - \Phi(z_{ijk})]} \right) L_{ij}(\boldsymbol{\theta}) \phi(z_{ijk}) \frac{\partial z_{ijk}}{\partial \boldsymbol{\eta}} g(\boldsymbol{\theta}_{(2)}) d\boldsymbol{\theta}_{(2)} \right\} \\ &\quad \times g(\theta_{(3)}) d\theta_{(3)},\end{aligned} \tag{10}$$

where

$$L_{ij}(\boldsymbol{\theta}) = \prod_{k=1}^{n_{ij}} \left[\Phi(z_{ijk})\right]^{1-\mathrm{y}_{ijk}} \left[1 - \Phi(z_{ijk})\right]^{\mathrm{y}_{ijk}}, \tag{11}$$

$$\begin{aligned} E_{i\theta_{(3)}} &= \prod_{j=1}^{n_i} \int_{\boldsymbol{\theta}_{(2)}} L_{ij}(\boldsymbol{\theta}) g(\boldsymbol{\theta}_{(2)}) d\boldsymbol{\theta}_{(2)} \\ &= \prod_{j=1}^{n_i} e_{ij\theta_{(3)}}, \end{aligned} \tag{12}$$

and

$$\frac{\partial z_{ijk}}{\partial \boldsymbol{\alpha}'} = \boldsymbol{w}'_{ijk} \qquad \frac{\partial z_{ijk}}{\partial (\mathrm{v}(\boldsymbol{T}))'} = (\boldsymbol{\theta}'_{(2)} \otimes \boldsymbol{x}'_{ijk}) \mathsf{J}'_r \qquad \frac{\partial z_{ijk}}{\partial \sigma_{(3)}} = \theta_{(3)},$$

where $\mathrm{v}(\boldsymbol{T})$ contains the unique elements of the Cholesky factor $\boldsymbol{T}$ and J_r is the transformation matrix of Magnus (1988) that eliminates the elements above the main diagonal.

As in the two-level case described by Gibbons and Bock (1987) and Gibbons et al. (1994), the method of scoring can be used to provide MMLEs and numerical integration on the transformed $\boldsymbol{\theta}$ space can be performed (Stroud and Sechrest, 1966). An advantage of numerical integration is that alternative prior distributions for the random effects can be considered. Thus, for example, we can compare parameter estimates for a normal versus rectangular prior to determine the degree to which our estimates are robust to deviation from the assumed normality of the prior distribution for the random effects.

3. A Three-Level Logistic Regression Model

In the previous discussion, we have focused on a three-level random effects probit regression model; however, many researchers are more familiar with the logistic regression model. Fortunately, modification of the response function and associated likelihood equations is trivial as Gibbons et al. (1994) and Hedeker and Gibbons (1994) have shown for two-level logistic regression models.

Following Gibbons et al. (1994), we replace the normal response function $\Phi(z_{ijk})$ with

$$\Psi(z_{ijk}) = \frac{1}{1 + \exp[-z_{ijk}]} \tag{13}$$

and the normal density function $\phi(z_{ijk})$ with the product

$$\Psi(z_{ijk})(1 - \Psi(z_{ijk})). \tag{14}$$

As in the normal case, we let $\gamma = 0$; however, the residual variance corresponding to the standard logistic distribution is $\pi^2/3$. Application of the logistic response function is attractive in many cases in which the response probability is small because the logistic distribution has greater tail probability than the normal distribution.

4. Illustration

The Television School and Family Smoking Prevention and Cessation Project (TVSFP) study (Flay et al., 1988) was designed to test independent and combined effects of a school-based social resistance curriculum and a television-based program in terms of tobacco use prevention and cessation. The study involved seventh grade students from 135 classrooms from 28 schools, where the schools were randomized to one of four study conditions: (a) a social resistance classroom curriculum, (b) a media (television) intervention, (c) a social resistance classroom curriculum combined with a mass-media intervention, and (d) a no-treatment control group. These conditions form a 2×2 design of social resistance classroom curriculum (CC = yes or no) by mass-media intervention (TV = yes or no). A tobacco and health knowledge scale (THKS) was used in classifying subjects as knowledgeable or not. Data from 1600 students with pre- and postintervention data were available. The resulting dataset was unbalanced with 1 to 13 classrooms per school and 2 to 28 students per classroom. Student frequencies for positive and negative THKS results, broken down by condition subgroups, are given in Table 1.

Two three-level probit regression models were fit to these data. In the first analysis, pre- and postintervention THKS responses were treated as a within-subject effect (i.e., level two) and class-

Table 1
Tobacco and health knowledge scale postintervention results subgroup frequencies (and percentages)

Subgroup		THKS score		
CC	TV	Pass	Fail	Total
No	No	175 (41.6)	246 (58.4)	421
No	Yes	201 (48.3)	215 (51.7)	416
Yes	No	240 (63.2)	140 (36.8)	380
Yes	Yes	231 (60.3)	152 (39.7)	383
Total		847 (52.9)	753 (47.1)	1600

Table 2

Tobacco and health knowledge scale pre- vs. postntervention (binary) scores comparison of two- and three-level random effect probit model estimates (standard errors)

	Fixed effect model	Two-level REPM class	Two-level REPM student	Three-level REPM
Constant	−0.3116***	−0.3377***	−0.3778***	−0.3916***
	(0.062)	(0.104)	(0.076)	(0.116)
Pre- vs. post	0.0986	0.1007	0.1198	0.1181
	(0.088)	(0.103)	(0.098)	(0.100)
CC (pre)	−0.1095	−0.1427	−0.1333	−0.1608
	(0.091)	(0.135)	(0.110)	(0.151)
TV (pre)	−0.0776	−0.1154	−0.0918	−0.1251
	(0.089)	(0.129)	(0.107)	(0.144)
CC × TV (pre)	−0.0026	0.0523	0.0030	0.0534
	(0.129)	(0.184)	(0.157)	(0.208)
CC (post–pre)	0.6585***	0.6922***	0.7995***	0.8065***
	(0.128)	(0.143)	(0.144)	(0.143)
TV (post–pre)	0.2484**	0.2632**	0.2997**	0.3045***
	(0.124)	(0.123)	(0.137)	(0.121)
CC × TV (post–pre)	−0.2428	−0.2691	−0.3066	−0.3204*
	(0.181)	(0.184)	(0.197)	(0.182)
Class S.D.		0.2872***		0.3103***
		(0.039)		(0.049)
Student S.D.			0.6798***	0.6048***
			(0.059)	(0.066)
Log L	−2108.63	−2082.12	−2077.89	−2061.92

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

room was treated as a cluster (level three). In the second analysis, postintervention THKS scores were considered and subjects were clustered within classrooms (level two) and schools (level three). In both cases, THKS knowledge was modeled in terms of CC, TV, and CC by TV interaction. In the second model, preintervention knowledge was also used as a covariate.

Results from the first analysis are presented in Table 2. The first column of Table 2 lists results for a probit regression analysis of student-level data ignoring clustering of students and treating each student and measurement occasion as an independent observation. This analysis indicates the positive effect of the social resistance classroom curriculum as well as the television part of the intervention, but it indicates no interaction. As compared to the random effect models, the standard errors are clearly underestimated, suggesting that we have greater precision than is actually the case given the dependence of measurements within students and students within classrooms. Results are somewhat similar for the two two-level models, where standard errors for the model with a subject random effect are somewhat smaller than those for the model with classroom random effect. For the three level model (i.e., measurements within subjects and subjects within classrooms), the standard errors are typically the largest, as expected. However, the three-level model indicates an even stronger effect for the TV intervention and a result that approaches significance for the interaction. The three-level analysis suggests that, while TV intervention is effective in increasing THKS scores for those not receiving the CC component, it has a slight negative effect on those exposed to both components (see Table 1). Similarly, while the CC intervention is effective in increasing THKS scores, the effect is more pronounced for those not receiving the TV component than for those receiving it. Both two-level models provide significant improvement in fit relative to the fixed effect model ($\chi^2_1 = 53.02$, $p < .0001$ and $\chi^2_1 = 61.48$, $p < .0001$ for classroom- and student-level random effect models, respectively). Similarly, the three-level model provided significant improvement in fit relative to the two two-level models ($\chi^2_1 = 40.40$, $p < .0001$ and $\chi^2_1 = 31.946$, $p < .0001$ for classroom- and student-level random effect models, respectively). A three-level model with two random effects at the subject level (i.e., baseline and post–pre change) did not significantly improve the fit relative to the three-level model with a single subject level random effect ($\chi^2_2 = 0.06$, $p =$ n.s.). The intraclass (classroom) correlation equals .066 (i.e., $.3103^2/[.3103^2 + .6048^2 + 1]$) and the intrastudent correlation equals .250 (i.e., $.6048^2/[.3103^2 + .6048^2 + 1]$). Thus, approximately 6.6% of the variance is attributable to classrooms and 25.0% is attributable to students.

To aid in interpreting the estimated model parameters, estimated response proportions are computed from the latent response vectors $\boldsymbol{y}_i$, which are normally distributed with mean $\boldsymbol{W}_i\boldsymbol{\alpha}$ and variance–covariance matrix $\boldsymbol{X}_i\boldsymbol{\Sigma}_{\beta^*}\boldsymbol{X}_i' + \sigma_\varepsilon \boldsymbol{I}_{N_i}$. For example, for students in the no-treatment control condition, the estimated response proportion equals $\Phi(-.3916/(1+.3103^2+.6048^2)^{1/2}) =$ .373 at preintervention and $\Phi((-.3916+.1181)/(1+.3103^2+.6048^2)^{1/2}) = .397$ at postintervention. Similar pre- and postintervention estimates are .324 and .621 for CC, .335 and .469 for TV, and .303 and .593 for CC plus TV. These estimates corroborate what the estimated model parameters indicate, namely, that there is a considerable difference in change between controls and CC students and less of a difference comparing controls to either TV or CC plus TV students. Also note the close agreement with the observed postintervention response proportions listed in Table 1.

Results from the second analysis (i.e., random class and school effects on postintervention THKS knowledge) are presented in Table 3. The picture is somewhat different when analysis focuses on postintervention THKS knowledge. Here, the fixed effect model identified a significant main effect of TV and a CC by TV interaction; however, neither of these two effects were significant for two-level or three-level random effect models. All three random effects models (i.e., the two-level classroom, two-level school, and three-level school by classroom models) provided significant improvement in fit relative to the fixed effect model ($\chi^2_1 = 18.80$, $p < .0001$; $\chi^2_1 = 12.60$, $p < .0004$; and $\chi^2_2 = 20.62$, $p < .0001$, respectively); however, the three-level model did not provide significant improvement in fit relative to the two-level (classroom) model ($\chi^2_1 = 1.82$, $p =$ n.s.). Estimates of the intraunit correlations based on the three-level model are .059 for classrooms and .026 for schools.

Tables 4 and 5 present results for three-level models with alternate response function (i.e., logistic) and prior distribution (normal versus rectangular). Table 4 displays results for the first example (i.e., clustered and longitudinal—class and pre- versus postintervention), whereas Table 5 displays results for the second example (i.e., two levels of clustering—school and classroom).

Comparison of Tables 2 and 3 with Tables 4 and 5 reveal that probit and logistic regression models lead to virtually identical conclusions regarding significance of fixed and random effects in the three-level models. The MMLEs and standard errors for the fixed effects are approximately 40% larger for the logistic regression model; however, estimates of the random effect standard deviations were quite similar when both models used a normal prior distribution. Fit of the probit and logistic models were virtually identical.

Table 3

Tobacco and health knowledge postintervention (binary) scores comparison of two- and three-level random effect probit model estimates (standard errors)

	Fixed effect model	Two-level REPM class	Two-level REPM school	Three-level REPM
Constant	-0.4209^{***}	-0.4412^{***}	-0.4394^{***}	-0.4410^{***}
	(0.069)	(0.100)	(0.153)	(0.154)
Pre	0.5335^{***}	0.5310^{***}	0.5106^{***}	0.5211^{***}
	(0.068)	(0.071)	(0.075)	(0.079)
CC	-0.5913^{***}	0.5916^{***}	0.6698^{***}	0.6277^{***}
	(0.092)	(0.128)	(0.226)	(0.228)
TV	0.1931^{**}	0.1718	0.2369	0.2078
	(0.089)	(0.124)	(0.172)	(0.171)
CC × TV	-0.2596^{**}	−0.2224	−0.3591	−0.2921
	(0.128)	(0.178)	(0.259)	(0.260)
Class S.D.		0.3005^{***}		0.2532^{***}
		(0.059)		(0.086)
School S.D.			0.2143^{***}	0.1690^{**}
			(0.071)	(0.086)
Log L	−1050.26	−1040.86	−1043.96	−1039.95

$^{***}\,p < 0.01$, $^{**}\,p < 0.05$, $^{*}\,p < 0.10$.

Table 4
Tobacco and health knowledge scale pre- vs. postintervention (binary) scores three-level random effect logistic regression models estimates (standard errors)

	Normal Prior	Rectangular prior
Constant	−0.6533***	−0.6681***
	(0.196)	(0.201)
Pre- vs. post	0.1981	0.1982
	(0.168)	(0.167)
CC (pre)	−0.2716	−0.2622
	(0.254)	(0.252)
TV (pre)	−0.2129	−0.1680
	(0.243)	(0.243)
CC × TV (pre)	0.0863	0.0416
	(0.350)	(0.349)
CC (post–pre)	1.3450***	1.3408***
	(0.241)	(0.240)
TV (post–pre)	0.5095***	0.5093***
	(0.203)	(0.203)
CC × TV (post–pre)	−0.5242**	−0.5148**
	(0.306)	(0.305)
Class S.D.	0.5228***	0.1727***
	(0.082)	(0.024)
Student S.D.	1.0135***	0.3230***
	(0.115)	(0.024)
Log L	−2061.68	−2061.48

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

When a rectangular prior was used, estimates of the fixed effects and their standard errors exhibited little if any effect; however, MMLEs and standard errors of the random effects decreased by approximately two thirds. In both examples, fit of the models with alternative prior distributions was virtually identical.

Taken as a whole, these results are quite encouraging because they demonstrate that, while a normal prior distribution is a good choice, fit of the model, parameter estimates, and standard er-

Table 5
Tobacco and health knowledge scale postintervention (binary) scores three-level random effect logistic regression models estimates (standard errors)

	Normal prior	Rectangular prior
Constant	−0.7172***	−0.7365***
	(0.252)	(0,266)
Pre	0.8502***	0.8498***
	(0.131)	(0.130)
CC	1.0179***	1.0585***
	(0.375)	(0.391)
TV	0.3345	0.3458
	(0.280)	(0.290)
CC × TV	−0.4625	−0.5046
	(0.426)	(0.442)
Class S.D.	0.4154***	0.1381***
	(0.141)	(0.040)
School S.D.	0.2742**	0.0914***
	(0.140)	(0.039)
Log L	−1039.97	−1039.75

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

rors of at least the structural parameters are reasonably robust to model misspecification. As in the case of fixed effect models, selection of probit versus logistic response functions appears to have more to do with custom or practice within a particular discipline than differences in statistical properties.

5. Discussion

Random effect models can be extremely useful in analysis of discrete clustered multivariate data. This paper extends these models to analysis of three-level data (i.e., two levels of clustering). The three-level random effects probit and logistic regression models presented here provide one solution to this problem. The approach taken here advances previous work of Longford (1988, 1994) and Goldstein (1991), which provide approximate solutions based on first- and second-order Taylor series expansions of the logarithm of the conditional likelihood required to linearize these models so that closed-form solutions of the likelihood equations exist. In practice, most three-level clustered problems involving dichotomous responses will require only two random effects (i.e., one for each level of clustering). For longitudinal data, typically one to three random effects are required to model the time trends (e.g., random intercept and linear trend), leaving one additional random effect due to clustering. In all of these cases, the method described here is appropriate and numerical evaluation of the likelihood equations is computationally tractable. As an example, all of the illustrations described here were run on a 90-MHz Pentium processor in 15 minutes or less. An additional advantage of the numerical solution is that it can accommodate alternate prior distributions, including empirical estimation of the prior distribution (Bock and Aitkin, 1981).

Alternatively, the models presented here and even more complicated models could be evaluated using Gibbs sampling (Geman and Geman, 1984; Gelfand et al., 1990b; Gelfand and Smith, 1990; Tanner, 1991). Although potentially more computationally intensive, the Gibbs sampler should have no practical limitation on number of random effects, whereas the numerical solution is computationally intractable for models with more than five or six random effects.

The three-level model presented in this article can be extended for ordinal responses in a similar manner as our ordinal extension of the two-level model (Hedeker and Gibbons, 1994). For an ordinal response with C categories, a series of threshold values $(\gamma_1, \gamma_2, \dots, \gamma_{C-1})$ are added to the model. A response then occurs in category c if the latent response process y exceeds the threshold value γ_{c-1} but does not exceed the threshold value γ_c. The effects of the model covariates are assumed to be the same for each threshold, and the model can be specified using either the probit or logistic formulation.

Acknowledgements

This work was supported by a Research Scientist Award from the National Institute of Mental Health (K05-MH01254) and by National Institute of Mental Health grant R01-MH48153. The authors benefitted greatly from discussions with Professor R. Darrell Bock and would like to thank Professor Brian Flay for use of the example data. We also thank the editor, associate editor, and an anonymous reviewer for several helpful comments and suggestions.

Résumé

Des modèles à effet aléatoire ont été récemment développés pour l'analyse de données binaires récoltées lors d'études "clusterisées" et longitudinales; ceux-ci permettent de traiter des problèmes à deux niveaux, par exemple sujets répartis en différents clusters ou classifications répétées par sujet. Malheureusement, ces modèles ne s'appliquent pas aux problèmes à trois niveaux que l'on rencontre fréquemment en pratique. Par exemple, les essais cliniques multicentriques longitudinaux impliquent des mesures répétées pour chaque sujet et ces sujets sont répartis dans différents sites. Cette combinaison de données "clusterisées" et longitudinales constitue classiquement le problème à trois niveaux en biométrie. De même, dans les études de prévention, on compare plusieurs programmes d'éducation à la santé visant à diminuer le risque d'assuétude (par exemple, prévention et arrêt du tabagisme) et pour lesquels la randomisation des conditions expérimentales s'effectue au niveau des écoles alors que la prévention est menée au niveau des classes. Par le passé, les méthodes statistiques applicables aux données binaires collectées dans les problèmes à trois niveaux ont, soit ignoré un des niveaux de regroupement, traité comme un effet fixe, soit utilisé des développements en série de Taylor de premier et second degrés du logarithme de la vraisemblance conditionnelle afin de linéariser ces modèles et estimer leurs paramètres par les approches classiquement utilisées pour les données quantitatives. Des études récentes ont montré que ces solutions approximatives comportaient un biais considérable et présentaient peu d'avantages par rapport à la régression logistique classique qui ignore la structure hiérarchique des données. Dans cet article, nous généralisons au cas de trois niveaux les résultats obtenus pour les modèles probit et logistique à effets aléatoires à deux niveaux. L'estimation des paramètres est obtenue à partir de la maximisation de la vraisemblance marginale complète, en ayant recours à des procédés d'intégration numérique afin de tenir compte des effets aléatoires multiples. Afin d'illustrer la méthode, nous l'avons appliquée aux données collectées dans 135 classes de 28 écoles dans le cadre d'une étude visant à comparer deux programmes d'arrêt du tabagisme.

REFERENCES

Bock, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill.

Bock, R. D. and Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika* **46,** 443–459.

Conaway, M. R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association* **84,** 53–61.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association* **76,** 341–353.

Flay, B., Brannon, B., Johnson, C., Hansen, W., Ulene, A., Whitney-Saltiel, D., Gleason, L., Sussman, S., Gavin, M., Kimarie, G., Sobol, D., and Spiegel, D. (1988). The Television School and Family Smoking Prevention and Cessation Project: I. Theoretical basis and program development. *Preventive Medicine* **17,** 585–607.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculation of marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85,** 972–985.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6,** 721–741.

Gibbons, R. D. and Bock, R. D. (1987). Trend in correlated proportions. *Psychometrika* **52,** 113–124.

Gibbons, R. D. and Hedeker, D. (1994). Application of random effects probit regression models. *Journal of Clinical and Consulting Psychology* **62,** 285–296.

Gibbons, R. D., Hedeker, D., Charles, S. C., and Frisch, P. (1994). A random effects probit model for predicting medical malpractice claims. *Journal of the American Statistical Association* **89,** 760–767.

Gilmour, A. R., Anderson, R. D., and Rae, A. L. (1985). The analysis of binomial data by generalized linear mixed model. *Biometrika* **72,** 593–599.

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

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

Koch, G., Landis, J., Freeman, J., Freeman, H., and Lehnen, R. (1977). A general methodology for the analysis of experiments with repeated measurements of categorical data. *Biometrics* **33,** 133–158.

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

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

Longford, N. T. (1988). A quasi-likelihood adaptation for variance components analysis. *Proceedings of the Statistical Computing Section of the American Statistical Association.* American Statistical Association, Alexandria, VA.

Longford, N. T. (1994). Logistic regression with random coefficients. *Computational Statistics Data Analysis* **17,** 1–15.

Magnus, J. R. (1988). *Linear Structures.* London: Charles Griffin.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.

Qaqish, B. F. and Liang, K. Y. (1992). Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. *Biometrics* **48,** 939–950.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary response. *Journal of the Royal Statistical Society, Series A* **158,** 73–89.

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

Stroud, A. H. and Sechrest, D. (1966). *Gaussian Quadrature Formulas.* Englewood Cliffs, New Jersey: Prentice Hall.

Tanner, M. A. (1991). *Tools for Statistical Inference.* New York: Springer-Verlag.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association* **80,** 513–524.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

Received March 1996; *revised January and May* 1997; *accepted June* 1997.