Fitting stratified proportional odds models by amalgamating conditional likelihoods

Bhramar Mukherjee[1,*,†], Jaeil Ahn[1], Ivy Liu[2], Paul J. Rathouz[3] and Brisa N. Sánchez[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48103, U.S.A.*
[2]*School of Mathematics, Statistics, and Computer Science, Victoria University of Wellington, Wellington, New Zealand*
[3]*Department of Health Studies, University of Chicago, Chicago, IL 60637, U.S.A.*

SUMMARY

Classical methods for fitting a varying intercept logistic regression model to stratified data are based on the conditional likelihood principle to eliminate the stratum-specific nuisance parameters. When the outcome variable has multiple ordered categories, a natural choice for the outcome model is a stratified proportional odds or cumulative logit model. However, classical conditioning techniques do not apply to the general K-category cumulative logit model ($K>2$) with varying stratum-specific intercepts as there is no reduction due to sufficiency; the nuisance parameters remain in the conditional likelihood. We propose a methodology to fit stratified proportional odds model by amalgamating conditional likelihoods obtained from all possible binary collapsings of the ordinal scale. The method allows for categorical and continuous covariates in a general regression framework. We provide a robust sandwich estimate of the variance of the proposed estimator. For binary exposures, we show equivalence of our approach to the estimators already proposed in the literature. The proposed recipe can be implemented very easily in standard software. We illustrate the methods via three real data examples related to biomedical research. Simulation results comparing the proposed method with a random effects model on the stratification parameters are also furnished. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: proportional odds; stratified data; conditional likelihood

1. INTRODUCTION

Models for ordinal data started receiving attention in the 1960s and 1970s [1, 2]. However, the most popular model for ordinal data was inspired by McCullagh [3] by modeling the log odds

corresponding to the cumulative probabilities, called the *cumulative logits*. Stratified ordinal data often arise in sociological and clinical studies, with the level of stratification depending on the study design. An extreme degree of stratification occurs when one collects matched pair data that frequently arise in 'before and after' experiments, in the context of Rasch's item-response model [4] with the object to describe 'subject-specific' rather than 'population-averaged' effects. The individual or the matched set then acts as a stratum. On the other hand, modest degree of stratification is seen to be present in frequency matched data and data from multi-center clinical trials. In many instances, the outcomes are measured in ordered categories: such as degree of seriousness of an injury, stages of progression of a disease, or the extent of relief due to treatments. A proper statistical analysis should account for stratum heterogeneity in the ordinal outcome model used for such studies in order to avoid the risk of biased and inconsistent estimation.

In inference with matched pair data with dichotomous responses, conditioning on the complete sufficient statistics for the stratum-specific parameters eliminates these nuisance parameters from the likelihood. The maximum likelihood estimates (MLEs) obtained from the resultant conditional likelihood are consistent and efficient in the sense of [5, 6]. However, in a stratified proportional odds model there is no such reduction due to sufficiency [7–9]. When the data are sparse, random effects approach and Bayesian approach have problems with convergence [10, 11]. A naive inference procedure that tries to estimate all parameters from directly maximizing the unconditional likelihood produces biased and inconsistent estimates of the parameters of interest. On the other hand, ignoring the stratification effects when they are present, e.g. using a constant intercept model across strata, is also known to severely underestimate the cumulative odds ratios.

McCullagh [7, 8] and Agresti and Lang [9] (referred as AL from now on) propose a novel strategy by simultaneously fitting conditional models to all possible binary collapsings of the ordinal response. McCullagh [8] considers this approach for a binary covariate, whereas AL consider a categorical covariate. The maximum likelihood (ML) estimation process in AL involves an improved Newton–Raphson algorithm for a class of generalized log-linear models and is based on the cell counts of the collapsed table. The purpose of this paper is to extend this idea to a general ordinal regression setup with any set of covariates, categorical or continuous, and provide an alternative simple fitting procedure for obtaining the parameter estimates and their standard errors. For a binary exposure, we illustrate the equivalence of our approach to the one proposed in AL and also to a Mantel–Haenszel-type estimator proposed in [12], although each of them is conceived from different standpoints. We call our method amalgamated conditional logistic regression (ACLR).

It is also natural to think about an alternative mixed effects approach with a random distribution on the stratum-specific nuisance parameters while fitting a stratified proportional odds model [13]. The proposed ACLR method has certain advantages over the random effects approach. With finely stratified data, misspecifying the random effects distribution may cause potential bias in the parameter estimates. For a stratified *logistic* regression model with matched pair data, necessary and sufficient conditions on the random effects distribution for consistency of the estimates from the marginal likelihood are known [14]. However, an analogous result is not known for the cumulative logit model for matched pair data. The ACLR method we propose is more robust in the sense that it does not make any assumptions on the distribution of the nuisance parameters. Computationally, our method is simpler than a random effects model (REM), which requires integration of the random stratification parameters with respect to their stochastic nuisance distributions to obtain the marginal likelihood of data. Since the integral involved is not in closed form, the integrated likelihood is approximated by Gauss–Hermite quadrature and then optimized in terms of the unknown model

parameters [15]. However, the mixed effects approach typically has better mean-squared-error properties than ACLR and provides smaller standard errors for the parameter estimates.

We compare and contrast these plausible alternative strategies with the proposed ACLR method through several data examples and a simulation study. The first example we consider is a multi-center randomized clinical trial with center-specific heterogeneity. The outcome of interest is doctor's evaluation of asthma patients, ordered into four categories based on severity of the disease; the goal is to characterize treatment effects after adjusting for covariates. The second example is an item response data set, part of the clinical trial data used in the first example. Here each individual has two responses: a self-evaluation score of asthma severity, and another score from a medical practitioner, both ordered into four categories. The goal is to characterize the differences in self-assessment and physician's assessment. The third example shows the application of these approaches to analyze pair-matched case–control data where the disease states have multiple ordered categories.

The remainder of this paper is organized as follows. In Section 2 we present the model, notation, and the proposed methods under random sampling of individuals. In Section 3 we consider the special case of a single binary covariate and show the equivalence of our approach with alternative approaches. We also discuss two special cases: (a) the generalized item response model as discussed in AL and (b) matched case–control data with multiple disease states. Section 4 presents real data examples, whereas Section 5 contains a small-scale simulation study under varying distribution of the stratification parameters.

2. THE STRATIFIED PROPORTIONAL ODDS MODEL

One of the most popular models for ordinal data was proposed by McCullagh [3] by modeling the log odds corresponding to the cumulative probabilities, called the *cumulative logits* described as follows. Let us consider stratified data with an ordinal response. Let Y_{ij} denote the response for the jth observation in the ith stratum, where $j = 1, \ldots, N_i$ and $i = 1, \ldots, N$. Also, let Y_{ij} be a K-category ordinal variable with categories scaled from $0, 1, \ldots, K-1$. Let $\mathbf{X}_{ij}$ denote the observed vector of covariates for observation j in stratum i. The cumulative logit model is given by

$$\text{logit}[P(Y_{ij} \geqslant r | \mathbf{X}_{ij})] = \alpha_{ri} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij}, \quad j = 1, \ldots, N_i, \ i = 1, \ldots, N, \ r = 1, \ldots, K-1 \quad (1)$$

where $\alpha_{1i} > \alpha_{2i} > \cdots > \alpha_{K-1,i}$ for all $i = 1, \ldots, N$ and $P(Y_{ij} \geqslant 0) = 1$. The parameters $\{\alpha_{ri}\}$ are usually nuisance parameters of little interest. The total number of these nuisance parameters increases in direct proportion with the sample size for the generalized Rasch model considered in AL. For other types of stratified designs, such as the multi-center clinical trial data we consider in Example 1, the number of nuisance parameters (centers) does not increase with sample size. However, the sparsity of the observations in each stratum may cause potential problems in ML estimation. The above model applies simultaneously to all $K-1$ cumulative probabilities, and it assumes an identical effect $\boldsymbol{\beta}$ of the predictors for each cumulative probability. This particular type of stratified cumulative logit model, with the p-dimensional exposure effect $\boldsymbol{\beta}$ the same for all r, is often referred to as a stratified *proportional odds model* [3].

When $K = 2$, the stratified proportional odds model (1) reduces to a stratified logistic regression model. One can then eliminate the nuisance parameters α_{ri} using conditional ML. However, as mentioned in the Introduction, the conditional ML method cannot be applied to the proportional

Table I. Example of a table showing that how a stratum with
five observations can be collapsed into three strata.

Strata	Y	X_1		$Y \geqslant 1$	X_1	$Y \geqslant 2$	X_1	$Y \geqslant 3$	X_1
	0	1.5		0	1.5	0	1.5	0	1.5
	1	2.0	$\Longrightarrow$	1	2.0	0	2.0	0	2.0
1	2	1.5		1	1.5	1	1.5	0	1.5
	2	1.8		1	1.8	1	1.8	0	1.8
	3	1.2		1	1.2	1	1.2	1	1.2

odds model (1) with $K > 2$, because there are no reduced sufficient statistics for the nuisance parameters α_{ri}. In many applications, α_{ri} is modeled as an additive effect of the threshold parameters α_r and the stratum-specific parameters γ_i, i.e. $\alpha_{ri} = \alpha_r + \gamma_i$. Note that this is an alternative parametrization of the generalized Rasch model considered in AL.

For each r ($= 1, \ldots, K - 1$), the proportional odds model is a logistic regression model. If we consider all possible collapsings of the ordinal response Y_{ij} into a binary scale ($\geqslant r$, $< r$), each stratum has $K - 1$ such collapsed versions that are dependent within strata. The estimator obtained from each separate collapsing for a given r is unbiased for $\boldsymbol{\beta}$, but not efficient. This paper proposes a conditional approach by first naively treating the $K - 1$ different collapsed versions within a single stratum as independent. That is, the original stratified data set with N strata can be extended into a larger stratified data set with $N \times (K - 1)$ strata. The extended stratified data set has only binary responses. Therefore, the usual conditional ML method can be used and the conditional ML estimators for $\boldsymbol{\beta}$ for each collapsing, and for a given r, remain consistent. The naive working independence assumption has been used in many problems including generalized estimating equations (GEEs) [16], in forming the simple estimate of the ordinal odds ratio for matched pairs [9], and the ordinal Mantel–Haenszel estimator [12]. Table I presents how a single stratum with ordinal responses from 0 to 3 and a single continuous covariate can be collapsed into three different strata, each with a binary response.

Let us now describe our notation and model structure. Let M_{ri} be the number of observations in stratum i having $Y_{ij} \geqslant r$ (i.e. $M_{ri} = \sum_{j=1}^{N_i} I(Y_{ij} \geqslant r)$). The conditional likelihood for the rth collapsing in stratum i is

$$L_{ri} = \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{S}_{ri})}{\sum_{\mathbf{q}_i \in \Omega_{ri}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)}$$

where $\mathbf{S}_{ri} = \sum_{j=1}^{N_i} \mathbf{X}_{ij} I(Y_{ij} \geqslant r)$ is the sum of covariate vectors over M_{ri} observations, and Ω_{ri} is a collection of all possible sum vectors of the form $\mathbf{X}_{il_1} + \cdots + \mathbf{X}_{il_{M_{ri}}}$, in which $(l_1, \ldots, l_{M_{ri}})$ is a subset of size M_{ri} from $(1, \ldots, N_i)$ chosen without replacement. Therefore, the cardinality of Ω_{ri} is $\binom{N_i}{M_{ri}}$. The conditional likelihood L_{ri} is simply the probability of having the observed data, given that there are exactly M_{ri} observations having $Y_{ij} \geqslant r$. In a finite population sampling framework, L_{ri} could also be interpreted as the probability that a random sample of size M_{ri} is selected from the finite population of size N_i with probability proportional to $\exp(\boldsymbol{\beta}^{\top} \mathbf{q}_i)$ where $\mathbf{q}_i$ is the sum of the $\mathbf{X}_{ij}$'s selected in the sample.

Under the naive independence assumption for the extended data set, the conditional likelihood over all strata and all collapsings is

$$L = \prod_{i=1}^{N} \prod_{r=1}^{K-1} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{S}_{ri})}{\sum_{\mathbf{q}_i \in \Omega_{ri}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)} \tag{2}$$

The pseudo-score function can thus be expressed as

$$\mathbf{U}(\boldsymbol{\beta}) = \left(\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)_{p \times 1} = \sum_{i=1}^{N} \sum_{r=1}^{K-1} \left(\mathbf{S}_{ri} - \frac{\sum_{\mathbf{q}_i \in \Omega_{ri}} \mathbf{q}_i \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)}{\sum_{\mathbf{q}_i \in \Omega_{ri}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)} \right) = \sum_{i=1}^{N} \sum_{r=1}^{K-1} \mathbf{U}_{ri}(\boldsymbol{\beta}) \tag{3}$$

where p denotes the length of $\boldsymbol{\beta}$. Therefore, the estimates of $\boldsymbol{\beta}$ can be obtained by fitting conditional logistic regression (CLR) to the extended stratified data set using any standard software implementing CLR. It produces the asymptotic variance–covariance matrix $\mathbf{I}$ having the form

$$\mathbf{I} = \left(\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)_{p \times p}^{-1} = \left(\sum_{i=1}^{N} \sum_{r=1}^{K-1} \mathbf{D}_{ri} \right)^{-1}$$

The (s, t)th element in $\mathbf{D}_{ri}$ (a $p \times p$ matrix) is

$$-\frac{\partial^2 \log L_{ri}(\boldsymbol{\beta})}{\partial \beta_s \partial \beta_t} = \left(\frac{\sum_{\mathbf{q}_i \in \Omega_{ri}} q_{is} q_{it} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)}{\sum_{\mathbf{q}_i \in \Omega_{ri}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)} \right)$$
$$- \left(\frac{\sum_{\mathbf{q}_i \in \Omega_{ri}} q_{is} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)}{\sum_{\mathbf{q}_i \in \Omega_{ri}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)} \right) \left(\frac{\sum_{\mathbf{q}_i \in \Omega_{ri}} q_{it} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)}{\sum_{\mathbf{q}_i \in \Omega_{ri}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)} \right) \tag{4}$$

where $\mathbf{q}_i = (q_{i1}, \ldots, q_{ip})^{\mathrm{T}}$ is a $p \times 1$ vector representing the sum of the observed covariate vectors in stratum i. The estimate of the asymptotic variance–covariance matrix $\mathbf{I}$ furnished by any standard CLR fitting program is based on the observed Hessian matrix

$$\left(\sum_{i=1}^{N} \sum_{r=1}^{K-1} \hat{\mathbf{D}}_{ri} \right)^{-1}$$

with $\hat{\boldsymbol{\beta}}$ replacing $\boldsymbol{\beta}$ in (4), where $\hat{\boldsymbol{\beta}}$ is the solution to the pseudo-score equations obtained by setting $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$. Note that the ACLR estimate of $\boldsymbol{\beta}$ is consistent under sparse data situations as due to the proportional odds structure each of the collapsed model is a stratified logistic model for which conditional ML estimates are known to be consistent. For details of a similar argument, see [17].

The estimator of the asymptotic variance and covariance matrix produced by the above amalgamation of conditional ML method is not consistent, as the extended data set has strata that are dependent across different collapsings. We propose a sandwich estimator of the variance–covariance matrix based on the following well-known result widely used in the GEEs literature.

Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\mathbf{G}_n(\boldsymbol{\theta}) = (1/n) \sum_{i=1}^{n} \mathbf{G}(\boldsymbol{\theta}, Y_i) = \mathbf{0}$. Then $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$, and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{A}^{-1} \mathbf{B} (\mathbf{A}^{\mathrm{T}})^{-1})$, where

$$\mathbf{A} = \mathbf{A}(\boldsymbol{\theta}) = E \left[\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}(\boldsymbol{\theta}, Y) \right] \quad \text{and} \quad \mathbf{B} = \mathbf{B}(\boldsymbol{\theta}) = E[\mathbf{G}(\boldsymbol{\theta}, Y) \mathbf{G}(\boldsymbol{\theta}, Y)^{\mathrm{T}}] = \mathrm{cov}\{\mathbf{G}(\boldsymbol{\theta}, Y)\}$$

Using the above result, an empirical estimate for the sandwich variance is obtained as

$$\tilde{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \sum_{r=1}^{K-1} \hat{\mathbf{D}}_{ri} \right)^{-1} \left[\sum_{i=1}^{N} \text{cov} \left(\sum_{r=1}^{K-1} \mathbf{U}_{ri}(\hat{\boldsymbol{\beta}}) \right) \right] \left(\sum_{i=1}^{N} \sum_{r=1}^{K-1} \hat{\mathbf{D}}_{ri} \right)^{-1}$$

where $\mathbf{U}_{ri}(\boldsymbol{\beta})$ is defined in (3). Replacing $\text{cov}(\sum_{r=1}^{K-1} \mathbf{U}_{ri}(\hat{\boldsymbol{\beta}}))$ by the empirical covariance matrix yields

$$\tilde{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \sum_{r=1}^{K-1} \hat{\mathbf{D}}_{ri} \right)^{-1} \left[\sum_{i=1}^{N} \left(\sum_{r=1}^{K-1} \mathbf{U}_{ri}(\hat{\boldsymbol{\beta}}) \right) \left(\sum_{r=1}^{K-1} \mathbf{U}_{ri}(\hat{\boldsymbol{\beta}}) \right)^{\mathrm{T}} \right] \left(\sum_{i=1}^{N} \sum_{r=1}^{K-1} \hat{\mathbf{D}}_{ri} \right)^{-1} \quad (5)$$

Remark 1
Note that, instead of ignoring the correlation of the collapsings within a stratum, we modify the correction term in the middle of the usual sandwich estimator by considering the contribution to the score function from the sums over the collapsings, the sums being independent across strata.

Remark 2
The pseudo-score equation (3) is evidently not the optimum estimating function for estimating $\boldsymbol{\beta}$. Following [8] it is conjectured that the efficient score function should be a weighted combination of the CLR score functions corresponding to each collapsing of the data, namely

$$\sum_{i=1}^{N} \sum_{r=1}^{K-1} w_{ri}(\boldsymbol{\beta}) \mathbf{U}_{ri}(\boldsymbol{\beta})$$

Obtaining the semi-parametric efficient score function in this problem is an interesting open question in its own right.

We correct for possible small sample bias in the sandwich estimator following [18]. Note that the score function in (3) can be expressed in the form

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{r=1}^{K-1} (\mathbf{S}_{ri} - p_{ri}(\boldsymbol{\beta}))$$

where

$$p_{ri}(\boldsymbol{\beta}) = \frac{\sum_{\mathbf{q}_i \in \Omega_{ri}} \mathbf{q}_i \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)}{\sum_{\mathbf{q}_i \in \Omega_{ri}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{q}_i)}$$

On the basis of a second-order Taylor's expansion of the score function about $\hat{\boldsymbol{\beta}}$, with details as given in [18], we form a set of bias-corrected scores for each stratum, and for each collapsing as

$$\mathbf{U}_{ri}^{*}(\hat{\boldsymbol{\beta}}) = \text{Diag} \left\{ I_p - \hat{\mathbf{D}}_{ri} \left(\sum_{i=1}^{N} \sum_{r=1}^{K-1} \hat{\mathbf{D}}_{ri} \right)^{-1} \right\}^{-1/2} (\mathbf{S}_{ri} - p_{ri}(\hat{\boldsymbol{\beta}}))$$

Table II. Comparison of the variance approximation in (5) (uncorrected sandwich estimator of variance) and (6) (bias-corrected sandwich estimator of variance) with actual sampling variance of the ACLR estimator under three different simulation settings.

Type	General (i)		Matched (ii)	Item response (iii)
	β_1	β_2	β_1	β_1
Simulation-based sample variance	0.34	0.14	0.25	0.28
Uncorrected sandwich estimator	0.32	0.13	0.25	0.29
Bias-corrected sandwich estimator	0.33	0.13	0.25	0.29

Setting (i) is stratified data with two covariates, one continuous, one binary with 50 strata each having five observations, response on an ordinal scale of 1–4. Setting (ii) is matched case–control data with two sub-types within cases and a single binary covariate for 200 matched pairs. Setting (iii) is item response data with two repeated responses on 100 subjects. In each case we report the actual variance of the regression parameter computed over 500 simulation runs, average value of uncorrected sandwich estimator of variance and average value of bias-corrected sandwich estimator of variance.

The corrected empirical covariance, which now replaces the middle term in (5), is given by

$$\sum_{i=1}^{N} \left(\sum_{r=1}^{K-1} \mathbf{U}_{ri}^*(\hat{\boldsymbol{\beta}}) \right) \left(\sum_{r=1}^{K-1} \mathbf{U}_{ri}^*(\hat{\boldsymbol{\beta}}) \right)^{\mathrm{T}} \tag{6}$$

A small-scale simulation study in Table II presents a comparison of the variance estimator in (5), the bias-corrected version (6) and the actual observed variance of the estimators across many replications under certain model settings. Since the bias correction seems to make very minor differences in the variance approximation, we continue to use the variance estimator obtained from (5). However, small sample bias in the sandwich estimator could be an important issue if the simulation designs were 'unbalanced' in the sense of [19], that is, if different number of clusters contribute to the estimation of the different parameters of interest, then one would prefer to use the bias-corrected estimator. The Wald-type ratios $\hat{\beta}_k/\{\mathrm{se}(\hat{\beta}_k)\}$ are approximately normally distributed test statistics for testing $H_0 : \beta_k = 0$. Figure 1 provides a sampling distribution of this test statistic for four different simulation settings, which will be discussed later in Section 5.

3. SPECIAL CASE: MATCHED PAIR DATA WITH ONE BINARY EXPOSURE

Consider the simple case of matched pair data with a single binary exposure and ordinal outcome. Let Y_{ij}, $i = 1, \ldots, N$, $j = 1, 2$, denote the response category of jth observation in matched set i. For a typical 'before' and 'after' experiment, the matched set will be the subject itself. Thus, Y_{ij} is a K-category ordinal variable with categories scaled from $0, \ldots, K-1$, where the category 0 denotes the baseline response category. Let X_{ij} denote the binary exposure having value 0 or 1. In Rasch's item response model considered in AL, Y_{ij} are repeated ordinal responses on the same subject, whereas X_{ij} is the indicator variable corresponding to an item.

In our conditional likelihood formulation, only matched sets with discordant Y and X observations contribute to the likelihood; thus, we restrict attention to those sets only. To this end, without
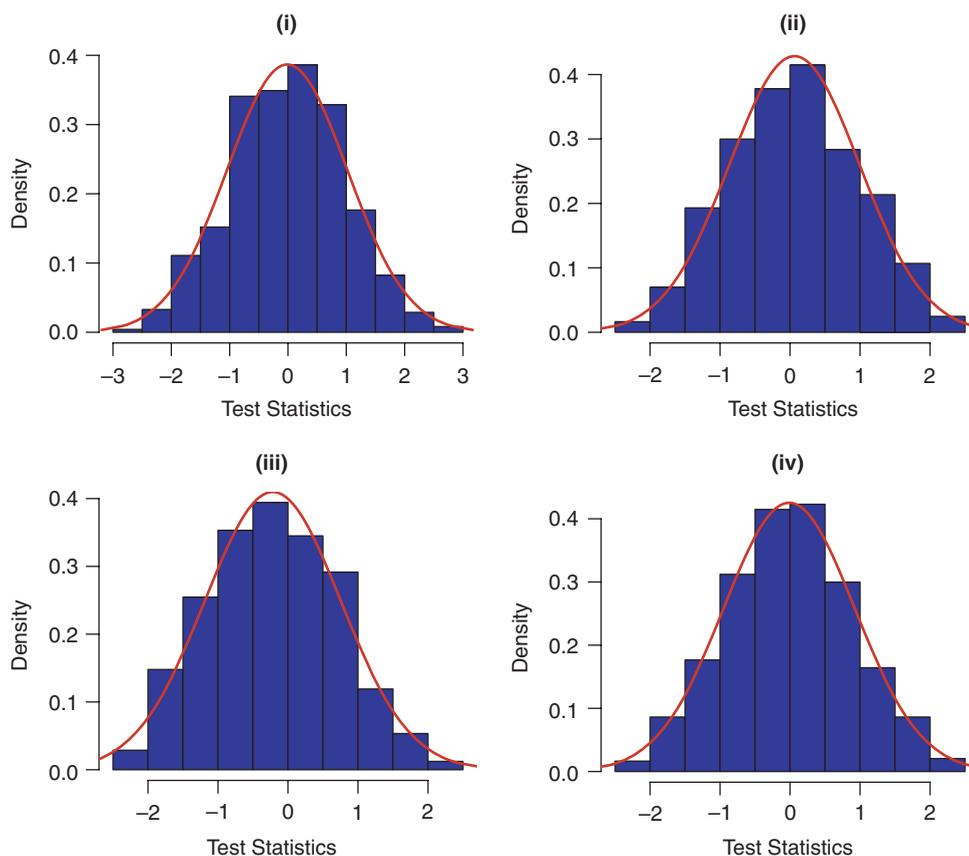
Figure 1. Sampling distribution for Wald statistic corresponding to ACLR estimator, where (i) corresponds to β_1 in the moderately stratified setting for 100 strata with five observations as described in Table VII. True $\beta_1 = -0.5$. (ii) Corresponds to β_2 for moderately stratified setting for 100 strata with five observations as described in Table VII. True $\beta_1 = 1.0$. (iii) Matched case–control data with 200 matched pairs and a single binary response. True $\beta = 1$. (iv) Item response data with 100 subjects with two repeated responses. True $\beta = 0.5$. The value of α_i was taken as constant -3 in each simulation.

loss of generality, let Y_{i1} be the response corresponding to $X = 1$, i.e. $X_{i1} \equiv 1$ and Y_{i2} be the response corresponding to $X = 0$, i.e. $X_{i2} \equiv 0$. For a particular stratum, let $Y_{i1} = s$ and $Y_{i2} = t$ with $s, t = 0, \ldots, K-1, s \neq t$. Then for the rth collapsing with the new binary response as $I(Y_{ij} \geqslant r)$, and for $s < t$, the likelihood contribution of the ith strata can be expressed as

$$L_{ri} = 1 \quad \text{for } r \leqslant s \text{ and } r \geqslant t+1$$

$$= \frac{1}{1 + \exp(\beta)} \quad \text{for } s+1 \leqslant r \leqslant t$$

If $s > t$, the likelihood contribution of ith stratum in rth collapsing changes to

$$L_{ri} = 1 \quad \text{for } r \leqslant t \text{ and } r \geqslant s+1$$

$$= \frac{\exp(\beta)}{1 + \exp(\beta)} \quad \text{for } t+1 \leqslant r \leqslant s$$

Thus, the entire likelihood in (2) can be represented as

$$L = \prod_{s<t} \left[\frac{1}{1 + \exp(\beta)} \right]^{(t-s)n_{st}} \prod_{s>t} \left[\frac{1}{1 + \exp(\beta)} \right]^{(s-t)n_{st}}$$

where n_{st} is the number of matched sets with $Y_{i1} = s$ and $Y_{i2} = t$. Taking derivatives and solving the score equations, the estimator reduces to

$$\beta = \log \left(\frac{\sum_{s>t} (s-t) n_{st}}{\sum_{s<t} (t-s) n_{st}} \right) \tag{7}$$

Agresti and Lang [9] derive an asymptotic variance of this estimator based on the multinomial distribution of the counts n_{st}, which coincides with our sandwich estimator in (5) upon simplification, and is given by

$$\hat{\text{var}}(\hat{\beta}) = \frac{\sum_{i<j} (j-i)^2 n_{ij}}{[\sum_{i<j} (j-i) n_{ij}]^2} + \frac{\sum_{i>j} (i-j)^2 n_{ij}}{[\sum_{i>j} (i-j) n_{ij}]^2} \tag{8}$$

Liu and Agresti [12] prove that an alternative Mantel–Haenszel approach to matched pair data with ordinal response also provides the same estimator and estimated asymptotic variance. These equivalences in special cases are reassuring. However, compared with the other methods, our method provides the most flexible regression model, which can incorporate any number of discrete and continuous covariates.

Remark 3
Another application of this method is for matched case–control data with multiple ordinal disease states [10]. For a 1:1 matched data set, which has a very similar likelihood structure as Rasch's item response model, let, without loss of generality, Y_{i1} be the case observation, which could take values in $1, \ldots, K-1$ and $Y_{i2} \equiv 0$, denote the control observation. The conditional likelihood (2) simplifies to

$$L = \prod_{i=1}^{N} \left\{ \frac{\exp(\beta X_{i1})}{\sum_{j=1}^{2} \exp(\beta X_{ij})} \right\}^{Y_{i1}}$$

The pseudo-score function is derived as

$$U(\beta) = \sum_{i=1}^{N} Y_{i1} \left(X_{i1} - \frac{\sum_{j=1}^{2} X_{ij} \exp(\beta X_{ij})}{\sum_{j=1}^{2} \exp(\beta X_{ij})} \right)$$

The expression within parentheses is identical to the score function for standard CLR with the binary response. For ordinal response, these score functions are simply weighted by the ordinal value for Y_{i1}. The empirical sandwich variance estimate as in (5) becomes

$$\left(\sum_{i=1}^{N} Y_{i1}^2 \left[X_{i1} - \frac{\sum_{j=1}^{2} X_{ij} \exp(\beta X_{ij})}{\sum_{j=1}^{2} \exp(\beta X_{ij})} \right]^2 \right)$$

$$\times \left(\sum_{i=1}^{N} Y_{i1} \frac{[\sum_{j=1}^{2} X_{ij} \exp(\beta X_{ij})]^2 - [\sum_{j=1}^{2} \exp(\beta X_{ij})][\sum_{j=1}^{2} X_{ij}^2 \exp(\beta X_{ij})]}{[\sum_{j=1}^{2} \exp(\beta X_{ij})]^2} \right)^{-2}$$

Again, for binary X's, upon simplification of the above two expressions, one retrieves the Agresti–Lang and Liu–Agresti estimators of β.

4. EXAMPLES

In this section, we illustrate our methods for three data sets, representing common situations where stratified ordinal data may arise in practice. We consider several possible analysis strategies, along with the ACLR method. We consider a REM with $\alpha_i \overset{\text{iid}}{\sim} N(\mu_\alpha, \sigma_\alpha^2)$ akin to the methods proposed by Hedeker and Gibbons [15] and implemented by SAS procedure NLMIXED. We furnish the unconditional MLE assuming distinct fixed effects α_i for each stratum. The unconditional MLE is subject to a classic Neyman–Scott phenomenon [20] in the presence of fine stratification and often fails to converge. An analysis of the data ignoring the ordinal categories and running a usual CLR analysis with a simple dichotomy of zero *vs* non-zero responses is also presented (labeled as binary CLR).

Apart from these conditional models, one could alternatively fit a marginal response model to stratified ordinal data using a GEEs approach. As pointed out in [21], the GEE approach can account for the correlation between observations within a strata through robust variance estimation, but the point estimate is exactly the same as that obtained in an unmatched analysis, treating the observations independent. Indeed, using the `polr` option in R package VGAM, we noted that fitting simple constant intercept proportional odds logistic regression model to the data, ignoring stratification, produces identical point estimates as the GEE estimates with slightly different standard errors. We include the GEE estimates from the marginal model for our data analysis. However, we keep in mind that the GEE estimates reflect a population-averaged effect of the covariate, whereas the random effects and the ACLR method describe the change in the risk of outcome for a stratum when exposure status changes within the stratum. Thus, they are not directly comparable.

Example 1 (*Multi-center clinical trial on severity of Asthma*)
We use a data set provided by Merck Research Laboratories for a double-blind, parallel-group, preliminary clinical study conducted at 21 centers. Patients suffering from asthma were randomly assigned to three different treatments (2 mg active drug, 10 mg active drug, and placebo). At the end of the study, the doctors described the patients' change in condition using an ordinal scale

Table III. Data for Example 1: doctors' evaluations of patients suffering from asthma on a 1–4 scale in 21 clinical centers.

Center	Drug	Response				Center	Drug	Response			
		1	2	3	4			1	2	3	4
1	2 mg	0	1	2	1	2	2 mg	0	0	1	1
	10 mg	0	2	0	2		10 mg	0	0	1	2
	Placebo	0	0	0	4		Placebo	0	1	1	1
3	2 mg	0	0	2	2	4	2 mg	0	2	3	1
	10 mg	0	0	1	0		10 mg	0	2	2	2
	Placebo	0	1	4	1		Placebo	0	0	1	1
5	2 mg	0	1	1	0	6	2 mg	0	2	0	0
	10 mg	1	0	0	2		10 mg	0	1	0	0
	Placebo	1	0	0	2		Placebo	0	0	0	2
7	2 mg	0	0	2	2	8	2 mg	1	0	0	1
	10 mg	0	0	2	1		10 mg	0	3	0	0
	Placebo	0	0	2	1		Placebo	0	0	1	1
9	2 mg	0	0	2	1	10	2 mg	0	2	1	1
	10 mg	1	0	2	0		10 mg	0	1	0	0
	Placebo	0	0	1	0		Placebo	0	1	0	1
11	2 mg	0	1	1	0	12	2 mg	1	0	0	0
	10 mg	0	1	1	1		10 mg	1	1	0	0
	Placebo	0	0	0	3		Placebo	0	2	0	0
13	2 mg	0	1	1	2	14	2 mg	1	1	3	0
	10 mg	0	1	0	1		10 mg	1	0	1	0
	Placebo	0	0	0	5		Placebo	0	0	1	0
15	2 mg	0	0	3	2	16	2 mg	0	2	2	1
	10 mg	0	2	3	0		10 mg	2	1	2	0
	Placebo	0	1	2	1		Placebo	1	1	1	1
17	2 mg	0	1	0	0	18	2 mg	1	1	1	1
	10 mg	1	1	1	2		10 mg	0	1	0	0
	Placebo	0	1	1	3		Placebo	1	0	1	3
19	2 mg	0	3	0	0	20	2 mg	0	0	2	1
	10 mg	0	1	3	0		10 mg	0	1	0	3
	Placebo	0	1	1	1		Placebo	0	1	1	0
21	2 mg	0	1	1	2						
	10 mg	1	1	0	1						
	Placebo	0	1	0	0						

Note: Response is scaled from better (1) to worse (4).

from better to worse (1–4). Table III shows the results of the doctors' evaluations associated with the treatments and is presented as 21 separate 3×4 tables. Such a study might use many clinics because of the time it takes each clinic center to recruit patients. Therefore, the three-way table

might then have many strata but few observations per stratum, leading to a sparse data set with ordinal outcomes. For this particular data set, each clinic has from 5 to 14 subjects, leading to a total of 197 subjects in the study. We analyze the data set by different methods with treatment as the factor of interest but with several other covariates such as age, height, weight, and sex in the model. The results of this analysis are presented in Table IV. For this data set, although there are certain minor numerical differences, all the methods provide very similar inference regarding the treatment effect and suggest that higher drug doses tend to be negatively associated with worse symptoms (higher end of the ordinal scale). Using the ACLR method, the 2 mg dose has a log cumulative OR of -0.72, whereas the 10 mg dose has a log cumulative OR of -1.07. This suggests, for each center, that the estimated odds when the evaluation score for the 2 mg drug falls below any fixed level is $\exp(0.72) = 2.05$ times the estimated odds of placebo, whereas the estimated odds when the evaluation score for the 10 mg dose falls below any fixed level is $\exp(1.07) = 2.91$ times the estimated odds of placebo. Thus, the 10 mg dose has higher efficacy in providing relief to asthma patients.

The estimates of the unknown parameters corresponding to the random effects distribution are $\hat{\mu}_\alpha = 4.8$ and $\hat{\sigma}_\alpha = 0.43$, suggesting a reasonably small center to center variability, which explains why estimates from unmatched GEE analysis are similar to ACLR and REM. Note that a naive dichotomization of data, ignoring the ordinal scale, provides quite imprecise point estimates. This loss of efficiency due to ignoring the full ordinal scale is also noted in [22].

Example 2 (Item response data on asthma severity score)
In the same data set as described in Example 1, each of the 197 subjects had two measurements on asthma severity scores. In addition to the assessment by a doctor, each patient provided a self-evaluated asthma severity score. Treating each subject as a stratum in which two measurements are available, we would like to examine the response differences, that is, the differences between self-reported score and physician's evaluation. For simplicity, we ignore the presence of other covariates in the model and stratification effect due to centers, inclusion of which will lead to a more complex random effects structure with subject effects nested within center effects. To this end, we can fit the cumulative Rasch model in AL

$$\text{logit}[P(Y_{ij} \geqslant r | \mathbf{X}_{ij})] = \gamma_r + \alpha_i + \beta_j, \quad r = 1, \ldots, K-1, \ \ i = 1, \ldots, N \tag{9}$$

where $j = 1, 2$, and we use the constraint $\gamma_1 = \beta_1 \equiv 0$. There are 197 subject-specific nuisance parameters in this situation. In our analysis, the first response is physician reported score and the second response is individual reported score; thus, there is a single parameter of interest, $\beta_2 = \beta$, say, to be estimated from data. This cumulative odds ratio β reflects the odds of having scores on the higher end of the scale for self-reporting when compared with physician's evaluation. For this situation, in Table V, we note that the estimate of β is 1.56 by our method, which is close to the estimate provided by REM. This suggests that self-reported scores tend to be reported toward the higher (worse) end of the scale. When compared with doctor's evaluation, self-reported responses have an odds of $\exp(1.56) = 4.76$ of being on the higher end of the scale. However, note that the unmatched GEE analysis with constant stratum effect highly underestimates β, whereas the unconditional MLE fails to converge. The estimates of the parameters corresponding to the random effects distribution turn out to be $\hat{\mu}_\alpha = 3.74$ and $\hat{\sigma}_\alpha = 3.42$, reflecting that there is substantial heterogeneity across subjects in this data set. The analysis with a naive dichotomization of the scale using binary CLR again loses efficiency when compared with the ordinal models.

Table IV. Analysis of multi-center clinical data in Example 1. The response was a four-category ordinal variable of physician's evaluation of asthma severity, with higher score representing worse symptoms. The covariates in the model were treatments (three categories: placebo, 2 and 10 mg), age, sex, height, and weight.

Proportional odds model-based methods		β_1 (2 mg)	β_2 (10 mg)	β_3 (age)	β_4 (height)	β_5 (weight)	β_6 (sex)
ACLR	Estimate	−0.72	−1.07	−0.02	−0.08	0.01	0.03
	SE	0.41	0.33	0.01	0.08	0.00	0.45
	CI	(−1.52, 0.08)	(−1.71, −0.43)	(−0.04, 0.01)	(−0.24, 0.09)	(−0.00, 0.02)	(−0.85, 0.91)
(Bias corrected)	SE	0.43	0.34	0.01	0.09	0.01	0.47
	CI	(−1.56, 0.12)	(−1.74, −0.40)	(−0.04, 0.01)	(−0.25, 0.10)	(−0.00, 0.02)	(−0.89, 0.95)
GEE	Estimate	−0.74	−1.09	−0.01	−0.04	0.01	0.01
	SE	0.33	0.34	0.01	0.06	0.01	0.37
	CI	(−1.39, −0.10)	(−1.75, −0.42)	(−0.03, 0.01)	(−0.16, 0.08)	(−0.00, 0.02)	(−0.73, 0.72)
Unconditional	Estimate	−0.82	−1.12	−0.01	−0.05	0.01	−0.01
MLE	SE	0.35	0.36	0.01	0.07	0.01	0.40
	CI	(−1.51, −0.13)	(−1.82, −0.41)	(−0.04, 0.01)	(−0.18, 0.08)	(−0.01, 0.02)	(−0.80, 0.78)
REM	Estimate	−0.75	−1.09	−0.01	−0.04	0.01	−0.01
	SE	0.34	0.35	0.01	0.06	0.01	0.38
	CI	(−1.45, −0.05)	(−1.81, −0.36)	(−0.04, 0.02)	(−0.17, 0.09)	(−0.01, 0.02)	(−0.80, 0.78)
Naive dichotomization							
Binary CLR	Estimate	−0.76	−1.69	−0.09	−0.31	0.03	−1.51
	SE	0.89	0.89	0.04	0.16	0.02	0.97
	CI	(−2.51, 0.99)	(−3.42, 0.05)	(−0.16, −0.01)	(−0.62, 0.00)	(−0.01, 0.07)	(−3.42, 0.40)

Table V. Analysis of item-response data as described in Example 2, with two repeated responses per subject, one corresponding to self-evaluation score, and the other corresponding to physician's evaluation score on severity of asthma.

Proportional odds model-based methods		Item-response data β
ACLR	Estimate	1.56
	SE	0.30
	CI	(0.97, 2.15)
(Bias corrected)	SE	0.30
	CI	(0.97, 2.15)
GEE	Estimate	0.61
	SE	0.11
	CI	(0.40, 0.81)
REM	Estimate	1.54
	SE	0.32
	CI	(0.90, 2.18)
Naive dichotomization		
Binary CLR	Estimate	1.65
	SE	0.49
	CI	(0.69, 2.61)

The coefficient β represents the relative cumulative log-odds ratio of reporting higher values of severity scores for an individual evaluation, when compared with physician's evaluation. Unconditional MLE did not converge for this data set.

Example 3 (*Matched case–control data with multiple ordered disease states*)
We consider a matched case–control data set coming from a low birth-weight study conducted by the Baystate Medical Center in Springfield, Massachusetts. The data set is discussed in [23] and is used as an illustrative example in Chapter 7. Low birth weight, defined as birth weight less than 2500 g, is a cause of concern for a newborn as infant mortality and birth defect rates are very high for low birth weight babies. The data were matched according to the age of the mother. A woman's behavior during pregnancy (smoking habits, diet, and prenatal care) can greatly alter the chances of having a low birth-weight baby. The goal of the study was to determine whether these variables were 'risk factors' in the clinical population served by Baystate Medical Center. Using the actual birth-weight observations, we divided the cases, namely, the low birth-weight babies into two categories, *very* low (weighing less than 2000 g) and low (weighing between 2000 and 2500 g) and tried to assess the impact of smoking habits of mother (smoker *vs* non-smoker) on the chance of falling in the two low birth-weight categories. We consider two data sets based on this low birth-weight study with two different matching ratios: a 1:1 matched data set with 56 matched sets and a 1:3 matched data set with 29 matched sets. Analysis of the 1:1 and 1:3 data sets are presented in Table VI, where the parameter β represents the smoking effect for the model. In this situation, we note that for the 1:1 data set, the estimates from our method and REM differ substantially, with REM estimates being identical to the ones obtained by unmatched analysis of data. The parameters of the random effects distribution were estimated as $\hat{\mu}_\alpha = -0.35$ and $\hat{\sigma}_\alpha = 4.9 \times 10^{-6}$. The point estimates obtained by our method and binary CLR are higher than REM estimates. For the 1:3 data set, both methods provide similar inference. This difference between REM and ACLR

Table VI. Analysis of low birth-weight study, one is 1:1 matched data set, and the other is 1:3 matched data set as described in Example 3.

Proportional odds model-based methods		β	
		1:1 Matched	1:3 Matched
ACLR	Estimate	0.97	0.87
	SE	0.44	0.43
	CI	(0.11, 1.83)	(0.02, 1.72)
(Bias corrected)	SE	0.43	0.43
	CI	(0.12, 1.82)	(0.04, 1.71)
GEE	Estimate	0.82	0.90
	SE	0.38	0.42
	CI	(0.07, 1.57)	(0.07, 1.72)
Unconditional MLE	Estimate	1.99	1.27
	SE	0.57	0.51
	CI	(0.88, 3.11)	(0.26, 2.27)
REM	Estimate	0.82	0.90
	SE	0.37	0.43
	CI	(0.10, 1.54)	(0.01, 1.79)
Naive dichotomization			
Binary CLR	Estimate	1.01	0.96
	SE	0.41	0.44
	CI	(0.20, 1.82)	(0.10, 1.82)

The estimates correspond to the smoking effect of mother on low birth-weight babies.

from 1:1 to 1:3 matching is most likely due to the degree of stratification, as will be noted in our simulation study. The effect of smoking is more pronounced in our method although all methods suggest elevated odds of having low birth-weight baby for a smoker mother.

Figure 2 summarizes the 95 per cent Wald confidence interval (CI) of the regression parameter β for all three examples. Note that the width of CI using GEE and REM is often slightly less than ACLR as they use all observations, not only discordant observations in each stratum and thus gain precision. The naive binary CLR produces much wider CI compared with the ordinal methods.

5. SIMULATION STUDY

In the above three data analysis examples, we saw similarities and differences between our method and other approaches. Since in a real data situation, one is not certain about the truth regarding parameter values, we designed a simulation study that explores, under different sample sizes, the performance of our method over an array of distributions for the stratum effects. As pointed out by a referee, the bias due to misspecification of the proportional odds structure should also be evaluated since it is a restrictive modeling assumption. In a given real example, the validity of the proportional odds assumption should first be verified by available diagnostic tools and goodness-of-fit statistics [23, 24] before using the proposed methods. However, in the present simulation study, we assume that the proportional odds model holds and only evaluate the variation across distributions governing the nuisance parameters. In particular, we consider the following four
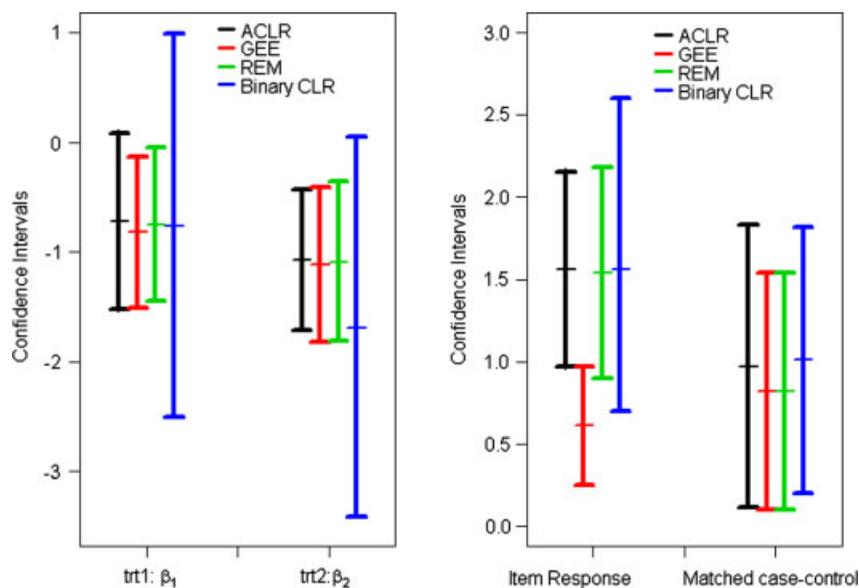
Figure 2. Results for the real data analysis: the left panel shows point estimates and 95 per cent Wald CI corresponding to dose effects in Example 1 and the right panel shows the estimate and 95 per cent Wald CI of the single cumulative log odds-ratio parameter in Examples 2 and 3 on item response and 1:1 matched case–control data set, respectively.

settings with (i) $\alpha_i \equiv \alpha_0 = -3$, (ii) $\alpha_i \overset{\text{iid}}{\sim} N(-3, 2^2)$, (iii) $\alpha_i \overset{\text{iid}}{\sim} 0.4 N(-3.5, 0.4^2) + 0.6 N(-2.5, 0.4^2)$, and (iv) α_i following a mixture of discrete and continuous distribution, assuming a constant value of -3.5 with probability 0.4 and following a N$(-2.5, 0.4^2)$ distribution with probability 0.6. The true values of the category-specific threshold parameters γ_1, γ_2, and γ_3 were chosen as 2.5, 1.4, and 0.5, respectively, in each simulation setting. We consider two experimental scenarios, reflecting different study designs. In Table VII we consider a moderately stratified data situation with five observations in each strata, responses having four ordinal values and with two covariates, one binary ($p = 0.4$) and the other continuous, from a N$(1, 2)$ distribution. The corresponding regression coefficients are set at $\beta_1 = -0.5$ and $\beta_2 = 1.0$, respectively. In Table VIII, we generate highly stratified item response data with two repeated responses per subject (each response having four ordinal categories) with the cumulative log odds ratio on the 'difficulty' scale being $\beta = 0.5$.

Tables VII and VIII present the results with bias, MSE, and the coverage probability of 95 per cent Wald-type CIs corresponding to the different estimates. Note that in the constant stratum effect situation (i) all ordinal methods perform comparably, with REM and GEE being more precise than ACLR. However, in the varying stratum effect situation, with a normal distribution on the stratum effects as in (ii), our method and REM stand superior to GEE in terms of bias and coverage probability. This is expected as GEE essentially provides an unmatched analysis of the data, which attenuates the effect of the covariate toward the null value. REM appears to have a slight edge over ours in terms of MSE for smaller number of strata. In spite of small MSE, the GEE method accrues bias in simulation setting (ii) and has poor coverage probability, which is again as expected. With a non-normal distribution as generated in (iii) and (iv), our method has less bias than REM for

Table VII. Table presenting MSE, bias (parenthesis), and coverage probability of 95 per cent Wald-type CI based on 1000 simulations under four settings for distribution of α_i as mentioned in Section 5 with moderately stratified data set containing five observations in each stratum. The response has four categories with two covariates, a binary covariate X_1 and a continuous covariate X_2. True value of $\beta_1 = -0.5$ and $\beta_2 = 1$. The number of strata is varied from 20 to 100.

Distribution for α_i	Method	β_1			β_2		
		20 Strata 5 observations	50 Strata 5 observations	100 Strata 5 observations	20 Strata 5 observations	50 Strata 5 observations	100 Strata 5 observations
(i) Constant	ACLR	0.28 (−0.01) 0.92	0.10 (−0.01) 0.93	0.05 (0.00) 0.94	0.04 (0.06) 0.92	0.02 (0.04) 0.94	0.01 (0.01) 0.95
	GEE	0.18 (−0.02) 0.95	0.07 (−0.02) 0.94	0.04 (0.00) 0.95	0.03 (0.04) 0.94	0.02 (0.01) 0.95	0.01 (0.01) 0.95
	REM	0.19 (−0.04) 0.96	0.07 (−0.03) 0.95	0.04 (−0.01) 0.95	0.04 (0.06) 0.95	0.01 (0.04) 0.94	0.01 (0.02) 0.95
	Binary CLR	0.69 (−0.06) 0.96	0.19 (−0.02) 0.96	0.09 (0.00) 0.96	0.21 (0.16) 0.96	0.04 (0.06) 0.97	0.02 (0.02) 0.95
(ii) Normal	ACLR	0.32 (−0.02) 0.92	0.13 (−0.01) 0.94	0.07 (0.00) 0.94	0.09 (0.09) 0.94	0.02 (0.03) 0.94	0.01 (0.01) 0.94
	GEE	0.20 (0.18) 0.94	0.10 (0.18) 0.88	0.06 (0.18) 0.82	0.14 (−0.35) 0.28	0.13 (−0.35) 0.03	0.14 (−0.37) 0.00
	REM	0.29 (0.00) 0.94	0.11 (−0.01) 0.94	0.05 (−0.01) 0.95	0.05 (0.05) 0.95	0.02 (0.02) 0.96	0.01 (0.01) 0.96
	Binary CLR	0.97 (−0.05) 0.96	0.27 (−0.03) 0.96	0.12 (0.00) 0.95	0.33 (0.14) 0.95	0.05 (0.06) 0.97	0.02 (0.02) 0.97
(iii) Mixture of normals	ACLR	0.31 (−0.02) 0.93	0.11 (−0.02) 0.95	0.05 (0.00) 0.95	0.06 (0.03) 0.92	0.02 (0.02) 0.95	0.01 (0.01) 0.95
	GEE	0.20 (0.09) 0.93	0.07 (0.07) 0.93	0.04 (0.09) 0.93	0.04 (−0.14) 0.77	0.03 (−0.16) 0.51	0.03 (−0.18) 0.23
	REM	0.25 (0.05) 0.95	0.09 (0.04) 0.94	0.04 (0.02) 0.95	0.04 (0.04) 0.96	0.01 (0.03) 0.95	0.01 (0.02) 0.94
	Binary CLR	0.71 (−0.06) 0.97	0.22 (−0.02) 0.96	0.10 (0.00) 0.95	0.18 (0.15) 0.97	0.05 (0.06) 0.95	0.02 (0.02) 0.96
	ACLR	0.30 (−0.03)	0.09 (−0.01)	0.05 (−0.01)	0.06 (0.05)	0.02 (0.02)	0.01 (0.01)

Table VII. *Continued.*

Distribution for α_i	Method	20 Strata 5 observations	50 Strata 5 observations	100 Strata 5 observations	20 Strata 5 observations	50 Strata 5 observations	100 Strata 5 observations
(iv) Mixture of normal and a constant	GEE	0.93 0.18 (0.08) 0.95	0.95 0.07 (0.07) 0.95	0.95 0.04 (0.08) 0.93	0.94 0.04 (−0.12) 0.82	0.94 0.03 (−0.15) 0.59	0.94 0.03 (−0.16) 0.30
	REM	0.22 (−0.06) 0.96	0.08 (−0.03) 0.96	0.04 (−0.02) 0.95	0.04 (0.07) 0.96	0.01 (0.03) 0.94	0.01 (0.01) 0.94
	Binary CLR	0.69 (−0.05) 0.97	0.19 (−0.06) 0.96	0.10 (−0.02) 0.95	0.29 (0.17) 0.98	0.04 (0.05) 0.96	0.02 (0.02) 0.93

Table VIII. Table presenting MSE, bias (parenthesis), and coverage probability of 95 per cent Wald-type CI based on 1000 simulations under four settings for distribution of α_i as mentioned in Section 5 with highly stratified item-response data set containing two item-specific responses per subject.

Distribution for α_i	Method	MSE (bias) coverage	50 Strata with 2 observations	100 Strata with 2 observations	200 Strata with 2 observations
	ACLR		0.19 (0.04) 0.94	0.08 (0.00) 0.96	0.04 (0.01) 0.96
(i) Constant	GEE		0.16 (0.03) 0.95	0.07 (0.00) 0.95	0.04 (0.01) 0.96
	REM		0.28 (0.06) 0.95	0.10 (0.02) 0.96	0.05 (0.02) 0.96
	Binary CLR		0.60 (0.05) 0.98	0.27 (0.01) 0.96	0.13 (0.03) 0.96
	ACLR		0.24 (0.01) 0.94	0.12 (0.01) 0.95	0.05 (0.00) 0.96
(ii) Normal	GEE		0.13 (-0.19) 0.86	0.08 (-0.19) 0.85	0.06 (-0.19) 0.85
	REM		0.26 (0.00) 0.96	0.11 (0.01) 0.95	0.05 (0.00) 0.95
	Binary CLR		0.55 (0.03) 0.97	0.25 (0.02) 0.95	0.11 (0.00) 0.95
	ACLR		0.19 (0.02) 0.95	0.09 (0.00) 0.95	0.04 (0.00) 0.97
(iii) Mixture of normals	GEE		0.18 (0.12) 0.86	0.10 (0.11) 0.83	0.07 (0.10) 0.84
	REM		0.30 (0.06) 0.97	0.13 (0.06) 0.97	0.05 (0.05) 0.98
	Binary CLR		0.51 (0.03) 0.98	0.27 (0.01) 0.96	0.12 (0.02) 0.96
	ACLR		0.17 (0.05) 0.96	0.09 (0.02) 0.95	0.05 (0.01) 0.95
(iv) Mixture of a normal and constant	GEE		0.13 (0.11) 0.94	0.08 (0.12) 0.92	0.04 (0.09) 0.93
	REM		0.28 (0.10) 0.94	0.13 (0.11) 0.96	0.07 (0.11) 0.96
	Binary CLR		0.61 (0.06) 0.97	0.59 (-0.02) 0.98	0.29 (0.03) 0.97

The response has four categories. True value of the log cumulative odds-ratio $\beta = 0.5$. The number of strata is varied from 50 to 200.

the highly stratified item response data set in Table VIII, and the bias in our method goes away with the increase in the number of strata, whereas the bias due to model misspecification on $\{\alpha_i\}$ remains in REM estimates even for relatively large number of strata. With more observations per stratum, the REM estimates become less biased under the non-normal distributions in (iii) and (iv) as noted in Table VII. To summarize, our proposed ACLR method appears to be more robust to variations in stratum-specific distributions in highly sparse data situations. This can be seen for the

item response data especially with a large number of strata, leading to many nuisance parameters. The proposed ACLR method also has reasonably good MSE properties for sparse data situations.

One can note the inadequacy of the GEE model or unmatched analysis for estimating β in the simulation settings with α_i truly varying. The lack of efficiency due to naive dichotomization in binary CLR model is also evident when compared with the ordinal models. The GEE model and REM use all the observations in each stratum, not only the discordant observations as in ACLR. As a result, they often have higher precision. But when striking a compromise between robustness and precision, the ACLR approach appears to be a reasonable choice under varying stratification heterogeneity.

Remark 4
In simulation studies, the results of which are not included in the current text, we noted two additional advantages of the ACLR method over the REM approach, which are worth mentioning. If the intercept parameters α_{ri} in equation (1) do not have the additive structure, $\alpha_{ri} = \alpha_r + \gamma_i$, but have a multiplicative (or some other) structure, the REM method incurs bias, the ACLR methods is not sensitive to this structure as the α_{ri} parameters do not appear in the conditional likelihood (2). In addition, if the stratum effects are generated from a distribution whose parameters depend on the regression covariate X, the REM approach suffers appreciably compared with ACLR in terms of bias.

6. DISCUSSION

In this paper, we present a regression method for estimating the parameters in a highly stratified proportional odds model. The direct MLEs for this model are not consistent in sparse-strata situations, that is, when the number of strata increases in direct proportion to sample size. The proposed estimator is consistent for the two standard types of asymptotics: (a) when sample size within each stratum increases and the number of strata is fixed (the large-strata case) and (b) when the number of strata increases proportional to the overall sample size (the sparse-strata case). We derive a sandwich estimate of the variance of the estimator and propose Wald-type tests and CIs based on this approximation. We show the equivalence of our approach with a Mantel–Haenszel estimator proposed by Liu and Agresti [12] and an alternative estimator proposed by Agresti and Lang [9] for generalized Rasch's item-response model. The above approaches are limited in their applications as they can accommodate only a modest number of categorical covariates while our method provides the flexibility of a general regression model. Compared with a random effects approach, our method is more robust as it does not assume a specific form of the stochastic distribution for the stratification parameters and is computationally much simpler. The proposed method seems particularly suitable for pair-matched data, representing the highest degree of stratification. Our simulation study also indicates the loss of efficiency by simple dichotomization of response instead of using the full ordinal scale. One could possibly consider a different model structure for the ordinal response, such as an adjacent category logit model or a continuation-ratio logit model [25], that we have not considered. The major difference in those cases is that the optimal solution, the conditional MLE, is available in both of those situations.

Admittedly, the proposed estimator is not fully efficient. For the large-strata situation as in (a) above, Liu and Agresti show that the loss of efficiency is surprisingly less in using this type of naive estimators when compared with the full ML estimators. For large-strata case, Clayton [26]

derived a weighted estimator of β based on the separate collapsings. The weights were chosen so as to minimize the variance of the estimator when $\beta = 0$. It is unclear how this method could be adapted to derive valid estimates under sparse data situation as the estimated weights are highly unstable in such a case. Furthermore, the choice of weights is not critically important to the efficiency for these types of estimators [17]. We also noted the same phenomena when we tried to use weights suggested from the Bayesian model averaging perspective [27], where estimates from each collapsing were weighted by the posterior probability of the collapsed model as approximated by the Bayes information criterion. This weighting hardly made difference in terms of mean-squared error performance of the estimator. However, from a theoretical perspective, it is indeed an interesting open question to obtain the optimum weights. Techniques such as the ones developed in [28] could potentially lead to the semiparametric efficient score function.

We have proposed a robust estimation strategy, which is relatively easy to implement, works well for sparse data and is not sensitive to the choice of the random distribution, which generates the nuisance parameters α_{ri} and utilizes the full ordinal scale. We reiterate that the proposed method relies on the proportional odds structure, which is a strong modeling assumption in itself, and thus is subject to model misspecification. In a real example, the validity of the proportional odds assumption should be carefully assessed. R code for implementing the above method and SAS code for implementing REM and GEE are available at http://www.sph.umich.edu/bhramar/public_html/research.

REFERENCES

1. Snell EJ. A scaling procedure for ordered categorical data. *Biometrics* 1964; **20**:592–607.
2. Bock RD, Jones LV. *The Measurement and Prediction of Judgement and Choice*. Holden-Day: San Francisco, CA, 1968.
3. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society*, *Series B* 1980; **42**:109–142.
4. Rasch G. On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, vol. 4, University of California Press, 1961; 321–334.
5. Godambe VP. Conditional likelihood and unconditional optimum estimating equations. *Biometrika* 1976; **63**: 277–284.
6. Rathouz PJ. Likelihood methods for missing covariate data in highly stratified studies. *Journal of the Royal Statistical Society*, *Series B* 2003; **65**:711–723.
7. McCullagh P. A logistic model for paired comparisons with ordered categorical data. *Biometrika* 1977; **64**: 449–453.
8. McCullagh P. On the elimination of nuisance parameters in the proportional odds model. *Journal of the Royal Statistical Society*, *Series B* 1984; **46**:250–256.
9. Agresti A, Lang JB. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* 1993; **80**:527–534.
10. Mukherjee B, Liu I, Sinha S. Analysis of matched case–control data with multiple ordered disease states: possible choices and comparisons. *Statistics in Medicine* 2007; **26**:3240–3257.
11. Liu I, Wang D. Diagnostics for stratified clinical trials in proportional odds models. *Communications in Statistics*: *Theory and Methods* 2007; **36**:211–220.
12. Liu I-M, Agresti A. Mantel–Haenszel-type inference for cumulative odds ratios. *Biometrics* 1996; **52**:1222–1234.

13. Hartzel J, Liu IM, Agresti A. Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials. *Computational Statistics and Data Analysis* 2001; **35**:429–449.
14. Neuhaus JM, Kalbfleisch JD, Hauck WW. Conditions for consistent estimation in mixed models for binary matched-pairs data. *Canadian Journal of Statistics* 1994; **22**:139–148.
15. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
16. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
17. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989; 274.
18. Fay MP, Graubard BI, Freedman LS, Midthune DN. Conditional logistic regression with sandwich estimators: application to meta-analysis. *Biometrics* 1998; **54**:195–208.
19. Fay MP, Graubard BI. Small sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001; **57**:1198–1206.
20. Neyman J, Scott EL. Consistent estimates based on partially consistent observations. *Econometrika* 1948; **16**:1–22.
21. Hu FB, Goldberg J, Hedeker D, Henderson WG. Modelling ordinal responses from co-twin control studies. *Statistics in Medicine* 1998; **17**:957–970.
22. Liu I, Agresti A. The analysis of ordered categorical data: an overview and a survey of recent developments. *Test* 2005; **14**:1–73.
23. Hosmer DW, Lemeshow S. *Applied Logistic Regression* (2nd edn). Wiley: New York, 2000.
24. Lipsitz SR, Fitzmaurice GM, Molenberghs G. Goodness-of-fit tests for ordinal response regression models. *Applied Statistics* 1996; **45**:175–190.
25. Thompson WA. On the treatment of grouped observations in life studies. *Biometrics* 1977; **33**:463–470.
26. Clayton DG. Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika* 1974; **61**:525–531.
27. Madigan D, Raftery A. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 1994; **89**:1535–1546.
28. Tsiatis AA, Ma YY. Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* 2004; 835–848.