ELSEVIER Computational Statistics & Data Analysis 35 (2001) 429–449

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials

Jonathan Hartzel[a], I-Ming Liu[b], Alan Agresti[a,*]

[a] *Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, USA*
[b] *Department of Statistics, National Chung Hsing University, Taipei, Taiwan, ROC*

Received 1 August 1999; received in revised form 1 June 2000

Abstract

Standard models for a set of contingency tables with ordered response categories assume a common effect within or between tables, described by a certain type of odds ratio. In practice, heterogeneity usually occurs among such odds ratios, even if its extent is minor in magnitude. This article presents models that summarize the effect while simultaneously describing the degree of heterogeneity. For cases in which the levels of the stratification factor are a sample, such as many multi-center clinical trials, we recommend the use of random effects models. These treat the true stratum-specific ordinal log odds ratios as a sample with some unknown mean and standard deviation. For the random effects distribution, we consider both normality and a nonparametric approach. In using these more realistic models permitting heterogeneity, it can be more difficult to establish significance of effects because of the extra variability inherent in the model. The primary focus is three-way contingency tables with an ordinal response and a stratification factor, but we also briefly discuss models for describing heterogeneity within contingency tables.

Keywords: Adjacent-categories logit; Association models; Cumulative logit; Gauss–Hermite quadrature; Nonparametric mixture model; Odds ratio; Proportional odds model; Random effect; Treatment-by-center interaction

* Corresponding author.
E-mail address: aa@stat.ufl.edu (A. Agresti).

0167-9473/01/$ - see front matter
PII: S 0167-9473(00)00020-7

Table 1
Clinical trial relating treatment to response for eight centers

		Response		
Center	Treatment	Much better	Better	Unchanged/worse
1	Drug	13	7	6
	Placebo	1	1	10
2	Drug	2	5	10
	Placebo	2	2	1
3	Drug	11	23	7
	Placebo	2	8	2
4	Drug	7	11	8
	Placebo	0	3	2
5	Drug	15	3	5
	Placebo	1	1	5
6	Drug	13	5	5
	Placebo	4	0	1
7	Drug	7	4	13
	Placebo	1	1	11
8	Drug	15	9	2
	Placebo	3	2	2

1. Introduction

For comparing groups on a categorical response with stratified data, a common starting point for modeling assumes a lack of interaction in the sense that certain odds ratios relating group and response are the same for each stratum. Tests exist for checking homogeneity of odds ratios, such as goodness-of-fit tests for the corresponding models. In practice, however, the true relationship usually has some heterogeneity, even if it is minor in magnitude and perhaps not even significant in the sample according to a statistical test. In a clinical trial conducted to compare treatments among several study centers, for instance, the true treatment effects may have the same direction for each center but might vary somewhat due to unmeasured factors such as differential mean socioeconomic status or age among subjects at different centers. In such cases, it may be more relevant to estimate the degree of the heterogeneity (treatment-by-center interaction) than simply to test whether it exists.

Table 1, part of a data set analyzed by one of us (I. Liu) during a summer internship at Merck pharmaceutical company, is an example of this type for an ordinal response variable. This table shows some preliminary results for eight of the centers

from a double-blind, parallel-group clinical study. The purpose of the study was to compare an active drug with placebo in the treatment of patients suffering from asthma. Patients were randomly assigned to the treatments. (The original study was a double-blind, parallel-group study to compare the effect of three doses of the active drug and placebo in chronic asthmatic patients. For simplicity of exposition here, we compare only the effect of the active drug, which is the combination of the three different dose groups, and placebo. Therefore, Table 1 shows a drug group with almost three times the number of patients as the placebo group.) At the end of the study, investigators described their perception of the patient's change in condition, using the ordinal scale (much better, better, unchanged or worse). The focus of the study was comparison of the treatments and investigation of potential treatment-by-center interaction. We will present analyses that compare the treatments while simultaneously modeling the association variability among centers.

We study potential heterogeneity using models that contain parameters describing the variability in odds ratios among strata. We consider two ways of doing this for ordinal response variables. One approach, a standard one, uses fixed effects modeling. Our main emphasis is on a second approach that uses random effects terms to describe the variability in conditional associations. This approach is natural when the strata are a sample, such as a sample of clinics or geographical areas. The analysis results in a simple summary consisting of a mean and standard deviation estimate for the variation across strata of an ordinal measure of association. We present these approaches for cumulative logit models, for which the ordinal measure is a cumulative log odds ratio, and for adjacent-categories logit models, for which the ordinal measure is a local log odds ratio. For the random effects models we use both a parametric version, assuming a normal distribution for the random effects, and a distribution-free version.

Section 2 reviews the standard fixed effects models and introduces the random effects approach with models that imply that ordinal log odds ratios are a sample from some distribution. Section 3 discusses model fitting and likelihood-ratio tests of no effect and of homogeneity, based on models with random effects having normal or unspecified distribution. Section 4 illustrates the various methods applied to Table 1. Section 5 extends the modeling of heterogeneity by also describing variability *within* each stratum with respect to rows and columns, using generalized loglinear models.

2. Describing heterogeneity with fixed and random effects in ordinal logit models

Let Y denote an ordinal response variable with c levels, X an explanatory variable with r levels, and Z a stratification factor with L levels. Let n_{ijk} denote the count at level i of X, j of Y, and k of Z. Let $\pi_{j|ik} = P(Y = j|X = i, Z = k)$. For $i = 1, \ldots, r$ and $k = 1, \ldots, L$, we assume that $(n_{i1k}, \ldots, n_{ick})$ has a multinomial distribution with probabilities $(\pi_{1|ik}, \ldots, \pi_{c|ik})$ and that samples from different levels of X and/or Z are independent. Let β denote a parameter describing the effect of X on Y. We summarize the effect by evaluating how β varies among levels of Z.

2.1. Cumulative logit models allowing heterogeneity

We illustrate first using the cumulative logit form of model. The jth cumulative probability on Y is $\pi^*_{j|ik} = \pi_{1|ik} + \cdots + \pi_{j|ik}$, for $j = 1, \ldots, c$. Suppose that X is itself ordinal, with fixed monotone scores $\{x_i\}$ for its levels, or binary with indicator $(x_1 = 0, x_2 = 1)$. A simple version of the cumulative logit form of model is

$$\text{logit}(\pi^*_{j|ik}) = \log\left(\frac{\pi_{1|ik} + \cdots + \pi_{j|ik}}{\pi_{j+1|ik} + \cdots + \pi_{c|ik}}\right) = \alpha_j - \gamma_k - \beta x_i, \quad j = 1, \ldots, c-1, \quad (1)$$

for all i and k. Identifiability requires a constraint such as $\alpha_1 = 0$. When $\{x_i\}$ are equally spaced with $\{x_{i+1} - x_i = 1\}$, in each stratum β denotes the log odds ratio for any of the $c-1$ collapsings to a 2×2 table of the $2 \times c$ table consisting of rows i and $i+1$. That is, at any level k of Z, the odds that Y falls above level j multiplies by $\exp(\beta)$ for each unit increase in X. We refer to $\exp(\beta)$ as the common *cumulative odds ratio* for the $X - Y$ conditional association. It is standard to estimate β using maximum likelihood (ML), although a Mantel–Haenszel style estimate is useful when the data are sparse (Liu and Agresti, 1996).

A limitation of model (1) is that it assumes homogeneous effects across levels of Z. To permit heterogeneity, one can generalize this model to

$$\text{logit}(\pi^*_{j|ik}) = \alpha_j - \gamma_k - \beta_k x_i, \quad j = 1, \ldots, c-1. \quad (2)$$

For each stratum, this model and model (1) assume the *proportional odds* structure (McCullagh, 1980) of a common effect of X for all categories j at which one can form a cumulative probability. In practice, it is often too restrictive to require the same 'cutpoints' $\{\alpha_j\}$ in each stratum, but the proportional odds structure also holds for the more general model with $\alpha_j - \gamma_k$ in (2) replaced by α_{jk}. The maximum likelihood fit is then equivalent to fitting a proportional odds model separately in each stratum. More general models yet (see, e.g., Section 5) do not assume the proportional odds structure or else have a nonlinear form that also allows the dispersion to depend on the predictor (McCullagh, 1980). One could use alternative link functions in these various types of models, but we illustrate with the logit link in this article.

We now consider random effects versions of models (1) and (2) that treat the strata as a sample. A random intercept version of the homogeneity model (1) is

$$\text{logit}(\pi^*_{j|ik}) = \alpha_j - c_k - \beta x_i, \quad j = 1, \ldots, c-1. \quad (3)$$

where $\{c_k\}$ are independent observations from a $N(\gamma, \sigma_c)$ distribution (Alternatively, one could remove constraints on $\{\alpha_j\}$ and then set $\gamma = 0$). This model is an extension of the random-intercept logistic-normal model for binary data (Pierce and Sands, 1975). In our experience, the ML estimate of the fixed effect β for this model and its standard error are very similar to those for model (1).

A more substantial and more useful extension permits heterogeneity in the conditional associations. A random effects version of the heterogeneous effects model (2) is

$$\text{logit}(\pi^*_{j|ik}) = \alpha_j - c_k - b_k x_i, \quad j = 1, \ldots, c-1, \quad (4)$$

where $\{b_k\}$ are independent observations from a $N(\beta, \sigma_b)$ distribution and $\{c_k\}$ are independent observations from a $N(\gamma, \sigma_c)$ distribution. As in the fixed effects case, one can also consider more complex cutpoint structure that replaces $\alpha_j - c_k$ by a vector of random effects $\{a_{jk},\ j = 1, \ldots, c-1\}$. Although such a model may fit better, it often presents computational problems for sparse data such as Table 1. In addition, this usually has little impact on the question of main focus – namely, on estimating the mean and variability in the association effects $\{b_k\}$.

The normality assumption for the random effects in these models can rarely be checked closely, especially when the number of strata L is not especially large, such as in many multi-center clinical trials. An attractive alternative that avoids this parametric assumption assumes a mixing distribution of unspecified form for the random effect (e.g., Aitkin, 1996, 1999). That is, one uses nonparametric maximum likelihood (NPML) to jointly estimate the regression parameters and the random effects distribution. Section 3 provides details of an EM algorithm for fitting this model. In either the parametric or nonparametric approach with the interaction model (4), the primary focus would usually be on estimating the expected cumulative log odds ratio β, and the variability in the log odds ratios across strata, such as described by their standard deviation σ_b.

2.2. Alternative ordinal models allowing heterogeneity

Similar approaches apply for modeling heterogeneity with other types of ordinal odds ratios. For instance, one could model heterogeneity in local odds ratios, which result both from certain logit and loglinear models. For expected cell counts $\{\mu_{ijk} = E(n_{ijk})\}$ with X and Y both ordinal, the loglinear model of heterogeneous linear-by-linear association (Agresti and Kezouh, 1983) is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \beta_k x_i y_j, \tag{5}$$

for fixed scores $\{x_i\}$ and $\{y_j\}$. With $\{x_i = i\}$ and $\{y_j = j\}$, all local log odds ratios in stratum k equal β_k (i.e., there is heterogeneous uniform association). This parameter also follows from the adjacent-categories logit model analog of (2),

$$\log\left(\frac{\pi_{j|ik}}{\pi_{j+1|ik}}\right) = \alpha_j - \gamma_k - \beta_k x_i, \quad j = 1, \ldots, c-1. \tag{6}$$

(With $\{y_j = j\}$, the loglinear model (5) is equivalent to the extended version of this model with $\alpha_j - \gamma_k$ replaced by α_{jk}.) A random effects version of model (6) is

$$\log\left(\frac{\pi_{j|ik}}{\pi_{j+1|ik}}\right) = \alpha_j - c_k - b_k x_i, \quad j = 1, \ldots, c-1, \tag{7}$$

with $\{b_k\}$ and $\{c_k\}$ specified as in (4), treated either as normal or in a distribution-free manner.

When Y has $c = 2$ categories, cumulative logit and adjacent-categories logit models simplify to the same model. In that case, a variety of models have been used to describe heterogeneity in odds ratios among several 2×2 tables. See, for instance, Skene and Wakefield (1990), Liu and Pierce (1993), and Agresti and Hartzel (2000).

When $r>2$ and X is nominal or treated as nominal by allowing arbitrary effects rather than assuming a linear trend, one replaces b_k in the heterogeneity models by a vector of correlated random effects $(b_{k1},\ldots,b_{k,r-1})$ that are coefficients of dummy variables.

2.3. Choice of scores and correlation structure for random effects

We have not yet fully stated the cumulative logit model (4) and the adjacent-categories logit model (7) containing random intercepts and slopes, since we have not indicated the correlation structure for (b_k,c_k). For simplicity, one might simply take these random effects to be uncorrelated. However, the results then depend on the location of the scores for the predictor variable x. For instance, suppose we assume model (4) but shift the scores to $\{x_i+d\}$ for some fixed $d\neq 0$. The new model then has form

$$\text{logit}(\pi^*_{j|ik})=\alpha_j-c_k-b_k(x_i+d)=\alpha_j-c^*_k-b_kx_i,$$

where $c^*_k=c_k+b_kd$. Then, (b_k,c^*_k) are correlated even if (b_k,c_k) are not.

Thus, two models in which the scores in one are a location shift of the scores in the other are not equivalent when one forces the random effects to be uncorrelated; they will provide different estimates of the parameters (β,σ_b) of interest, which is undesirable. The models are equivalent, producing the same estimates of (β,σ_b), if one allows the random effects to be correlated in the two cases. We recommend fitting such models allowing correlated random effects. For relatively small numbers of centers such as in Table 1, however, the estimates of variance components and correlations are typically very imprecise for either case.

3. Model fitting and inference

We now discuss model fitting and inference for the random effects models presented in the previous section. Because the response is multinomial rather than binomial, these models are special cases of multivariate generalized linear mixed models (MGLMMs) for ordinal responses (Tutz and Hennevogl, 1996). We present the estimation methods in terms of the MGLMM and note the applications to the previous models. We then present an EM algorithm for fitting the nonparametric random effects version of the model.

3.1. Multivariate generalized linear mixed models

In a general setting with clustered data, let $\boldsymbol{y}_{ik}$ be the response vector for the ith observation in the kth cluster and let $\boldsymbol{u}_k$ be a vector of random effects. In a MGLMM one assumes that (1) conditionally on $\boldsymbol{u}_k$ the observations are independent with conditional distribution $f(\boldsymbol{y}_{ik}|\boldsymbol{u}_k)$, a member of the multivariate exponential family, with conditional mean and linear predictor given by

$$\boldsymbol{\mu}_{ik}=E(\boldsymbol{y}_{ik}|\boldsymbol{u}_k)=\boldsymbol{h}(\boldsymbol{\eta}_{ik}),\quad \boldsymbol{\eta}_{ik}=\boldsymbol{Z}_{ik}\boldsymbol{\alpha}+\boldsymbol{W}_{ik}\boldsymbol{u}_k,$$

and (2) the $\boldsymbol{u}_k$ are independent with a $N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ distribution. To express the ordinal models as MGLMMs, treating strata as clusters, we first re-express the ordinal response Y for subject s with treatment i in stratum k as a response vector $\boldsymbol{y}_{sik} = (y_{si1k}, \ldots, y_{siqk})$:

$$y_{sijk} = \begin{cases} 1 & \text{if } Y = j, \quad j = 1, \ldots, q = c - 1, \\ 0 & \text{otherwise,} \end{cases}$$

with corresponding response probabilities $\boldsymbol{\pi}'_{ik} = (\pi_{1|ik}, \ldots, \pi_{q|ik})$. Then, denote the multinomial proportions for the n_{ik} subjects having treatment i in stratum k by $\boldsymbol{y}_{ik} = (\sum_s \boldsymbol{y}_{sik})/n_{ik}$. Since the distribution $f(\boldsymbol{y}_{ik}|\boldsymbol{u}_{ik})$ of $n_{ik}\boldsymbol{y}_{ik}$ is multinomial and thus in the multivariate exponential family, the models of the previous two sections are MGLMMs (Fahrmeir and Tutz, 1994, p. 69).

The relationship between the mean $\boldsymbol{\pi}_{ik}$ and the linear predictor $\boldsymbol{\eta}_{ik}$ for cumulative logit models is defined by the response function $\boldsymbol{h}(\boldsymbol{\eta}_{ik})$,

$$\pi_{1|ik} = h_1(\boldsymbol{\eta}_{ik}) = \frac{1}{1 + \exp(-\eta_{i1k})},$$

$$\pi_{j|ik} = h_j(\boldsymbol{\eta}_{ik}) = \frac{1}{1 + \exp(-\eta_{ijk})} - \frac{1}{1 + \exp(-\eta_{i,j-1,k})}, \quad j = 2, \ldots, c - 1.$$

The relationship for adjacent-categories logit models has the response function

$$\pi_{j|ik} = h_j(\boldsymbol{\eta}_{ik}) = \frac{\exp[-(j - c)\eta_{ijk}]}{1 + \sum_{j=1}^{q} \exp[-(j - c)\eta_{ijk}]}.$$

Then, for instance, for the heterogeneity models (4) and (7), the model matrix $\boldsymbol{Z}_{ik}$ for the fixed effects $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_{c-1})$ is the $(c - 1) \times (c - 1)$ identity matrix, the model matrix for the random effects $\boldsymbol{u}'_k = (c_k,\ b_k)$ is the $(c - 1) \times 2$ matrix

$$\boldsymbol{W}_{ik} = \begin{bmatrix} -1 & -x_i \\ \vdots & \\ -1 & -x_i \end{bmatrix},$$

and $\boldsymbol{u}_k \sim N[(\gamma, \beta)', \boldsymbol{\Sigma}]$ in the normal case.

3.2. ML model fitting with normal random effects

Let $g(\boldsymbol{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma})$ denote the multivariate normal density function with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$. The likelihood function for a MGLMM with stratified multinomial responses for r groups has the form

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{k=1}^{L} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|\boldsymbol{u}_k; \boldsymbol{\alpha}) \right] g(\boldsymbol{u}_k; \boldsymbol{\beta}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{u}_k. \tag{8}$$

Maximum likelihood estimates are obtained by maximizing (8), which involves evaluating intractable integrals. For the case of cumulative logit models, Tutz and Hennevogl (1996) utilized the EM algorithm along with either Monte Carlo or Gauss–Hermite quadrature approximations for the integrals. Hedeker and Gibbons (1994) considered both cumulative logit and cumulative probit models and directly

maximized the likelihood after approximating the integrals by Gauss–Hermite quadrature. Here, we also consider the adjacent-categories logit model, we utilize a different method for approximating the integrals in (8), and we consider a nonparametric as well as the normal parametric approach. We directly maximize (8) but approximate the integrals by adaptive Gauss–Hermite quadrature (Liu and Pierce, 1994; Pinheiro and Bates, 1995). Adaptive quadrature centers the Gauss–Hermite nodes with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode. From our experience, this dramatically reduces the number of quadrature points needed to approximate the integrals effectively.

To approximate the integrals for the kth stratum in (8), we first calculate the mode, $\hat{\boldsymbol{\mu}}_k$, of the integrand $\prod_{i=1}^{r}[f(\boldsymbol{y}_{ik}|\boldsymbol{u}_k;\boldsymbol{\alpha})]g(\boldsymbol{u}_k;\boldsymbol{\beta},\boldsymbol{\Sigma})$ and center the original Gauss–Hermite nodes about that point. We then scale the centered nodes according to the curvature of the integrand around the mode. An estimate of the curvature at the mode of the integrand can be obtained by inverting the negative of the second derivative matrix of the integrand evaluated at the estimated mode. We use numerical second derivatives for the estimation of the curvature, $\hat{\boldsymbol{Q}}_k$. Denote the dimension of the random effects vector $\boldsymbol{u}_k$ by m. The adaptive quadrature nodes for the kth stratum are then

$$\boldsymbol{z}_{k\boldsymbol{l}}^{*} = \hat{\boldsymbol{\mu}}_k + \sqrt{2}\,\hat{\boldsymbol{Q}}_k^{1/2}\boldsymbol{z}_{\boldsymbol{l}}$$

where $\boldsymbol{l}=(l_1,\ldots,l_m)$, $\boldsymbol{z}_{\boldsymbol{l}}=(z_{l_1},\ldots,z_{l_m})$, and $\{z_l\}$ are the original Gauss–Hermite nodes. The m-dimensional integral approximation is then

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|\boldsymbol{u}_k;\boldsymbol{\alpha})\right] g(\boldsymbol{u}_k;\boldsymbol{\beta},\boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{u}_k$$
$$= |\hat{\boldsymbol{Q}}|^{1/2}2^{m/2}\sum_{\boldsymbol{l}} w_{\boldsymbol{l}}\left[\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|\boldsymbol{z}_{k\boldsymbol{l}}^{*};\boldsymbol{\alpha})\right] g(\boldsymbol{z}_{k\boldsymbol{l}}^{*};\boldsymbol{\beta},\boldsymbol{\Sigma})\exp(\boldsymbol{z}_{\boldsymbol{l}}'\boldsymbol{z}_{\boldsymbol{l}}),$$

where $w_{\boldsymbol{l}}=\prod_{t=1}^{m} w_{l_t}$ and $\{w_{l_t}\}$ are the original Gauss–Hermite weights. The upper limits of the multiple summations over $\boldsymbol{l}$ should be increased until the desired accuracy in the approximation is reached. However, using s quadrature points in each dimension requires summing over s^m terms. Due to the exponential growth in computational effort with each dimension, only dimensions of up to 5 or 6 are currently computationally feasible.

Maximization of the likelihood (8) can be carried out by standard methods, such as Newton–Raphson or quasi-Newton methods. Care must be taken when maximizing with respect to the unique parameters of the covariance matrix, since nonnegative definite matrices can occur during the maximization routine. With a modest amount of work the observed information matrix can be calculated to provide standard errors at convergence (Hartzel, 1999).

We developed an OX program (Doornick, 1998) using adaptive Gauss–Hermite quadrature to fit cumulative logit and adjacent-categories logit models with random effects. The recently released Version 8 of SAS contains PROC NLMIXED for fitting generalized linear mixed models using adaptive Gauss–Hermite quadrature. Although it is not obvious that one can fit ordinal models using this procedure, it is possible

to do so by defining the appropriate likelihood function, as we illustrate in the next section. Hedeker and Gibbons (1994) developed a FORTRAN program, MIXOR, for proportional odds models with random effects, available at Hedeker's web site (www.uic.edu/ ∼hedeker).

3.3. ML model fitting with the nonparametric approach

An alternative to the normality assumption in the MGLMM is to assume that $g(\boldsymbol{u}_k)$ is a discrete distribution with unknown support size P, mass points $\boldsymbol{m}=(m_1,\ldots,m_P)$, and probabilities $\boldsymbol{p}=(p_1,\ldots,p_P)$. Joint estimation of $\boldsymbol{\alpha}$, $\boldsymbol{m}$, and $\boldsymbol{p}$ can be implemented with an EM algorithm. We describe this for the case of a single random effect, $\boldsymbol{u}_k=u_k$, generalizing work of Aitkin (1999) for binary response models. Denote the complete log-likelihood that contains both the observed data $\boldsymbol{y}$ and unobserved data $\boldsymbol{u}$ by

$$\log L(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\alpha})=\sum_{k=1}^{L}\log\left[\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|u_k;\boldsymbol{\alpha})\right]+\sum_{k=1}^{L}\log g(u_k). \tag{9}$$

In the E-step at iteration $(s+1)$ the expectation of the complete log-likelihood (9) is calculated with respect to the conditional distribution $f(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\alpha}^{(s)},\boldsymbol{m}^{(s)},\boldsymbol{p}^{(s)})$, where $\boldsymbol{\alpha}^{(s)},\boldsymbol{m}^{(s)}$, and $\boldsymbol{p}^{(s)}$ are the working parameter estimates from the previous iteration. Using independence, Bayes Rule, and expressing $g(u_k)$ in terms of the masses $\boldsymbol{p}$ and mass points $\boldsymbol{m}$, one obtains

$$\begin{aligned}&E[\log L(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\alpha},\boldsymbol{m},\boldsymbol{p}|\boldsymbol{\alpha}^{(s)},\boldsymbol{m}^{(s)},\boldsymbol{p}^{(s)})]\\&\quad=\sum_{k=1}^{L}\sum_{l=1}^{P}\left[q_{kl}^{(s)}\log\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|m_l;\boldsymbol{\alpha})+q_{kl}^{(s)}\log p_l\right],\end{aligned} \tag{10}$$

where

$$q_{kl}^{(s)}=\frac{p_l^{(s)}\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|m_l^{(s)};\boldsymbol{\alpha}^{(s)})}{\sum_{l=1}^{P} p_l^{(s)}\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|m_l^{(s)};\boldsymbol{\alpha}^{(s)})}.$$

Here, $q_{kl}^{(s)}$ represents the estimated posterior probability that the response vector $(\boldsymbol{y}_{1k},\ldots,\boldsymbol{y}_{rk})$ for stratum k comes from component l. The $\{q_{kl}^{(s)}\}$ are calculated from the parameter estimates at the sth iteration.

The M-step consists of maximizing (10) with respect to $\boldsymbol{\alpha}$, $\boldsymbol{m}$, and $\boldsymbol{p}$. The second term of (10) is not a function of $\boldsymbol{\alpha}$ or $\boldsymbol{m}$ and can be maximized separately from the first term. Maximizing $\sum_k\sum_l q_{kl}^{(s)}\log p_l$ subject to $\sum_{l=1}^{P} p_l=1$ yields simply

$$\hat{p}_l^{(s)}=\sum_{k=1}^{L} q_{kl}^{(s)}/L.$$

Since $q_{kl}^{(s)}$ is known, the first term of (10), $\sum_k\sum_l q_{kl}^{(s)}\log\prod_i f(\boldsymbol{y}_{ik}|m_l;\boldsymbol{\alpha})$, is simply the log-likelihood of a weighted multivariate GLM. Hinde and Wood (1987) noted that m_l, $l=1,\ldots,P$, can be estimated by incorporating a P-level factor in the model in place of m_l. By absorbing the additional mass point parameters into $\boldsymbol{\alpha}^*=(\boldsymbol{\alpha}',m_1,\cdots,m_{P-1})'$ and adjusting the model and response matrices accordingly,

one can maximize the weighted multivariate GLM using the Fisher scoring algorithm with forms of the score function and expected information matrix given in Fahrmeir and Tutz (1994, p. 346).

The EM algorithm with a Fisher scoring algorithm embedded in each M-step can be summarized as follows:

0. Calculate initial values $\boldsymbol{\alpha}^{*(0)} = (\boldsymbol{\alpha}^{(0)'}, m_1^{(0)}, \ldots, m_{P-1}^{(0)})'$ and $\boldsymbol{p}^{(0)}$.
   For $s = 0, 1, 2, \ldots$
1. Calculate posterior probabilities $q_{kl}^{(s)}$, $k = 1, \ldots, L,\ l = 1, \ldots, P$, using $\boldsymbol{\alpha}^{*(s)}$ and $\boldsymbol{p}^{(s)}$. Calculate $\boldsymbol{p}^{(s+1)}$ using $q_{kl}^{(s)}$, $k = 1, \ldots, L,\ l = 1, \ldots, P$.
2. Carry out the Fisher scoring algorithm to obtain $\boldsymbol{\alpha}^{*(s+1)}$ using the weights $q_{kl}^{(s)}$, $k = 1, \ldots, L,\ l = 1, \ldots, P$.

For initial estimates of the regression parameters $\boldsymbol{\alpha}$ in $\boldsymbol{\alpha}^*$, one can use the ML estimates from the parametric random effects model or else estimates obtained by fitting a GLM to the original data, ignoring the random effect. A number of ways of obtaining initial estimates for the mass points $\boldsymbol{m}$ exist. Most of these utilize the residuals from fitting a GLM to the original data (Hinde and Wood, 1987; Follmann and Lambert, 1989). Aitkin (1996) suggested using the P nodes and weights from P-point Gaussian quadrature, which is the approach we used.

Convergence of the EM algorithm, which is often slow, can be determined by the change in successive deviances or by the absolute difference in parameter estimates. Usually it is adequate to monitor the deviance, but it is wise to check also the parameter estimates to check whether any seem to be heading to infinity. Convergence to a local maximum is possible, so trying different starting values is also recommended. Standard errors can be obtained through the calculation of the observed information matrix. One can calculate the observed information matrix using Louis' method (Louis, 1982) or, as we have done, by directly calculating first and second derivatives of the log-likelihood function,

$$\sum_{k=1}^{L} \log \sum_{l=1}^{P} p_l \left[\prod_{i=1}^{r} f(\boldsymbol{y}_{ik} | \boldsymbol{\alpha}^*) \right]$$

with respect to $\boldsymbol{p}$ and $\boldsymbol{\alpha}^*$ (Hartzel, 1999). Upon convergence of the NPML algorithm, the observed information matrix is evaluated at the maximum likelihood estimates of the fixed parameters and mixing distribution and then inverted to obtain an estimated variance–covariance matrix for the parameters. We developed an OX program (Doornick, 1998) to do these analyses.

In the EM algorithm described above, we assumed that the support size of the mixing distribution $g(u_k)$ is a fixed quantity, P, when it is in fact an unknown parameter. One approach for estimating P is to start with $P = 2$ and successively apply the algorithm while incrementing P until the optimal support size is reached. There are a number of ways to determine if the optimal support size has been reached. Typically an increase in the support size beyond the optimal value leads to multiplicities in mass points, or masses with zero probabilities. In conjunction with these occurrences, there is usually little to no change in the deviance between the successive fits. Thus one can determine convergence in P by comparing deviances

between fits. Occasionally, however, an increase in the support size beyond the optimal value will lead to singular matrices within the Fisher scoring algorithm as mass points take on identical values. The deviance would then be undefined, but the choice of P would be obvious.

A question we are currently investigating is whether there are random effects distributions for the models discussed in this article such that estimates obtained with the nonparametric approach differ substantively from those with the parametric normal approach. That is, can misspecification of the random effects distribution, assuming normality when reality is far from normal, result in serious bias? We do not expect this to happen for inferences about fixed effects, such as inferences about β in random-intercept forms of models such as (3); however, it could plausibly happen for inferences about the mean β of highly skewed distributions of effects $\{b_k\}$ in heterogeneity models such as (4).

3.4. Predicted values

Besides estimating the model parameters, one might also want predicted values for the center-specific cumulative log odds ratios $\{b_k\}$ in model (4). Point predictors for these random effects are based on their conditional expectation, given the observed data. For example, the prediction for b_k is

$$E(b_k|\boldsymbol{y}_k;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}}) = \frac{\int_{-\infty}^{\infty} b_k \int_{-\infty}^{\infty} [\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|b_k,c_k;\hat{\boldsymbol{\alpha}})] g(b_k,c_k;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}})\,\mathrm{d}c_k\,\mathrm{d}b_k}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [\prod_{i=1}^{r} f(\boldsymbol{y}_{ik}|b_k,c_k;\hat{\boldsymbol{\alpha}})] g(b_k,c_k;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}})\,\mathrm{d}c_k\,\mathrm{d}b_k}, \quad (11)$$

where $\boldsymbol{y}_k$ is the data vector for the kth center and $\hat{\boldsymbol{\Sigma}}$ has elements $\hat{\sigma}_b$, $\hat{\sigma}_c$, and $\hat{\sigma}_{bc}$. Both the numerator and denominator of (11) contain intractable integrals and thus some approximation is required. As in the estimation routine in Section 3.2, these integrals can be approximated using adaptive Gauss–Hermite quadrature. The modes and curvatures of the integrands in the numerator and denominator are found and adaptive quadrature is applied individually to each set of integrals. Although Eq. (11) seems to suggest that the prediction of b_k depends only on the data for center k, the parameter estimates depend on all the data; as a consequence, these predictions typically exhibit considerable shrinkage from the estimates for corresponding fixed effects models, especially when the sample size is small.

Standard errors for the predicted random effects are also calculated conditionally. As this conditional variance is a function of the estimated parameter vector, additional steps are needed to take into account the sampling variability associated with that estimate. Booth and Hobert (1998) have proposed a conditional mean squared error of prediction (CMSEP) criterion for predictions of linear combinations of the fixed and random effects that does this.

3.5. Testing homogeneity versus heterogeneity

The traditional way to check the homogeneity of odds ratios is the fixed effects approach of testing that $\beta_1 = \cdots = \beta_L$ in model (2) for cumulative odds ratios and

in model (6) for local odds ratios. The likelihood-ratio test statistic, the difference between the deviances for the model with this restriction and the more general model, follows an asymptotic chi-squared distribution with $L-1$ degrees of freedom. Alternatively one can check homogeneity in terms of other measures of ordinal association (e.g., Uesaka, 1993).

For the parametric normal random effects models (4) for cumulative odds ratios and (7) for local odds ratios, one can check homogeneity with the likelihood-ratio test of whether the normal distribution of $\{b_k\}$ is degenerate, which we express as H_0: $\sigma_b = 0$. Under H_0, the likelihood-ratio statistic,

$$-2[\max \log L(\boldsymbol{\alpha}, \beta, \gamma, \sigma_b = 0, \sigma_c) - \max \log L(\boldsymbol{\alpha}, \beta, \gamma, \sigma_b, \sigma_c)], \tag{12}$$

is approximately an equal mixture of degenerate at 0 (which occurs when $\hat{\sigma}_b = 0$) and chi-squared distributed with 1 degree of freedom. The value 0 occurs when the maximized likelihoods are identical under the null and the alternative, and hence their ratio equals 1. Thus, when $\hat{\sigma}_b > 0$ and the observed test statistic has value $t > 0$, the P-value for this large-sample test is $(1/2)P(\chi_1^2 > t)$, half the P-value that applies for χ_1^2 asymptotic tests (such as tests about fixed effect components). Although analogous results may hold with the nonparametric approach, as yet there is not any definitive research on the distribution of likelihood-ratio statistics when comparing such a model to one without the discrete random effect. For instance, one difficulty is that the number of mass points is itself a parameter. The estimated value of this may increase with the sample size, yet standard asymptotics treat the number of parameters as a fixed and known constant as $n \to \infty$.

4. Example: Ordinal multi-center clinical trial data

We now return to the application of comparing treatments when data are collected from several centers of some type, such as medical clinics. Table 1 is an example of this type. We analyze these data using fixed and random effects models permitting heterogeneity.

4.1. Example: fixed effects models

For a baseline, we begin with model (1) with $x_1 = 0, x_2 = 1$, which assumes a common cumulative log odds ratio β for each center. The treatment effect estimate is $\hat{\beta} = 0.93$ with standard error of 0.28. The data are sparse, but the deviance and Pearson goodness-of-fit statistic values of $G^2 = 53.7$ and $X^2 = 53.5$ with $df = 22$ give some cause for concern about the adequacy of this model.

The heterogeneity model (2) allows the cumulative log odds ratio to vary among centers and yields $\{\hat{\beta}_k\}$ as shown in Table 2,ranging from $\hat{\beta}_2 = -1.62$ to $\hat{\beta}_1 = 3.03$. The likelihood-ratio statistic testing homogeneity of associations by comparing this model with the homogeneity model (1) equals 24.8 ($df = 7$), giving strong evidence (P-value < 0.001) against a common cumulative odds ratio. The heterogeneity model itself shows potential lack of fit ($G^2 = 28.9$, $df = 15$). It follows from model-fitting

Table 2
Summary of center-specific cumulative log odds ratio estimates and standard errors for treatment effects with fixed and random effects heterogeneity models applied to Table 1[a]

	Fixed effects model (2)		Random effects model (4)	
Effect	Estimate	Std. Error	Estimate	Std. Error
Center 1	3.03 (2.74)	0.87 (0.88)	2.35	0.75
Center 2	−1.62 (−1.67)	0.95 (1.00)	−0.62	0.92
Center 3	0.20 (0.30)	0.55 (0.65)	0.32	0.52
Center 4	0.71 (0.81)	0.85 (0.93)	0.76	0.72
Center 5	2.84 (2.26)	0.95 (0.95)	2.11	0.83
Center 6	−1.06 (−0.94)	1.21 (1.14)	−0.10	0.94
Center 7	1.76 (1.55)	0.87 (0.87)	1.53	0.73
Center 8	0.83 (0.89)	0.82 (0.81)	0.84	0.73

[a]Parenthesized values for Model (2) result from better-fitting version of model with separate cutpoint structure for each center.

discussed in Section 5 that this is due to requiring the same relative distances between cutpoints in each center rather than the proportional odds assumption. That is, the more general model with $\alpha_j - \gamma_k$ in (2) replaced by α_{jk} provides a good fit ($G^2 = 5.0$, $df = 8$). Table 2 also shows estimates and standard errors of $\{\hat{\beta}_k\}$ for this more general model. Results are similar. In our experience, the cutpoint structure for the model has little effect on inferences and substantive conclusions about associations, even though the two sets of parameters are not orthogonal. Since the main focus here is on describing heterogeneity of association rather than heterogeneity of response probabilities and since the data are so sparse for each center, the following discussion of random effects models primarily discusses models with the simpler cutpoint structure.

4.2. Example: random effects models

In fitting random effects models in which the strata are levels of a random effect, ideally one would prefer to have more than the 8 strata that Table 1 has. Moreover, for most data sets of this type the strata are not truly a random sample. However, we agree with the statement by Grizzle (1987), 'Although the clinics are not randomly chosen, the assumption of random clinic effect will result in tests and confidence intervals that better capture the variability inherent in the system more realistically than when clinic effects are considered fixed.' In addition, the random effects approach more naturally directs resulting inference toward the true population of interest rather than just these eight centers. The fixed effects models have the limitation that their inferences, strictly speaking, apply only to those centers. Thus, keeping in mind the limitations of a small number of (nonrandomly chosen) centers and sparse data and the fact that estimates of variance components would be rough for such samples even with a more random design, we now use Table 1 to illustrate such models. This subsection assumes normal distributions for random effects.

The homogeneity cumulative logit model (3), which assumes a distribution only for the center effects, has very similar results as the corresponding fixed effects model (1), with $\hat{\beta} = 0.95$ and standard error 0.28. (With the more complex a_{jk} cutpoint structure, using a vector of correlated random effects, $\hat{\beta} = 0.93$ and standard error 0.28). More realistically, to allow for interaction we use the heterogeneity model (4), which has $\hat{\beta} = 0.92$ with standard error 0.53. Although the treatment estimate is similar, the standard error is much larger. This larger standard error results from the extra variance component for $\{b_k\}$, described by $\hat{\sigma}_b = 1.22$. That is, model (4) predicts that cumulative log odds ratios vary among centers with a mean of 0.92 and a standard deviation of 1.22.

Given the heterogeneity model (4), the likelihood-ratio statistic for testing homogeneity of associations ($\sigma_b = 0$) equals 5.9, which corresponds to a P-value of 0.008 (half the tail probability for a χ^2_1 variate). As with the fixed effects models, there is strong evidence of heterogeneity. Recognizing the heterogeneity, we must be content with a less precise estimate of the overall association level. In our experience, the standard error of $\hat{\beta}$ is similar to that of the corresponding estimated fixed effect in a homogeneity model only when $\hat{\sigma}_b = 0$ or close to it.

Table 2 shows the predicted values of the cumulative log odds ratios according to model (4), based on estimating the expected value of b_k in this model given the data. The standard errors provided are based on a Laplace approximation to the conditional mean squared error of prediction. Since the random effects estimates 'borrow from the whole,' they show a considerable shrinkage compared to the estimates from the fixed effects model. For instance, the negative estimates of -1.62 and -1.06 shrink to -0.62 and -0.10. When datasets have small sample sizes per stratum, we believe that shrinkage of effect estimates is highly appealing for models permitting heterogeneity, since the stratum-specific estimates are then likely to exhibit more variability than the true parameters. In particular, stratum-specific estimates are infinite when either none of the sample pairs of observations are concordant or none are discordant.

4.3. Likelihood-ratio tests of treatment effects

Next, we consider the significance of the treatment effect, for various models. We begin with the homogeneous effects models, but only for illustrative purposes since they fit poorly. For the fixed effects model, the likelihood-ratio statistic for testing that $\beta = 0$ (i.e., conditional independence of response and treatment, given clinic) equals 11.5, with $df = 1$ ($P < 0.001$). For the model (3) with random center effects, the likelihood-ratio statistic equals 12.0 ($P < 0.001$). These tests, coupled with the positive sign for $\hat{\beta}$, provide strong evidence that the response tends to be better with drug than placebo, but we have seen that their assumption of association homogeneity is unrealistic. By contrast, if we use the better-fitting random effects heterogeneity model (4), the likelihood-ratio statistic for testing that the mean β of the cumulative log odds ratios is zero equals 2.5 with $df = 1$ ($P = 0.11$ for the alternative, $\beta \neq 0$). Thus, the evidence of a treatment effect is considerably weaker, and that effect is then a 'mean' effect rather than a common effect for each stratum. In using more realistic

Table 3
Estimated treatment log odds ratio and standard error, for cumulative logit and adjacent-categories logit models, with Table 1

			Cumulative logit			Adjacent-Cat. logit		
Effect	Center	Random effect distribution	Model	$\hat{\beta}$	Std. error	Model	$\hat{\beta}$	Std. error
Homogeneous	Fixed	—	(1)	0.932	0.278		0.656	0.193
	Random	Normal	(3)	0.947	0.276		0.654	0.190
		Nonparametric	(3)	0.938	0.282		0.654	0.191
Heterogeneous	Random	Normal	(4)	0.923	0.526	(7)	0.633	0.341
		Nonparametric	(4)	0.978	0.530	(7)	0.602	0.232

models permitting heterogeneity, it can be more difficult to establish significance of effects because of the extra variability inherent in the model.

4.4. Alternative models

All results discussed so far in this section refer to the cumulative logit form of model. Similar results occur with adjacent-categories logit models and inferences regarding local odds ratios. Table 3 summarizes the association parameter estimates for various models applied to Table 1. The $\{\hat{\beta}_k\}$ in fixed effects model (6) range from $\hat{\beta}_2 = -1.12$ to $\hat{\beta}_1 = 2.00$. The simpler homogeneity model has $\hat{\beta} = 0.66$ with standard error 0.19. The likelihood-ratio statistic for testing homogeneity ($\beta_1 = \cdots = \beta_8$) equals 21.7 ($df = 7$, P-value $= 0.003$), providing evidence of nonhomogeneous local odds ratios. With the adjacent-categories logit random effects heterogeneity model (7), $\hat{\beta} = 0.63$ with a standard error of 0.34. Again the standard error is considerably larger than for the homogeneity models. The variability among $\{b_k\}$ is described by $\hat{\sigma}_b = 0.77$.

Substantive results are similar with the nonparametric random effects models, as Table 3 shows. For these models the likelihood achieved its maximum using a discrete mixture distribution with few mass points. This coincides with results found by others for binomial response models (Follmann and Lambert, 1989; Aitkin, 1996,1999). For example, the cumulative logit homogeneity model (3) required a discrete mixing distribution with only three mass points (see Table 4). The nonparametric estimation algorithm in Section 3.3 can be used to fit the heterogeneous effects cumulative logit model (4) or adjacent-categories logit model (7) by including interaction terms between the treatment factor and the mass point dummy variables.

The nonparametric version of the cumulative logit heterogeneity model (4) required only 4 mass points for the joint distribution of (c_k, b_k), shown in Table 4. The log odds ratios were estimated to vary around a mean of 0.98 (std. error $= 0.53$) with $\hat{\sigma}_b = 1.25$, compared to a mean of 0.92 (std. error $= 0.53$) with $\hat{\sigma}_b = 1.22$ in the normal random effects model. The likelihood-ratio statistic for testing that $\beta = 0$

Table 4
Estimated mass points and probabilities (in parentheses) of mixing distribution for nonparametric random effects fitting of cumulative logit models (3) and (4) with Table 1

Model	Random effect	Support size (P)	Mixing distribution $\hat{m}_1$ ($\hat{p}_1$)	$\hat{m}_2$ ($\hat{p}_2$)	$\hat{m}_3$ ($\hat{p}_3$)	$\hat{m}_4$ ($\hat{p}_4$)
Homogeneous (3)	c_k	3	−0.59 (0.31)	−1.43 (0.46)	−2.38 (0.23)	
Heterogeneous (4)	(c_k, b_k)	4	(0.20, 0.16) (0.25)	(−1.32, 2.91) (0.25)	(−0.87, −0.04) (0.37)	(−2.28, 1.71) (0.13)

equals 4.3; again there is much weaker evidence of an effect than with the homogeneity model (3), for which the test statistic equals 11.7. The heterogeneity model shows slightly stronger evidence of a treatment effect with this approach than with the parametric one (for which the test statistic equals 2.5). However, we make this observation with some caution, since the chi-squared asymptotic theory has weaker validity in the nonparametric case; the number of parameters is unknown, since the support size is unknown, and indeed the ML fits could have different numbers of support points under the null and alternative (in which case at least one parameter for the larger support size falls on the boundary of the parameter space for the smaller support size). Results from a simulation study that we performed do suggest that the use of the likelihood-ratio test is reasonable for the NPML approach, and can provide at least approximate inferences even when the support sizes differ between the null and the alternative hypotheses (Hartzel, 1999).

4.5. Choosing a model form

In this section we have analyzed Table 1 with a variety of models, differing in terms of the type of logit (cumulative or adjacent categories), type of effects (fixed or random), whether the model allows heterogeneity among centers in the treatment effects, and in the random effects case the choice of distribution for the random effect (normal or nonparametric). We now discuss issues in choosing a model from among those resulting from the possible combinations of these factors. Testing goodness of fit is not the primary issue here, as several models can provide an adequate fit to any given data set.

Conditional on the other choices, substantive conclusions are usually the same for the two types of logit. We believe this choice is the least important of the four just mentioned. The cumulative logit is the most popular in the literature both for fixed effects and random effects models, and it relates naturally to a regression model for an underlying continuous response (McCullagh, 1980). The adjacent category logit is more natural if one wants conclusions in terms of odds ratios to apply to pairs of response categories rather than to cumulative probabilities and hence ordinal groupings of categories.

In applications such as multi-center clinical trials and meta analyses, in which the strata are a sample of possible ones, we prefer the random effects approach because of its scope of inference, applying more generally than to only the strata sampled. Even when the strata are not a sample, when the number of strata is large the random effects approach can be beneficial because of the smoothing effects on stratum-specific estimates and the natural summary of a mean and standard deviation for the treatment effects. Our experience with a variety of examples indicates that the random effects model assuming no interaction tends to provide similar results about the common treatment effect as the random effects interaction model does for the mean of the treatment effects when the variance component estimate for the treatment effects equals 0 or is close to 0. The latter model may provide a much wider confidence interval for the average effect when that variance component estimate is substantial, as we observed for Table 1, but this is reasonable because of the extra source of variability. Of the random effects models, we recommend allowing heterogeneity by using the interaction model. If one uses the simpler homogeneity model but there is actually substantial heterogeneity, the standard error of the estimated treatment effect will be unrealistically low.

When the primary interest is in summarizing the treatment effect, our hunch is that the choice of distribution for the random effect is not crucial. This appears to be the case for binary data (Neuhaus et al., 1992). Since some of the asymptotic inferential issues are still unresolved with the nonparametric approach, for now our preference is for the normal random effects model. The normal choice also has the advantage of extending naturally to multivariate random effects that may have a particular form of correlation structure. However, the potential effects of misspecification require closer study. Even if there is a minor effect on summary treatment effects, the effect may be greater on stratum-specific estimates, especially for sparse asymptotics in which the number of centers grows with the overall sample size.

Because of these considerations, our preferred choice for analyzing Table 1 is model (4) with a bivariate normal distribution for (b_k, c_k). Thus, we would summarize Table 1 by predicting that cumulative log odds ratios vary about a mean of 0.92 with a standard deviation of 1.22, and the estimated mean of 0.92 has a standard error of 0.53.

4.6. Fitting random effects models with SAS

Table 5 illustrates the use of PROC NLMIXED in Version 8 of SAS to fit random effects models such as (4). Though the multinomial distribution is not directly supported by NLMIXED, one can define the general likelihood function needed to fit the models considered here through the use of SAS programming statements. In Table 5 we enter the counts for each cell and create variables denoting the center, treatment group, and response value. Within the NLMIXED procedure, we define the cumulative logit probabilities as functions of the linear predictors η_1 and η_2. Since the first threshold and the mean of the random effects distribution are aliased, we set the first threshold to zero and use the BOUNDS statement to maintain the ordering of the remaining threshold parameter. The RANDOM statement defines the distribution

Table 5
Example of SAS code (Version 8) for using PROC NLMIXED to fit cumulative logit random effects heterogeneity model (4) to Table 1

```
data ordinal;
   do center = 1 to 8;
     do trt = 1 to 0 by -1;
        do resp = 3 to 1 by -1;
           input count @@;
           output;
        end;
     end;
   end;
   datalines;
    13  7  6 1 1 10  2  5  10 2 2 1
    11 23  7 2 8  2  7 11   8 0 3 2
    15  3  5 1 1  5 13  5   5 4 0 1
     7  4 13 1 1 11 15  9   2 3 2 2
run;

proc nlmixed data = ordinal qpoints = 15;
   ** To maintain the threshold ordering define thresholds such that  **;
   ** threshold 1 = 0 and threshold 2 = i2, where i2 > 0.              **;
   ** Use starting value of 0 for sig_cb                              **;
   bounds i2>0;    parms sig_cb = 0;
   eta1 =     c-b*trt;
   eta2 = i2-c-b*trt;
   if (resp = 1) then z = 1/(1+exp(-eta1));
       else if (resp = 2) then z = 1/(1+exp(-eta2))-1/(1+exp(-eta1));
       else z = 1-1/(1+exp(-eta2));
   if (z > le-8) then ll = count*log(z);    ** Check for small values of z **;
       else ll = - le100;
   model resp ~ general (ll);    ** Define general log-likelihood. **;
   random c b ~ normal ([gamma, beta],[sig_c*sig_c, sig_cb, sig_b*sig_b])
        subject = center out = out1; ** OUT1 contains predicted center-    **;
                                     ** specific cumulative log odds ratios **;
run;
```

and covariance structure of the random effects. For multiple random effects, the covariance matrix in the RANDOM statement consists of the lower triangle of the desired covariance structure. The OUT option in the RANDOM statement requests the predicted values of the random effects to be outputted to the named dataset. The standard errors of the predicted random effects are based on a Laplace approximation to the conditional mean squared error of prediction.

The QPOINTS = option in PROC NLMIXED forces SAS to use the specified number of quadrature points. We recommend specifying this to be upwards of 15 to 20 to ensure accurate approximation of the integrals, since in our experience the default in SAS was insufficiently large to approximate standard errors and predictions adequately. Running on a Pentium II, 400 MHz computer with 128 MB of RAM, the example in Table 5 needed about 8 seconds to obtain convergence using 10

quadrature points and about 19 seconds to obtain convergence using 20 quadrature points.

5. Extensions to modeling within-stratum and between-stratum heterogeneity

The models considered in this paper study between-strata heterogeneity but have a single parameter describing association within each stratum. Although it is also unrealistic to think that all odds ratios within a stratum are truly exactly equal, in practice it is often sufficient to summarize the overall association within a stratum and describe the variability in that overall association across strata. Alternatively, one could model the heterogeneity both within and between strata.

One can model both sources of heterogeneity in a traditional manner using fixed effects. For instance, let $\theta_{ij(k)}$ denote an odds ratio with split following row i and column j in stratum k, where this association may refer to local odds ratios, cumulative odds ratios, global odds ratios, or some other form. Then, one could consider an ANOVA-like model that describes how the association varies around some summary value as a function of the levels of the variables. To illustrate, consider a model of the form

$$\log \theta_{ij(k)} = \beta + \alpha_i + \gamma_j + \delta_k, \quad i = 1,\ldots,r-1, \;\; j = 1,\ldots,c-1, \;\; k = 1,\ldots,L, \tag{13}$$

where $\sum \alpha_i = \sum \gamma_j = \sum \delta_k = 0$. Here, β is an overall level of association, and the other parameters refer to departures due to rows, columns, and strata. There are $(r-1)(c-1)$ odds ratios in each of the L strata, so the residual df for testing fit equal

$$\begin{aligned} df &= (r-1)(c-1)L - [1 + (r-2) + (c-2) + (L-1)] \\ &= L(rc - c - r) - r - c + 4. \end{aligned}$$

For local log odds ratios with $L = 1$, this is the Goodman (1979) R + C model having additive row and column effects, with $df = (r-2)(c-2)$. One can generalize this model to allow, for instance, $\log \theta_{ij(k)} = \beta + \alpha_{ik} + \gamma_{jk} + \delta_k$, in which the row and column effects may vary by stratum. Its residual $df = L(r-2)(c-2)$. This model is equivalent to applying Goodman's R+C model separately to each stratum. Regardless of the type of odds ratio in this approach, one can express such models in the generalized loglinear model form

$$\boldsymbol{C} \log \boldsymbol{A\mu} = \boldsymbol{X\beta},$$

where $\boldsymbol{\mu}$ is the vector of expected frequencies. One can fit models in this class using ML methods presented in Lang and Agresti (1994). An S-plus function for fitting the model is available from Prof. J.B. Lang (Statistics Department, University of Iowa).

To illustrate, we again consider Table 1. It has only two rows, so row effects do not apply. For both types of odds ratio, the data are described adequately by a model that has only stratum effects, that is, $\log \theta_{ij(k)} = \beta_k$, with $G^2 = 5.0$ for cumulative odds ratios and $G^2 = 5.4$ for local odds ratios, each with $df = 8$. These models

with stratum effects are not equivalent to the cumulative logit model (2) or the adjacent-categories logit model (6), since there is no requirement that cutpoint parameters be the same in each stratum; they correspond instead to replacing $\alpha_j - \gamma_k$ in those models by α_{jk}. In each case, the overall estimate shows evidence of a positive treatment effect, the estimate equalling 2.3 standard errors (e.g., 0.74 with std. error $= 0.32$ for the cumulative log odds ratios), although again there is weaker evidence than with models that assume homogeneous associations.

When the strata are a sample, in the models described above one could replace fixed stratum effects by random effects. In model (13) for decomposing log odds ratios, for instance, one could replace $\{\delta_k\}$ by random effects $\{d_k\}$. We do not pursue this here, but in future research it might be worth considering this as well as other ways to model heterogeneity using random effects. One possibility may be to add a random effect to each cell in a loglinear model. For instance, consider the generalization of the heterogeneous linear-by-linear association model,

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \beta_k x_i y_j + \sigma Z_{ijk},$$

where $\{Z_{ijk}\}$ are independent standard normal variates. With $\{x_i = i\}$ and $\{y_j = j\}$, the local log odds ratios in stratum k are then normal with mean β_k and standard deviation 2σ. If we replace $\{\beta_k\}$ by $\{b_k\}$ from a normal distribution, then the model has a variance component for within-stratum heterogeneity and a separate variance component for between-strata heterogeneity. This is a nonstandard application of random effects in the sense that the cells are not a random sample, and it is analogous to an approach often used to handle overdispersion.

Acknowledgements

Liu's research was supported by the National Science Council in Taiwan, and Agresti's and Hartzel's research was partially supported by the NIH and NSF in the United States. The authors thank Dr. Brent Coull, Dr. Russell Wolfinger, and two referees for helpful comments.

References

Agresti, A., Hartzel, J., 2000. Tutorial in biostatistics: strategies for comparing treatments on a binary response with multi-center data. Statistics in Medicine 19, 1115–1139.

Agresti, A., Kezouh, A., 1983. Association models for multi-dimensional cross-classifications of ordinal variables. Comm. Statist. Part A-Theory & Methods 12, 1261–1276.

Aitkin, M., 1996. A general maximum likelihood analysis of overdispersion in generalized linear models. Statist. Comput. 6, 251–262.

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

Booth, J.G., Hobert, J.P., 1998. Standard errors of prediction in generalized linear mixed models. J. Amer. Statist. Assoc. 93, 262–272.

Doornick, J.A., 1998. Object-oriented matrix programming using Ox 2.0. Timberlake Consultants, Ltd, Kent, England.

Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, New York.

Follmann, D.A., Lambert, D., 1989. Generalizing logistic regression by nonparametric mixing. J. Amer. Statist. Assoc. 84, 295–300.

Goodman, L.A., 1979. Simple models for the analysis of association in cross-classifications having ordered categories. J. Amer. Statist. Assoc. 74, 537–552.

Grizzle, J.E., 1987. Letter to the editor. Controlled Clin. Trials 8, 392–393.

Hartzel, J., 1999. Random effects models for nominal and ordinal data. Unpublished Ph.D. Thesis, University of Florida.

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

Hinde, J.P., Wood, J.T., 1987. Binomial variance component models with a non-parametric assumption concerning random effects. In: Crouchley R. (Ed.), Longitudinal Data Analysis. Avebury, Aldershot, Hants, 1987, pp. 110–127.

Lang, J.B., Agresti, A., 1994. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. J. Amer. Statist. Assoc. 89, 625–632.

Liu, I.-M., Agresti, A., 1996. Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. Biometrics 52, 1223–1234.

Liu, Q., Pierce, D.A., 1993. Heterogeneity in Mantel-Haenszel-type models. Biometrika 80, 543–556.

Liu, Q., Pierce, D.A., 1994. A note on Gauss-Hermite quadrature. Biometrika 81, 624–629.

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser. B 44, 226–233.

McCullagh, P., 1980. Regression models for ordinal data. J. Roy. Statist. Soc. Ser. B Methodol. 42, 109–142.

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.

Pierce, D.A., Sands, B.R., 1975. Extra-Bernoulli variation in regression of binary data. Technical Report No. 46, Dept. of Statistics, Oregon State University, Corvallis, Oregon.

Pinheiro, J.C., Bates, D.M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. J. Comput. Graphical Statist. 4, 12–35.

Skene, A.M., Wakefield, J.C., 1990. Hierarchical models for multicentre binary response studies. Statist. Med. 9, 919–929.

Tutz, G., Hennevogl, W., 1996. Random effects in ordinal regression models. Comput. Statist. Data Anal. 22, 537–557.

Uesaka, H., 1993. Test for interaction between treatment and stratum with ordinal responses. Biometrics 49, 123–129.