STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:1433–1446 (DOI: 10.1002/sim.1522)

A mixed-effects multinomial logistic regression model

Donald Hedeker*,†

Division of Epidemiology & Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, U.S.A.

SUMMARY

A mixed-effects multinomial logistic regression model is described for analysis of clustered or longitudinal nominal or ordinal response data. The model is parameterized to allow flexibility in the choice of contrasts used to represent comparisons across the response categories. Estimation is achieved using a maximum marginal likelihood (MML) solution that uses quadrature to numerically integrate over the distribution of random effects. An analysis of a psychiatric data set, in which homeless adults with serious mental illness are repeatedly classified in terms of their living arrangement, is used to illustrate features of the model. Copyright © 2003 by John Wiley & Sons, Ltd.

KEY WORDS: nominal data; ordinal data; categorical data; multilevel data; logistic regression; maximum marginal likelihood; quadrature; clustering; repeated observations

1. INTRODUCTION

Nominal or polytomous response data are common in many fields of research. For example, the variable 'type of service use' is an outcome variable that is often measured in health services research. Types of health services utilization can include medical provider visit, hospital outpatient visit, emergency room visit, hospital inpatient stay and home health care visit. If observations are independent, the multinomial or polychotomous logistic regression model [1–5] can be used to assess the influence of explanatory variables on the nominal response variable. It is often the case, however, that subjects are observed nested within clusters (for example, schools, firms, clinics) or are repeatedly measured. In this case, use of the ordinary multinomial logistic regression model assuming independence of observations is problematic, since observations from the same cluster or subject are usually correlated.

For data that are clustered (or longitudinal), mixed-effects regression models have become increasingly popular, and several books have been written on this topic [6–8]. For dichotomous response data, several approaches adopting either a logistic or probit regression model and

* Correspondence to: Donald Hedeker, Division of Epidemiology & Biostatistics (M/C 923), School of Public Health, University of Illinois at Chicago, 1603 West Taylor Street, Room 955, Chicago, IL 60612-4336, U.S.A.
† E-mail: hedeker@uic.edu

Contract/grant sponsor: National Institutes of Mental Health; contract/grant number: MH56146.

various methods for incorporating and estimating the influence of the random effects have been developed [9–14]. Several articles [15–18] have discussed and compared some of these models and their estimation procedures. Extending these methods to ordinal response data has also been actively pursued [19–25].

For clustered nominal responses, there have been some developments as well. An early example is the model for nominal educational test data described by Bock [26]. This model includes a random subject effect and fixed item parameters for the item responses that are clustered within subjects. While Bock's model is a full-information maximum likelihood approach, using Gauss–Hermite quadrature to integrate over the random-effects distribution, it does not include covariates or multiple random effects. More general regression models of clustered nominal data have been considered by Goldstein (reference [6], chapter 7), Daniels and Gatsonis [27], and Revelt and Train [28]. These approaches use either more approximate or Bayesian methods to handle the integration over the random effects. Also, these models generally adopt a reference cell approach for modelling the nominal response variable in which one of the categories is chosen as the reference cell and parameters are characterized in terms of the remaining $C-1$ comparisons to this reference cell. Bock's model, alternatively, was written in terms of any set of $C-1$ comparisons across the nominal response categories. A recent paper by Hartzel *et al.* [29] synthesizes much of the work in this area, describing a general mixed-effects model for both clustered ordinal and nominal responses.

In this paper, a mixed-effects multinomial logistic regression model will be described that is appropriate for either clustered or longitudinal response data. This model will accommodate multiple random effects and, additionally, allow for a general form for model covariates. In terms of comparisons across the nominal outcome categories, both reference cell and more general category comparisons are described. A full maximum marginal likelihood solution is outlined for parameter estimation. In this solution, multi-dimensional quadrature is used to numerically integrate over the distribution of random effects, and an iterative Fisher scoring algorithm is used to solve the likelihood equations. An example of an analysis of longitudinal data will illustrate features of the mixed-effects model for nominal response data.

2. MIXED-EFFECTS MULTINOMIAL REGRESSION MODEL

Using the terminology of multilevel analysis [6], let i denote the level-2 units (clusters) and let j denote the level-1 units (nested observations). Assume that there are $i=1,\ldots N$ level-2 units and $j=1,\ldots,n_i$ level-1 units nested within each level-2 unit. Let y_{ij} be the value of the nominal variable associated with level-2 unit i and level-1 unit j. In the nominal case, we need to consider the values corresponding to the unordered multiple categories of the response variable. For this, let us assume that the C response categories are coded as $c=1,2,\ldots,C$.

Adding random effects to the usual multinomial logistic regression model, the probability that $y_{ij}=c$ (a response occurs in category c) for a given level-2 unit i, conditional on the random effects $\boldsymbol{\beta}$, is given by

$$p_{ijc}=P(y_{ij}=c\,|\,\boldsymbol{\beta})=\frac{\exp(z_{ijc})}{1+\sum_{h=1}^{C}\exp(z_{ijh})}\quad\text{for } c=2,3,\ldots,C \tag{1}$$

$$p_{ij1}=P(y_{ij}=1\,|\,\boldsymbol{\beta})=\frac{1}{1+\sum_{h=1}^{C}\exp(z_{ijh})} \tag{2}$$

where $z_{ijc} = \mathbf{w}'_{ij}\boldsymbol{\alpha}_c + \mathbf{x}'_{ij}\boldsymbol{\beta}_{ic}$. Here, $\mathbf{w}_{ij}$ is the $s \times 1$ covariate vector and $\mathbf{x}_{ij}$ is the design vector for the r random effects, both vectors being for the jth level-1 unit nested within level-2 unit i. Correspondingly, $\boldsymbol{\alpha}_c$ is an $s \times 1$ vector of unknown fixed regression parameters, and $\boldsymbol{\beta}_{ic}$ is an $r \times 1$ vector of unknown random effects for the level-2 unit i. The distribution of the random effects is assumed to be multivariate normal with zero mean vector and covariance matrix $\boldsymbol{\Sigma}_c$; it will be indicated later how this assumption can be relaxed.

It is convenient to standardize the random effects. For this, let $\boldsymbol{\beta}_{ic} = \mathbf{T}_c\boldsymbol{\theta}_i$, where $\mathbf{T}_c\mathbf{T}'_c = \boldsymbol{\Sigma}_c$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_c$. The reparameterized model is then

$$z_{ijk} = \mathbf{w}'_{ij}\boldsymbol{\alpha}_c + \mathbf{x}'_{ij}\mathbf{T}_c\boldsymbol{\theta}_i$$

As discussed by Bock [26], the model has a plausible interpretation. Namely, each nominal category is assumed to be related to an underlying latent 'response tendency' for that category. The category c associated with the observed response y_{ij} is then the category for which the response tendency is maximal. These latent response tendencies are assumed to be independently distributed following approximately normal distributions (that is, logistic distributions). As is well known, the logistic closely resembles the normal distribution, the primary difference being that the logistic places more probability in the tails of the distribution.

The model can also accommodate separate (that is, independent) random-effect variance terms for groups of either i or j units. For example, suppose that there is interest in allowing varying random-effect variance terms by gender. For this, $\mathbf{x}_{ij}$ is specified as a 2×1 vector of dummy codes indicating male and female membership, respectively. $\mathbf{T}_c$ is then a 2×1 vector of independent random-effect standard deviations for males and females, and the subject effect θ_i is a scalar that is pre-multiplied by the vector $\mathbf{T}_c$. This is also useful for educational testing models [26] where n item responses ($j = 1, 2, \ldots n$) are nested within N subjects ($i = 1, 2, \ldots N$) and a separate random-effect standard deviation (that is, an element of the $n \times 1$ vector $\mathbf{T}_c$) is estimated for each test item (that is, each j unit). Again, this is accomplished by specifying $\mathbf{x}_{ij}$ as an $n \times 1$ vector of dummy codes indicating the repeated items. For both cases, $\mathbf{T}_c$ is an $r \times 1$ vector that is pre-multiplied by the transpose of an $r \times 1$ vector of indicator variables $\mathbf{x}_{ij}$, and so $\mathbf{T}_c$ pre-multiplies a scalar random effect θ_i (instead of an $r \times 1$ vector of random effects $\boldsymbol{\theta}_i$).

2.1. More general category contrasts

The model as written above allows estimation of any pairwise comparisons among the C response categories. As characterized in Bock [26], it is beneficial to write the nominal model to allow for any set of $C - 1$ non-redundant contrasts. For this, the category probabilities are written as

$$p_{ijc} = \frac{\exp(z_{ijc})}{\sum_{h=1}^{C} \exp(z_{ijh})} \quad \text{for } c = 1, 2, \ldots, C \tag{3}$$

where now

$$z_{ijc} = \mathbf{w}'_{ij}\boldsymbol{\Gamma}\mathbf{d}_c + (\mathbf{x}'_{ij} \otimes \boldsymbol{\theta}'_i)\mathbf{J}'_{r*}\boldsymbol{\Lambda}\mathbf{d}_c \tag{4}$$

Here, $\mathbf{D}$ is the $(C-1) \times C$ matrix containing the contrast coefficients for the $C - 1$ contrasts between the C logits and $\mathbf{d}_c$ is the cth column vector of this matrix. The $s \times (C-1)$ parameter matrix $\boldsymbol{\Gamma}$ contains the regression coefficients associated with the s covariates for each of the

$C-1$ contrasts. Similarly, $\mathbf{\Lambda}$ contains the random-effect variance parameters for each of the $C-1$ contrasts. Specifically

$$\mathbf{\Lambda}=[\mathrm{v}(\mathbf{T}_1)\quad \mathrm{v}(\mathbf{T}_2)\quad \ldots\quad \mathrm{v}(\mathbf{T}_{C-1})]$$

where $\mathrm{v}(\mathbf{T}_c)$ is the $r^*\times 1$ vector ($r^*=r[r+1]/2$) of elements below and on the diagonal of the Cholesky (lower-triangular) factor $\mathbf{T}_c$, and J_{r^*} is the transformation matrix of Magnus [30] that eliminates the elements above the main diagonal. This latter matrix is necessary to ensure that the appropriate terms from the $1\times r^2$ vector resulting from the Kronecker product $(\mathbf{x}'_{ij}\otimes\boldsymbol{\theta}'_j)$ are multiplied with the $r^*\times 1$ vector resulting from $\mathbf{\Lambda}\mathbf{d}_c$.

Several special cases of the model are worth noting. If the random effects are independent, as described earlier, then the model simplifies to

$$z_{ijc}=\mathbf{w}'_{ij}\mathbf{\Gamma}\mathbf{d}_c+\mathbf{x}'_{ij}\mathbf{\Lambda}\mathbf{d}_c\theta_i \tag{5}$$

where $\mathbf{\Lambda}$ is the $r\times(C-1)$ matrix of r independent random-effects variance terms (that is, standard deviations) for each of the $C-1$ category contrasts. Similarly, for the case of a random-intercepts model, the model simplifies to

$$z_{ijc}=\mathbf{w}'_{ij}\mathbf{\Gamma}\mathbf{d}_c+\mathbf{\Lambda}\mathbf{d}_c\theta_i \tag{6}$$

with $\mathbf{\Lambda}$ as the $1\times(C-1)$ vector $\mathbf{\Lambda}=[\sigma_1\ \sigma_2\ \ldots\ \sigma_{C-1}]$. Finally, notice that if $\mathbf{D}$ equals

$$\mathbf{D}=\begin{bmatrix}0&1&0&\ldots&0\\0&0&1&\ldots&0\\.&.&.&\ldots&.\\0&0&0&\ldots&1\end{bmatrix}$$

the model simplifies to the earlier representation in (1) and (2).

The current formulation allows for a great deal of flexibility in the types of comparisons across the C response categories. For example, if the categories are ordered, an alternative to the cumulative logits of the commonly-used proportional odds model is to employ Helmert contrasts [31] within the nominal model. For this, with $C=4$, the following contrast matrix would be used:

$$\mathbf{D}=\begin{bmatrix}-\frac{3}{4}&\frac{1}{4}&\frac{1}{4}&\frac{1}{4}\\0&-\frac{2}{3}&\frac{1}{3}&\frac{1}{3}\\0&0&-\frac{1}{2}&\frac{1}{2}\end{bmatrix}$$

with the scale of each contrast set to equal unity in terms of the difference. Helmert contrasts of the logits are similar but not exactly the same as the comparisons within continuation-ratio logit models, as described within a mixed model formulation by Ten Have and Uttal [32]. The difference is that the Helmert contrasts above are applied to the category logits rather than the category probabilities as in continuation-ratio models.

2.2. Parameter estimation

Let $\mathbf{y}_i$ denote the vector of nominal responses from level-2 unit i (for the n_i level-1 units nested within). Then the probability of any $\mathbf{y}_i$, conditional on the random effects $\boldsymbol{\theta}$ (and given $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$), is equal to the product of the probabilities of the level-1 responses

$$\ell(\boldsymbol{y}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n_j} \prod_{c=1}^{C} (p_{ijc})^{y_{ijc}} \tag{7}$$

where $y_{ijc} = 1$ if $y_{ij} = c$, and 0 otherwise. Thus, associated with the response from a particular level-1 unit, $y_{ijc} = 1$ for only one of the C categories and zero for all others. The marginal density of the response vector $\mathbf{y}_i$ in the population is expressed as the following integral of the likelihood, $\ell(.)$, weighted by the prior density $g(.)$:

$$h(\mathbf{y}_i) = \int_{\boldsymbol{\theta}} \ell(\boldsymbol{y}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$$

where $g(\boldsymbol{\theta})$ represents the population distribution of the random effects.

For parameter estimation, the marginal log-likelihood from the N level-2 units can be written as $\log L = \sum_i^N \log h(\mathbf{y}_i)$. Then, using $\boldsymbol{\Delta}$ to represent either parameter matrix

$$\frac{\partial \log L}{\partial \boldsymbol{\Delta}'} = \sum_{i=1}^{N} h^{-1}(\mathbf{y}_i) \int_{\boldsymbol{\theta}} \left[\sum_{j=1}^{n_i} \mathbf{D}(\mathbf{y}_{ij} - \mathbf{p}_{ij}) \otimes \partial \boldsymbol{\Delta} \right] \ell(\boldsymbol{y}_i \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{8}$$

where

$$\partial \boldsymbol{\Gamma} = \mathbf{w}'_{ij}, \quad \partial \boldsymbol{\Lambda} = [\mathrm{J}_{r*}(\boldsymbol{\theta} \otimes \mathbf{x}_{ij})]' \tag{9}$$

Here, $\mathbf{y}_{ij}$ is the $C \times 1$ indicator vector, and $\mathbf{p}_{ij}$ is the $C \times 1$ vector obtained by applying (3) for each category. Notice that if the random effects are independent then $\partial \boldsymbol{\Lambda}$ simply equals $\mathbf{x}'_{ij}\theta$, while for the even simpler random-intercepts model $\partial \boldsymbol{\Lambda} = \theta$.

As described elsewhere (for example, see Hedeker and Gibbons [23]), Fisher's method of scoring can be used to provide the solution to these likelihood equations. At convergence, the ratio of the (maximum marginal likelihood) estimates to their standard errors can be used to construct asymptotic z-statistics (for example, Wald statistics). Additionally, the converged value of the log-likelihood can be used to construct likelihood-ratio tests.

2.3. Numerical quadrature

Numerical integration can be used to perform the integration over the random-effects distribution that is indicated in (8). Specifically, if the assumed distribution is normal, Gauss–Hermite quadrature can be used to approximate the above integral to any practical degree of accuracy [33]. The integration is approximated by a summation on a specified number of quadrature points Q for each dimension of the integration. If the random effects are assumed to follow a distribution other than the normal distribution, other points may be chosen and density weights substituted for those specified by Gauss–Hermite quadrature. For example, if a rectangular or uniform distribution is assumed, then Q points may be set at equal intervals over an appropriate range (for each dimension) and the quadrature weights are then set equal to $1/Q$. Other distributions are possible, including the possibility of empirically estimating the random-effect

distribution [34]. In the examples below, we will compare results assuming a normal distribution to those obtained under a uniform distribution to provide some information about the sensitivity of the results to the assumed normal distribution.

Model estimation using quadrature has been implemented for use in the MIXNO program [35].* At each iteration and for each level-2 unit, the solution goes over the Q^r quadrature points, with summation replacing the integration over the random-effect distribution. The conditional probabilities $\ell(\boldsymbol{y}_i \mid \boldsymbol{\theta})$ are obtained substituting the random-effect vector $\boldsymbol{\theta}$ by the current r-dimensional vector of quadrature points $\mathbf{B}_q$. The marginal density for each level-2 unit is then approximated as

$$h(\mathbf{y}_i) \approx \sum_{q}^{Q^r} \ell(\mathbf{y}_i \mid \mathbf{B}_q) A(\mathbf{B}_q)$$

At each iteration, computation of the first derivatives and information matrix then proceeds summing over level-2 units and quadrature points. In the summation over the Q^r quadrature points, the $\boldsymbol{\theta}$ random-effect vector is substituted by the current vector of quadrature points $\mathbf{B}_q$, and the evaluation of the density $g(\boldsymbol{\theta})$ is substituted by the current quadrature weight $A(\mathbf{B}_q)$. Following the summation over level-2 units and quadrature points, parameters are corrected according to the Fisher scoring solution, and the entire procedure is repeated until convergence.

2.4. Estimation of random effects and marginal probabilities

In some cases, it may be of interest to estimate values of the random effects $\boldsymbol{\theta}_i$ within the sample. A reasonable choice for this is the expected '*a posteriori*' (EAP) or empirical Bayes estimator $\bar{\boldsymbol{\theta}}_i$ [34]. For the univariate case, this estimator $\bar{\theta}_i$ is given by

$$\bar{\theta}_i = E(\theta_i \mid \mathbf{y}_i) = \frac{1}{h(\mathbf{y}_i)} \int_{\theta} \theta_i \ell(\boldsymbol{y}_i \mid \theta) g(\theta)\, \mathrm{d}\theta \tag{10}$$

The variance of this estimator is obtained similarly as

$$V(\bar{\theta}_i \mid \mathbf{y}_i) = \frac{1}{h(\mathbf{y}_i)} \int_{\theta} (\theta_i - \bar{\theta}_i)^2 \ell(\boldsymbol{y}_i \mid \theta) g(\theta)\, \mathrm{d}\theta \tag{11}$$

Upon convergence, these quantities can be obtained using one additional round of quadrature. They may then be used, for example, to evaluate the response probabilities for particular level-2 units. Also, Ten Have [24] suggests how these empirical Bayes estimates might be used in performing residual diagnostics.

An additional step is required to obtain estimated marginal probabilities. First, so-called 'subject-specific' probabilities [36, 37] are estimated for specific values of covariates and random effects $\boldsymbol{\theta}_i$ by applying (3) with $\hat{z}_{ijc} = \mathbf{w}'_{ij}\hat{\boldsymbol{\Gamma}}\mathbf{d}_c + (\mathbf{x}'_{ij} \otimes \boldsymbol{\theta}'_i)\mathrm{J}'_{r*}\hat{\boldsymbol{\Lambda}}\mathbf{d}_c$. Denoting these subject-specific probabilities as $\hat{p}_{\mathrm{ss}}$, marginal probabilities $\hat{p}_{\mathrm{m}}$ are then obtained by integrating over the random-effect distribution, namely $\hat{p}_{\mathrm{m}} = \int_{\boldsymbol{\theta}} \hat{p}_{\mathrm{ss}} g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$. Again, numerical quadrature can be used for this. Alternatively, for a random-intercepts model, the approximation described in Diggle *et al.* (reference [38], p. 142) does not require use of quadrature. For this, the

*This program and its manual can be obtained from http://www.uic.edu/~hedeker/mix.html.

estimated regression coefficients $\hat{\boldsymbol{\Gamma}}$ are divided by $\sqrt{\{(16\sqrt{3}/15\pi)^2\hat{\sigma}_c^2 + 1\}}$. These 'marginalized' coefficients $\hat{\boldsymbol{\Gamma}}^*$ are then directly used with $\hat{z}_{ijc} = \mathbf{w}'_{ij}\hat{\boldsymbol{\Gamma}}^*\mathbf{d}_c$ to produce estimated marginal probabilities. Results using either of these methods generally agree closely. In the example below, we illustrate model fit using the quadrature method.

2.5. Intraclass correlation

For a random-intercepts model it is often of interest to express the level-2 variance in terms of an intraclass correlation. One way to obtain this expression utilizes the underlying latent response tendencies, denoted as Y_{ijc}. Also, for simplicity, this will be done for the reference-cell formulation, though it applies to the more general contrast situation as well. The random-intercepts regression model for the latent variable Y_{ijc}, including level-1 residuals ε_{ijc}, is written as

$$Y_{ijc} = \mathbf{w}'_{ij}\boldsymbol{\alpha}_c + \sigma_c\theta_i + \varepsilon_{ijc} \quad c = 1, 2, \ldots, C \tag{12}$$

As mentioned earlier, for a particular ijth unit, the category c associated with the observed nominal response y_{ij} is the one for which Y_{ijc} is maximal. Since, in the present formulation, $c = 1$ is the reference category, $\boldsymbol{\alpha}_1 = \sigma_1 = 0$, and so the model can be rewritten as

$$Y_{ijc} = \mathbf{w}'_{ij}\boldsymbol{\alpha}_c + \sigma_c\theta_i + (\varepsilon_{ijc} - \varepsilon_{ij1}) \quad c = 2, \ldots, C \tag{13}$$

for the latent response tendency of category c relative to the reference category. It can be shown that the level-1 residuals ε_{ijc} for each category are distributed according to a type I extreme-value distribution (see Maddala, reference [39], p. 60). It can further be shown that the standard logistic distribution is obtained as the difference of two independent type I extreme-value variates (see McCullagh and Nelder, reference [40], pp. 20 and 142). As a result, the level-1 variance is given by $\pi^2/3$, which is the variance for a standard logistic distribution. The estimated intraclass correlations are thus calculated as $r_c = \hat{\sigma}_c^2/(\hat{\sigma}_c^2 + \pi^2/3)$, where $\hat{\sigma}_c^2$ is the estimated level-2 variance assuming normally distributed random intercepts. Notice that $C - 1$ intraclass correlations are estimated. As such, the cluster influence on the level-1 responses is allowed to vary across the nominal response categories.

3. HEALTH SERVICES RESEARCH EXAMPLE

The McKinney Homeless Research Project (MHRP) study [41, 42] in San Diego, CA, was designed to evaluate the effectiveness of using Section 8 certificates versus no certificates as a means of providing independent housing to the severely mentally ill homeless. Section 8 housing certificates were provided from the Department of Housing and Urban Development (HUD) to local housing authorities in San Diego. These housing certificates, which require clients to pay 30 per cent of their income towards rent, are designed to make it possible for low income individuals to choose and obtain independent housing in the community. A total of 361 clients took part in this longitudinal study employing a randomized factorial design. Clients were randomly assigned to one of two types of supportive case management (comprehensive versus traditional) and to one of two levels of access to independent housing (Section 8 certificates: yes or no). Eligibility for the project was restricted to individuals diagnosed with a severe and persistent mental illness who were either homeless or at high

Table I. Housing status across time: response proportions and sample sizes (n).

Group	Status	Time-point			
		Baseline	6 months	12 months	24 months
Control	street	0.555	0.186	0.089	0.124
	community	0.339	0.578	0.582	0.455
	independent	0.106	0.236	0.329	0.421
	n	180	161	146	145
Section 8	street	0.442	0.093	0.121	0.120
	community	0.414	0.280	0.146	0.228
	independent	0.144	0.627	0.732	0.652
	n	181	161	157	158

risk of becoming homeless at the start of the study. Individuals' housing status was classified at baseline and at 6, 12 and 24 months follow-ups.

Here, we focus on examining the effect of access to Section 8 certificates on housing outcomes across time. Specifically, at each timepoint subjects' housing status was classified as either streets/shelters, community housing, or independent housing. The observed sample sizes and response proportions by group are presented in Table I. These observed proportions indicate a general decrease in street living and an increase in independent living across time for both groups. The increase in independent housing, however, appears to occur sooner for the Section 8 group relative to the control group. Regarding community living, across time there is an increase for the control group and a decrease for the Section 8 group. There is some attrition across time; attrition rates of 19.4 per cent and 12.7 per cent are observed at the final time-point for the control and Section 8 groups, respectively. Since estimation of model parameters is based on a full-likelihood approach, the missing data are assumed to be 'ignorable' conditional on both the model covariates and the observed nominal responses [43]. In longitudinal studies, ignorable non-response falls under the 'missing at random' (MAR) assumption of Rubin [44], in which the missingness depends only on observed data. In what follows, since the focus is on describing the mixed-effects multinomial regression model, we will make the MAR assumption.

Several mixed-effects multinomial logistic regression models were fit to these data. The first two were random-intercept models assuming the random effects were normally and uniformly distributed, respectively. Results from these analyses are given in Tables II and III. Helmert contrasts were used for category comparisons: the first Helmert contrast compares non-street (that is, community and independent housing) to street housing, while the second Helmert contrast compares the two types of non-street housing (that is, independent versus community housing). These contrasts are well matched to primary study questions: (i) Do Section 8 certificates help subjects get off the streets?; (ii) Do Section 8 certificates help subjects get into independent rather than community housing for subjects off the street? Tables II and III list results for these two Helmert contrasts, respectively. For these analyses, the repeated housing status classifications were modelled in terms of time effects (6, 12 and 24 months

Table II. Housing status across time: 1289 observations within 361 subjects, mixed-effects multinomial regression estimates and standard errors (SE). Helmert contrast 1: community and independent versus street housing.

Term	Normal RE distribution		Uniform RE distribution	
	Estimate	SE	Estimate	SE
Intercept	**−1.564**	0.244	**−1.600**	0.230
$t1$ (6 month versus base)	**2.312**	0.322	**2.195**	0.319
$t2$ (12 month versus base)	**3.454**	0.484	**3.301**	0.475
$t3$ (24 month versus base)	**3.179**	0.387	**3.058**	0.382
Section 8 (yes = 1, no = 0)	*0.651*	0.334	*0.573*	0.311
Section 8 by $t1$	*0.934*	0.495	*0.898*	0.487
Section 8 by $t2$	−0.684	0.601	−0.647	0.586
Section 8 by $t3$	−0.324	0.517	−0.326	0.506
Subject SD	1.602	0.148	0.322	0.032
$-2 \log L$	2218.73		2224.74	

For fixed effects: bold indicates $p < 0.05$, italic indicates $0.05 < p < 0.10$.

Table III. Housing status across time: 1289 observations within 361 subjects, mixed-effects multinomial regression estimates and standard errors (SE). Helmert contrast 2: independent versus community housing.

Term	Normal RE distribution		Uniform RE distribution	
	Estimate	SE	Estimate	SE
Intercept	**−2.224**	0.326	**−2.255**	0.320
$t1$ (6 month versus base)	**0.741**	0.375	*0.690*	0.374
$t2$ (12 month versus base)	**1.268**	0.352	**1.230**	0.350
$t3$ (24 month versus base)	**1.839**	0.358	**1.830**	0.354
Section 8 (yes = 1 no = 0)	0.260	0.425	0.204	0.411
Section 8 by $t1$	**2.138**	0.505	**2.236**	0.501
Section 8 by $t2$	**2.465**	0.512	**2.584**	0.512
Section 8 by $t3$	**1.256**	0.509	**1.321**	0.504
Subject SD	1.463	0.166	0.336	0.034
$-2 \log L$	2218.73		2224.74	

For fixed effects: bold indicates $p < 0.05$, italic indicates $0.05 < p < 0.10$.

follow-ups compared to baseline), a group effect (Section 8 versus control), and group by time interaction terms.

In terms of statistical significance of the fixed effects, the two models yield similar conclusions. Thus, the random-effects distributional form does not seem to play an important role for these data, at least as characterized by these two distributional forms. Subjects in the control

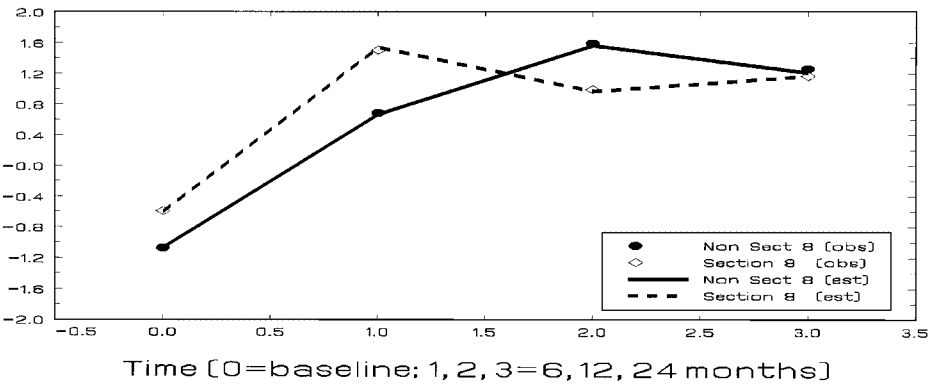

Figure 1. Independent+community versus street logits.

group increase both community and independent housing relative to street housing (Table II), and independent relative to street housing (Table III), at the three follow-ups as compared to baseline. In terms of group differences there are only marginally significant differences in the non-street versus street comparison (Table II). However, as Table III indicates, all Section 8 by time interactions are significant for the independent versus community housing comparison. Thus, the Section 8 group has greater increases than the control group at all follow-ups in independent housing, relative to community housing.

Considering the random subject effect in the normal random effects model, expressed as intraclass correlations, $r_1 = 0.44$ and $r_2 = 0.39$ for the two Helmert contrasts. Thus, the degree of subject influence is of moderate size and relatively similar for these two comparisons of the response categories. It should be noted that there are concerns in using the standard errors in constructing Wald test statistics for the variance terms (for example, the subject standard deviation), particularly when the population variance is near zero and the number of subjects is small [7]. As a result, statistical significance is not indicated for the random-effect variance parameters in the tables.

An analysis was also done to examine whether the random-effect variance terms varied significantly by treatment group. The deviance $(-2 \log L)$ for this model, assuming normally distributed random effects, equalled 2218.43, which was nearly identical to the value of 2218.73 (from Tables II and III) for the model assuming homogeneous variances across groups. The control group and Section 8 group estimates of the subject standard deviations were, respectively, 1.696 (SE = 0.212) and 1.499 (SE = 0.216) for non-street versus street comparison, and 1.471 (SE = 0.232) and 1.457 (SE = 0.244) for the independent versus community housing comparison. Thus, the homogeneity of variance assumption across treatment groups is reasonable.

Model fit to the observed data is depicted in Figures 1 and 2. The marginal observed logits are plotted with the 'marginalized' estimated logits of the mixed-effects model assuming normally distributed random effects. In terms of the non-street versus street comparison, Figure 1 shows the general increase across time for both groups. As the statistical tests indicated, the groups do not differ dramatically in terms of this logit over time. Figure 2, which illustrates the logits of independent versus community housing, clearly depicts the beneficial effect of Section 8 certificates at all follow-up time-points. Considering these plots along with the

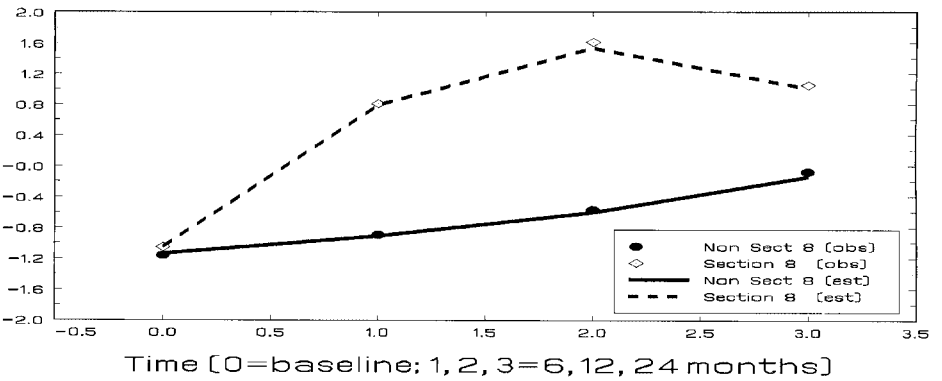

Figure 2. Independent versus community logits.

results of the mixed-effects analysis, it is seen that both groups are relatively successful in getting off the streets, but do so in somewhat different ways. The Section 8 group is much more likely to achieve independent housing than the control group, which is more likely to move towards community housing.

4. DISCUSSION

A mixed-effects multinomial logistic regression model is proposed for the analysis of multilevel nominal response data. Maximum marginal likelihood methods are used for parameter estimation. For this solution, quadrature is utilized to numerically integrate over the distribution of random effects. For multilevel data with two levels this model allows for multiple random effects at the second level. Fixed covariates can be included into the model at either level of the data.

Comparisons across the C nominal categories are specified either by selecting one category as the reference cell or by specifying a set of other $C-1$ category contrasts. The covariate effects are then estimated for each of the $C-1$ comparisons. The random-effect variance terms are also allowed to vary across the $C-1$ nominal category comparisons. The model makes an assumption that has been referred to as 'independence of irrelevant alternatives' in the econometric literature [39]. This is because the effect of an explanatory variable comparing two categories is the same regardless of the total number of categories considered. This assumption is generally reasonable when the categories are distinct and dissimilar, and unreasonable as the nominal categories are seen as substitutes for one another [45, 46]. Furthermore, McFadden [47] notes that the multinomial logistic regression model is relatively robust in many cases in which this assumption is implausible. In the present example, the outcome categories are fairly distinct and so the assumption would seem to be reasonable for these data. The possibility of relaxing this assumption for a more general mixed-effects multinomial regression model has recently been discussed by Hartzel *et al.* [29].

As noted, the solution via quadrature can involve summation over a large number of points when the number of random effects is increased. An issue, then, is the number of necessary quadrature points to use for accurate estimation of the model parameters. Based on models

for dichotomous and ordinal outcomes, respectively, Longford [8] and Jansen [20] note that estimation is affected very little when the number of points is five or greater for the unidimensional solution. Also, as suggested by Bock *et al.* [48] in the context of a dichotomous factor analysis model, the number of points in each dimension can be reduced as the dimensionality is increased. These authors noted that as few as three points per dimension were necessary for a five-dimensional solution. Alternatively, the EGRET program [49] uses 20 quadrature points by default for its unidimensional mixed-effects (binary) logistic regression model. In the present example, we compared results based on 10 versus 20 quadrature points and found little difference.

The use of Gibbs sampling and related methods [50] provides an alternative way of handling the integration over the random-effect distribution. While the quadrature solution is relatively fast and computationally tractable for models with few random effects, Gibbs sampling is more advantageous for models with many random effects. For example, if there is only one random effect, the quadrature solution requires only one additional summation over Q points relative to the fixed effects solution. For models with multiple random effects, however, the quadrature is performed over Q^r points (where r equals the number of random effects), and so becomes computationally burdensome for $r > 5$ or so. Attempting to overcome this problem, methods of adaptive quadrature have been developed [51–53] that use fewer number of points per dimension (for example, three or so) that are adapted to the location and dispersion of the distribution to be integrated. For dichotomous factor analysis models with five and eight factors (that is, random effects), Bock and Shilling [53] found similar results for adaptive quadrature as compared to a Gibbs sampling approach.

The example presented illustrates the usefulness of the mixed-effects approach for longitudinal categorical data. Mixed-effects models are also useful in analysis of clustered data, where individuals are observed nested within schools, hospitals or firms, for example. A further extension of the model is underway to allow for three-level data in order to accommodate clustered data where the clustered subjects are also repeatedly measured across time.

ACKNOWLEDGEMENTS

Thanks are due to Drs Richard Hough and Michael Hurlburt for use of the longitudinal data and for helpful comments in their analysis. This work was supported by National Institutes of Mental Health grant MH56146.

REFERENCES

1. Cox DR. *Analysis of Binary Data*. Chapman and Hall: London, 1970.
2. Bock RD. Estimating multinomial response relations. In *Contributions to Statistics and Probability*, Bose RC (ed.). University of North Carolina Press: Chapel Hill, NC, 1970.
3. Nerlove M, Press SJ. Univariate and multivariate log-linear and logistic models. Technical Report R-1306-EDA/NIH, Rand Corporation, Santa Monica, CA, 1973.
4. Plackett RL. *The Analysis of Categorical Data*. Charles Griffin: London, 1974.
5. Agresti A. *Categorical Data Analysis*. Wiley: New York, 1990.
6. Goldstein H. *Multilevel Statistical Models*. 2nd edn. Halstead Press: New York, 1995.
7. Bryk AS, Raudenbush SW. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications Inc.: Newbury Park, CA, 1992.
8. Longford NT. *Random Coefficient Models*. Oxford University Press: New York, 1993.
9. Stiratelli R, Laird NM, Ware JH. Random-effects models for serial observations with binary response. *Biometrics* 1984; **40**:961–971.

10. Anderson DA, Aitkin M. Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society* 1985; **47**:203–210.
11. Wong GY, Mason WM. The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association* 1985; **80**:513–524.
12. Gibbons RD, Bock RD. Trend in correlated proportions. *Psychometrika* 1987; **52**:113–124.
13. Conaway MR. Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association* 1989; **84**:53–61.
14. Goldstein H. Nonlinear multilevel models, with an application to discrete response data. *Biometrika* 1991; **78**:45–51.
15. Fitzmaurice GM, Laird NM, Rotnitzky AG. Regression models for discrete longitudinal responses. *Statistical Science* 1993; **8**:284–309.
16. Rodriguez G, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society*, *Series A* 1995; **158**:73–89.
17. Pendergast JF, Gange SJ, Newton MA, Lindstrom MJ, Palta M, Fisher MR. A survey of methods for analyzing clustered binary response data. *International Statistical Review* 1996; **64**:89–118.
18. Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, *Series B* 1996; **159**:505–513.
19. Harville DA, Mee RW. A mixed-model procedure for analyzing ordered categorical data. *Biometrics* 1984; **40**:393–408.
20. Jansen J. On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics* 1990; **39**:75–84.
21. Ezzet F, Whitehead J. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine* 1991; **10**:901–907.
22. Agresti A, Lang JB. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* 1993; **80**:527–534.
23. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
24. Ten Have TR. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* 1996; **52**:473–491.
25. Hedeker D, Mermelstein RJ. A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research* 1998; **33**:427–455.
26. Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 1972; **37**:29–51.
27. Daniels MJ, Gatsonis C. Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine* 1997; **16**:2311–2325.
28. Revelt D, Train K. Mixed logit with repeated choices: Household's choices of appliance efficiency level. *Review of Economics and Statistics* 1998; **80**:647–657.
29. Hartzel J, Agresti A, Caffo B. Multinomial logit random effects models. *Statistical Modelling* 2001; **1**:81–102.
30. Magnus JR. *Linear Structures*. Charles Griffin: London, 1988.
31. Bock RD. *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill: New York, 1975.
32. Ten Have TR, Uttal DH. Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Applied Statistics* 1994; **43**:371–384.
33. Stroud AH, Sechrest D. *Gaussian Quadrature Formulas*. Prentice Hall: Englewood Cliffs, NJ, 1966.
34. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: an application of the em algorithm. *Psychometrika* 1981; **46**:443–459.
35. Hedeker D. Mixno: a computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software* 1999; **4**(5):1–92.
36. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; **59**:25–35.
37. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
38. Diggle P, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: New York, 1994.
39. Maddala GS. *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press: Cambridge, U.K., 1983.
40. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd edn. Chapman and Hall, New York, 1989.
41. Hough RL, Harmon S, Tarke H, Yamashiro S, Quinlivan R, Landau-Cox P, Hurlburt MS. Supported independent housing: Implementation issues and solutions in the san diego mckinney homeless demonstration research project. In *Mentally Ill and Homeless: Special Programs for Special Needs*, Breakey WR, Thompson JW (eds). Harwood Academic Publishers: New York, 1997; 95–117.
42. Hurlburt MS, Wood PA, Hough RL. Providing independent housing for the homeless mentally ill: a novel approach to evaluating long-term longitudinal housing patterns. *Journal of Community Psychology* 1996; **24**:291–310.

43. Laird NM. Missing data in longitudinal studies. *Statistics in Medicine* 1988; **7**:305–315.
44. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
45. McFadden D. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, Zarembka P (ed.). Academic Press: New York, 1973.
46. Amemiya T. Qualitative response models: a survey. *Journal of Econometric Literature* 1981; **19**:483–536.
47. McFadden D. Qualitative response models. In *Advances in Econometrics*, Hildenbrand W (ed.). Cambridge University Press: Cambridge, U.K., 1980; 1–37.
48. Bock RD, Gibbons RD, Muraki E. Full-information item factor analysis. *Applied Psychological Measurement* 1988; **12**:261–280.
49. Corcoran C, Coull B, Patel A. *EGRET for Windows User Manual*. CYTEL Software Corporation: Cambridge, MA, 1999.
50. Tanner MA. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 3rd edn. Springer Verlag: New York, 1996.
51. Liu Q, Pierce DA. A note on gauss-hermite quadrature. *Biometrika* 1994; **81**:624–629.
52. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in nonlinear mixed-effects models. *Journal of Computational and Graphical Statistics* 1995; **4**:12–35.
53. Bock RD, Shilling S. High-dimensional full-information item factor analysis. In *Latent Variable Modeling and Applications to Causality*, Berkane M (ed.). Springer: New York, 1997; 163–176.

