Computational Statistics and Data Analysis 92 (2015) 126–133

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Detecting misspecification in the random-effects structure of cumulative logit models

Kuo-Chin Lin [a], Yi-Ju Chen [b,*]

[a] Department of Business Administration, Tainan University of Technology, Taiwan
[b] Department of Statistics, Tamkang University, Taiwan

ARTICLE INFO

Article history:
Received 5 September 2014
Received in revised form 22 June 2015
Accepted 1 July 2015
Available online 10 July 2015

Keywords:
Generalized linear mixed models
Longitudinal ordinal data
Misspecification
Reconstructed data

ABSTRACT

A common approach to analyzing longitudinal ordinal data is to apply generalized linear mixed models (GLMMs). The efficiency and validity of inference for parameters are affected by the random-effects distribution in GLMMs. A proposed test is developed based on the observed data and a reconstructed data set induced from the observed data for diagnosing the random-effects misspecification in cumulative logit models for longitudinal ordinal data, extending the idea presented by Huang (2009) for longitudinal binary data. The proposed test statistic has the quadratic form of the difference of maximum likelihood estimators between the observed data and the reconstructed data, and it follows a limiting chi-squared distribution when the model is correctly specified. The simulation studies are conducted to assess the performance of the proposed test, and a clinical trial example demonstrates the application of the proposed test.

1. Introduction

Recently there has been increasing use of the generalized linear mixed models (GLMMs) for longitudinal data in a variety of fields. The generalized linear models are a class of fixed-effects regression models for the different types of response variables including continuous, dichotomous and counts, and GLMMs are a generalization of the generalized linear models by incorporating random effects into the linear predictor. The random effects represent the heterogeneity between subjects, and are often assumed to be normally distributed. The estimation of model parameters in GLMMs containing the coefficients of fixed effects and the covariance matrix of random effects is an essential issue, and diagnosis of the random-effects misspecification in GLMMs is an obvious concern. Verbeke and Lesaffre (1997) studied the effect of misspecification of the random-effects distribution for continuous longitudinal data. Heagerty and Zeger (2000) pointed out that the estimators of regression coefficients in random-effects models are more sensitive to random-effects assumption than those in marginal models. Agresti et al. (2004) addressed the impact of random-effects misspecification in GLMMs on the possibly considerable loss of efficiency in the maximum likelihood estimates of the fixed effects. Litière et al. (2007, 2008) conducted simulation studies to evaluate the influence of the random-effects misspecification on type I and II errors of tests for the mean structure in logistic random-intercept models. The sensitivity of model parameter estimates arising from the random-effects misspecification becomes an important issue. Therefore, some diagnostic methods for detection of misspecification and avoidance of harmful effects from misspecification are developed.

* Corresponding author.
E-mail address: ychen@stat.tku.edu.tw (Y.-J. Chen).

http://dx.doi.org/10.1016/j.csda.2015.07.002
0167-9473/

Over the past few years, there is a considerable literature on the diagnosis of the random-effects misspecification in GLMMs. Tchetgen and Coull (2006) introduced a diagnostic test based on the difference between marginal maximum likelihood and conditional maximum likelihood estimators for the random-intercept logistic regression model for clustered binary responses. Waagepetersen (2006) assessed the distributional assumption for the random effects in Poisson-log normal and binomial-logit normal GLMMs utilizing a conditional simulation of the random effects, conditional on the observations for longitudinal data. Alonso et al. (2008) developed three diagnostic tools based on the eigenvalues of the variance–covariance matrix of the fixed-effects parameters estimates for detecting the random-effects structure of GLMMs with binary responses. Huang (2009) provided a diagnostic method for detecting the random-effects misspecification in GLMMs for clustered binary responses using the comparison of the estimation of parameters based on the observed data and a reconstructed data set induced from the observed data. Abad et al. (2010) proposed two diagnostic approaches for detection of misspecification in GLMMs for longitudinal binary data through two representations of the model information matrix. So far there has been limited study regarding the detection of random-effects misspecification in GLMMs for longitudinal ordinal data, leading to our motivation to develop a diagnostic test for misspecification in the random-effects structure of cumulative logit models with proportional odds assumption described by McCullagh (1980), which is a commonly used approach for the analysis of longitudinal ordinal data.

The main challenge for assessing the random-effects misspecification in GLMMs is that the random effects are not observed; namely, data realization or surrogate observation is unavailable for the random effects. As the observed data solely are unable to be used for detecting the misspecification of the random-effects distribution, a reconstructed data set induced from the observed data is created to accompany the observed data to form the diagnostic test statistic. The observed and reconstructed data for GLMMs with cumulative logit are introduced in Section 2. The sampling distribution of the proposed test statistic is derived in Section 3. The performance of the proposed test in terms of relative bias, empirical size and power is evaluated by simulations, and the practical implementation of the proposed test is illustrated by a real data set in Section 4. Finally, a summary discussion and some concluding remarks are given.

2. GLMMs with cumulative logit

Based on the difference of parameter estimates between the observed data and a reconstructed data, Huang (2009) proposed a diagnostic approach for the random-effects misspecification in GLMMs for clustered binary responses. Although we extend the diagnostic method proposed by Huang (2009) to the ordinal responses, we have to deal with the complexity of ordinal responses in the theoretical result and computational simulation. Three types of logit functions used for the analysis of ordinal data are cumulative logit, adjacent-category logit and continuation-ratio logit. Among of them, the cumulative logit model with proportional odds assumption is commonly employed for ordinal responses. In this article we focus on the proportional odds model to illustrate the proposed test for detecting the random-effects misspecification.

2.1. Observed ordinal data

Consider a longitudinal study consisting of ordinal responses with C categories and p-dimensional covariate vectors, $(Y_{ij}, \mathbf{X}_{ij})$, where Y_{ij} denotes the jth response for subject i, and $\mathbf{X}_{ij}$ represents a $p\times 1$ vector of (discrete or continuous) covariates for subject i at occasion j, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. For simplicity, we assume $n_i \equiv J$. Denote $\mathbf{y}_{ij} = (y_{ij}^{(1)}, \ldots, y_{ij}^{(C)})'$ as a vector of C indicator variables, where $y_{ij}^{(k)} = 1$ if $Y_{ij} = k$, and 0 otherwise, $k = 1, \ldots, C$. The conditional mean for GLMMs is defined by

$$\mathrm{E}(Y_{ij}|\, \mathbf{X}_{ij},\ \mathbf{Z}_{ij}, \mathbf{b}_i\,) = g(\mathbf{X}'_{ij}\,\boldsymbol{\beta} + \mathbf{Z}'_{ij}\,\mathbf{b}_i\,),$$

where g is a monotone differentiable inverse link function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p\times 1$ vector of unknown fixed parameters, $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})'$ is a $q\times 1$ vector of random effects, and $\mathbf{X}'_{ij} = (X_{ij1}, \ldots, X_{ijp})$ and $\mathbf{Z}'_{ij} = (Z_{ij1}, \ldots, Z_{ijq})$ are covariate vectors for the fixed and random effects, respectively. The random effects of $\mathbf{b}_i$ are often assumed to follow a multivariate normal distribution, $\mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0}, \mathbf{D})$, where the covariance matrix $\mathbf{D}$ depends on a unknown parameter vector $\boldsymbol{\gamma}$. The purpose of this article is to diagnose the appropriateness of the distributional assumption of $\mathbf{b}_i$ in GLMMs with cumulative logit link.

The design matrices $\mathbf{X}$ and $\mathbf{Z}$ can be partitioned into $\mathbf{X}_{N\times p} = [\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n]'$ and $\mathbf{Z}_{N\times(nq)} = [\mathbf{Z}_1 \bigotimes \mathbf{1}_1, \ldots, \mathbf{Z}_n \bigotimes \mathbf{1}_n\,]$, where the notation $\bigotimes$ is Kronecker product, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iJ})$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{iJ})'$, and $\mathbf{1}_i$ is an $n\times 1$ column vector having 1 in the ith entry and 0 elsewhere, $i = 1, \ldots, n$, $j = 1, \ldots, J$. Denote $\eta_{ij}^{(k)} = P(Y_{ij} \le k \mid \mathbf{b}_i) = \sum_{c=1}^{k} P(y_{ij}^{(c)} = 1 \mid \mathbf{b}_i) = \sum_{c=1}^{k} E(y_{ij}^{(c)} \mid \mathbf{b}_i)$ as the conditional cumulative probability of the outcome Y_{ij} given $\mathbf{b}_i$ for category k, $k = 1, \ldots, C$. A random-effects version of the cumulative logit model with identical odds ratios across the $C-1$ cutoffs is given by

$$\mathrm{logit}[\eta_{ij}^{(k)}] = \lambda_k - [\,\mathbf{X}'_{ij}\,\boldsymbol{\beta} + \mathbf{Z}'_{ij}\,\mathbf{b}_i\,], \tag{1}$$

where the threshold parameters $\lambda_1, \ldots, \lambda_{C-1}$ satisfy $\lambda_1 \le \cdots \le \lambda_{C-1}$.

The conditional probability of a response falling in category k for subject i at occasion j is denoted by $P(y_{ij}^{(k)} = 1 \mid \mathbf{b}_i\,) = p_{ij}^{(k)} = \pi(\zeta_{ij}^{(k)}) - \pi(\zeta_{ij}^{(k-1)})$, where $\pi(\zeta_{ij}^{(k)}) = \exp(\zeta_{ij}^{(k)})/(1 + \exp(\zeta_{ij}^{(k)}))$, and $\zeta_{ij}^{(k)} = \lambda_k - [\,\mathbf{x}'_{ij}\,\boldsymbol{\beta} + \mathbf{z}'_{ij}\,\mathbf{b}_i\,]$, $k = 1, \ldots, C-1$.