Computational Statistics and Data Analysis 88 (2015) 173–186

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Multilevel cumulative logistic regression model with random effects: Application to British social attitudes panel survey data

Moon-tong Chan [a], Dalei Yu [b,c], Kelvin K.W. Yau [a,*]

[a] Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Hong Kong

[b] Statistics and Mathematics College, Yunnan University of Finance and Economics, Kunming 650221, China

[c] Yunnan Tongchuang Scientific Computing & Data Mining Center, Kunming 650221, China

ARTICLE INFO

*Article history:*
Received 21 January 2014
Received in revised form 25 February 2015
Accepted 27 February 2015
Available online 9 March 2015

*Keywords:*
Generalized linear mixed model
Multilevel model
Ordinal response
Random effect

ABSTRACT

A multilevel model for ordinal data in generalized linear mixed models (GLMM) framework is developed to account for the inherent dependencies among observations within clusters. Motivated by a data set from the British Social Attitudes Panel Survey (BSAPS), the random district effects and respondent effects are incorporated into the linear predictor to accommodate the nested clusterings. The fixed (random) effects are estimated (predicted) by maximizing the penalized quasi likelihood (PQL) function, whereas the variance component parameters are obtained via the restricted maximum likelihood (REML) estimation method. The model is employed to analyze the BSAPS data. Simulation studies are conducted to assess the performance of estimators.



## 1. Introduction

Ordinal responses are very common in different areas of scientific research (e.g. behavioral research, medical research and social attitude studies). However, due to the complexity of the model, corresponding statistical inference about the ordinal responses is in general difficult, especially in the situation where the data have complex correlation structure. A class of three-level mixed effects models for ordinal data was proposed by Raman and Hedeker (2005) and both proportional and non-proportional odds models were considered. Furthermore, for the ordinal data with the clustered structure, Liu and Hedeker (2006) proposed a mixed-effects item response theory model that allows for three-level multivariate ordinal outcomes and accommodates multiple random subject effects. For the model estimation, both Raman and Hedeker (2005) and Liu and Hedeker (2006) adopted Gauss–Hermite quadrature to integrate out the random effects numerically to obtain the marginal likelihood function. In the context of complex multilevel structures, Fielding and Yang (2005) developed generalized linear mixed models (GLMMs) for ordered responses using quasi-likelihood with up to the second-order terms in the Taylor expansion. Chakraborty and Das (2008), in a pharmacokinetic study, considered a latent nonlinear mixed effects model to summarize and analyze multivariate ordinal data. They employed Monte Carlo EM-based methods for parameter estimation and inference. Essentially, all of these methods fall into the domain of marginal models (Lee and Nelder, 2004). However, in many situations where the random cluster effects themselves are of research interest, it is useful to obtain the prediction of random effects under the framework of conditional likelihood inference (Jiang et al., 2001; Lee and Nelder, 2004).

* Corresponding author.
*E-mail address:* mskyau@cityu.edu.hk (K.K.W. Yau).

http://dx.doi.org/10.1016/j.csda.2015.02.018

The current study is motivated by the data collected from the British Social Attitudes Panel Survey (BSAPS). The survey consisted of data collection in four consecutive years (1983–1986). In each year, 54 polling districts were involved and 264 adults living at these addresses were surveyed. Therefore, respondents who reside in the same polling district are exposed to the same set of unobservable district effects. Also, the data collected from the same respondent are deemed to be correlated (detailed discussion is provided in Section 3). In addition to the assessment of the fixed effects, the random district effects as well as the random respondent effects are also of research interest. Accordingly, instead of integrating out the random effects as in the marginal models, we adopt the conditional likelihood approach (McGilchrist, 1994; Lee and Nelder, 1996, 2004) to analyze the data, where the estimation/prediction of fixed/random effects is obtained in principle by penalized quasi likelihood (PQL) estimation method (Breslow and Clayton, 1993). The PQL estimator carries over the spirit of best linear unbiased prediction (BLUP) into the non-normal framework (McGilchrist, 1994; McGilchrist and Yau, 1995) in a computationally attractive way. Moreover, its theoretical properties have been investigated in different perspectives, e.g. by Breslow and Clayton (1993), McGilchrist (1994), Lee and Nelder (1996), Yau and Kuk (2002), and Yu and Yau (2012). In addition, Jiang et al. (2001) investigated the consistency of the PQL estimators under the framework of maximum posterior estimation and proved that the PQL is asymptotically accurate under certain regularity conditions.

This paper is organized as follows. A multilevel cumulative logistic regression model with random effects is presented in Section 2. The proposed method is applied to analyze the BSAPS data in Section 3. Simulation studies are conducted to assess the performance of the estimators in Section 4. The last section provides some concluding discussions and model extensions.

## 2. Multilevel cumulative logistic regression model with random effects

### 2.1. Model formulation

Let $y_{ist}$ $(i = 1, 2, \ldots, m;\ s = 1, 2, \ldots, n_i;\ t = 1, 2, \ldots, T_{is})$ represent the ordinal response for the $t$-th observation of the $s$-th respondent in the $i$-th district, where $m$ is the number of districts, $n_i$ is the number of respondents within the $i$-th district, $T_{is}$ is the number of repeated observations for the $s$-th respondent in the $i$-th district, the total number of respondents being $n = \sum_{i=1}^{m} n_i$, and the total number of observations being $N = \sum_{i=1}^{m}\sum_{s=1}^{n_i} T_{is}$.

Denote $x_{ist}$ as the vector of covariates corresponding to the $t$-th observation for the $s$-th respondent in the $i$-th district. Based on the GLMM framework, two sets of random effects are considered: $a = (a_1, a_2, \ldots, a_m)^{\mathrm{T}}$, which represents the vector of random district effects, and $a_i$ is the $i$-th district effect; $b = \left(b_1^{\mathrm{T}}, b_2^{\mathrm{T}}, \ldots, b_m^{\mathrm{T}}\right)^{\mathrm{T}}$ is the vector of random respondent effects, where $b_i = \left(b_{i1}, b_{i2}, \ldots, b_{in_i}\right)^{\mathrm{T}}$, and $b_{is}$ is the $s$-th respondent effect in the $i$-th district. In addition, $a$ and $b$ are assumed to have a probability density function $\pi(a, b \mid \varphi)$, where $\varphi$ is the vector of variance component parameters. Conditional on $a$ and $b$, $y_{ist}$ are independent and the model is defined by

$$\begin{aligned}\operatorname{logit}P\left(y_{ist} \le j|\xi, a, b\right) &= \log\left[\frac{\gamma_j\left(x_{ist}|\xi, a, b\right)}{1-\gamma_j\left(x_{ist}|\xi, a, b\right)}\right], \quad j = 1, 2, \ldots, k,\\ &= \theta_j + \eta_{ist},\\ &= \theta_j + x_{ist}^{\mathrm{T}}\beta + a_i + b_{is}, \end{aligned} \tag{1}$$

where $\gamma_j\left(x_{ist}|\xi, a, b\right) = P\left(y_{ist} \le j|\xi, a, b\right),\ \xi = \left(\theta^{\mathrm{T}}, \beta^{\mathrm{T}}\right)^{\mathrm{T}},\ \theta_j$ is the intercept of the $j$-th cumulative logit, $\beta$ is the $p$-dimensional vector of fixed effects, and $k$ is the number of ordinal categories. The intercepts satisfy $\theta_1 < \theta_2 < \cdots < \theta_{k-1}$ such that $\gamma_j\left(x_{ist}|\xi, a, b\right) - \gamma_{j-1}\left(x_{ist}|\xi, a, b\right)$ remains positive, $\gamma_0\left(x_{ist}|\xi, a, b\right) = 0$ and $\gamma_k\left(x_{ist}|\xi, a, b\right) = 1$.

The penalized log-likelihood function is $l = l_1 + l_2$, with

$$l_1 = \sum_{i=1}^{m}\sum_{s=1}^{n_i}\sum_{t=1}^{T_{is}}\sum_{j=1}^{k}\delta_{ist,j}\log\left[\gamma_j\left(x_{ist}|\xi, a, b\right) - \gamma_{j-1}\left(x_{ist}|\xi, a, b\right)\right], \tag{2}$$

where $\delta_{ist,j} = 1$ if $y_{ist} = j$ and 0 otherwise, and $l_2 = \log\pi(a, b \mid \varphi)$ is the penalty term. Profiling on $\varphi$, the BLUP-type estimators/predictors of fixed/random effects are obtained by maximizing $l$ and the obtained estimators/predictors can be viewed as a natural extension of BLUP (Jiang, 2007). In fact, if $y$ is normally distributed (with conditional mean $\eta_{ist}$) and $\pi(a, b \mid \varphi)$ is the probability density function of normally distributed random effects vectors $a$, $b$, the resulting estimation method will reduce to BLUP. Detailed derivations of the BLUP estimation method were given in Robinson (1991). Further discussions and its extensions can be found in Jiang (2007). In principle, there are different choices for $\pi(a, b \mid \varphi)$, we consider two typical examples.

**Example 1.** If the random effects, $a$ and $b$, are assumed to follow $N\left(0, \sigma_1^2 I_m\right)$ and $N\left(0, \sigma_2^2 I_n\right)$, respectively, then,

$$l_2 = -\frac{1}{2}\left\{\left[m\log\left(2\pi\sigma_1^2\right) + \sigma_1^{-2}a^{\mathrm{T}}a\right] + \left[n\log\left(2\pi\sigma_2^2\right) + \sigma_2^{-2}b^{\mathrm{T}}b\right]\right\}, \tag{3}$$

and in this scenario, $\varphi = (\sigma_1^2, \sigma_2^2)^{\mathrm{T}}$.