Computational Statistics and Data Analysis 56 (2012) 629–644

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Conditional Akaike information criterion for generalized linear mixed models

Dalei Yu [a,b], Kelvin K.W. Yau [b,*]

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

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

ARTICLE INFO

Article history:
Received 21 January 2011
Received in revised form 8 September 2011
Accepted 12 September 2011
Available online 22 September 2011

Keywords:
Conditional Akaike information
Generalized linear mixed model
Model identification
Poisson time series
Variance component

ABSTRACT

In this study, a model identification instrument to determine the variance component structure for generalized linear mixed models (GLMMs) is developed based on the conditional Akaike information (cAI). In particular, an asymptotically unbiased estimator of the cAI (denoted as cAICC) is derived as the model selection criterion which takes the estimation uncertainty in the variance component parameters into consideration. The relationship between bias correction and generalized degree of freedom for GLMMs is also explored. Simulation results show that the estimator performs well. The proposed criterion demonstrates a high proportion of correct model identification for GLMMs. Two sets of real data (epilepsy seizure count data and polio incidence data) are used to illustrate the proposed model identification method.

1. Introduction

In recent years, generalized linear mixed models (GLMMS) have become a widely used tool in analyzing correlated/longitudinal non-normal data. Early works on the estimation and inference for GLMMS can be found in Schall (1991), Breslow and Clayton (1993), McGilchrist (1994) and Lee and Nelder (1996). The model selection issue arises immediately when there is a set of candidate models for fitting same set of data. Selection methods for fixed/random effects have been studied in the literature. Lee and Nelder (1996) proposed a method to test whether certain random component should be dropped. To assess one or more random effect terms should be included in the model, some researchers have proposed the use of a frequentist score test (Commenges and Jacqmin-Gadda, 1997; Lin, 1997). In the Bayesian framework, Chen et al. (2003) considered a class of information priors for model selection. Chen and Dunson (2003) proposed a hierarchical Bayesian model for random-effects model selection. Moreover, Azari et al. (2006) studied the selection of fixed effects in longitudinal data models while Cai and Dunson (2006) considered a fully Bayesian approach to study the problem of simultaneous selection of fixed and random effects in GLMMS. These studies mainly focus on the aspect of variable selection. However, in many situations, one needs a more general model selection criterion for the GLMMS.

Considering the parameter-driven Poisson models for time series of count data (Zeger, 1988; Chan and Ledolter, 1995), one of the commonly adopted approaches is to assume a process with stochastic autoregressive mean (categorized as parameter driven model). The model was firstly studied by Zeger (1988) to analyze the polio incidence data; see also Chan and Ledolter (1995) for more details. In these studies, the order of the latent autoregressive process is assumed to be one. Putting time series of count data in the framework of GLMM modeling, Yau and Kuk (2002) fitted the polio incidence data by robust methods with the latent autoregressive process also assumed to be order one. In principle, as noted by Chan and

* Corresponding author.
E-mail addresses: msyudl@cityu.edu.hk (D. Yu), mskyau@cityu.edu.hk (K.K.W. Yau).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.09.012

Ledolter (1995), one can consider a higher order autoregressive latent process. However, when the model involves higher orders, determining the optimal choice is not an easy task. The preceding discussion describes the fact that there are different kinds of model selection problems for GLMMs and a general model identification instrument is needed.

In a different but related stream of study, for richly parameterized models, including linear hierarchical models, random effect models, some smoothers and spatial models, the model selection problem is generally difficult. The reason is that it is not straightforward to determine the effective number of parameters and hence the degree of freedom of the model (Ye, 1998; Hodges and Sargent, 2001). Moreover, in GLMMs, the marginal distribution usually does not have an explicit analytic form. As a consequence, the corresponding definition of the Akaike information criterion (Akaike, 1973) is not that straightforward. What likelihood should be chosen if the marginal distribution is not readily available? How to evaluate the correct degree of freedom or model complexity? To address these problems, Ye (1998) proposed a generalized definition of the degree of freedom in complex modeling procedures. Lee and Nelder (1996) derived the degree of freedom from the scaled deviance for the lack of fit test in hierarchical generalized linear models (HGLMs), which extends the scaled residual sum of squares in linear models to HGLMs. Hodges and Sargent (2001) considered an extension of the notion of degree of freedom for richly parameterized models. Spiegelhalter et al. (2002) proposed a Bayesian measure of complexity for hierarchical models.

For the information based criterion, Vaida and Blanchard (2005) defined the conditional Akaike information (cAI) for mixed-effects models. The conditional Akaike information criterion (cAIC) was then derived directly according to this definition, assuming variance component parameters are known. They suggested using a plug-in estimator for the unknown variance component parameters in practice and argued that the impact is negligible. This cAIC is referred to as conventional cAIC in Greven and Kneib (2010). Liang et al. (2008) proposed a corrected version of cAIC to avoid the bias due to assuming known variance component parameters. Greven and Kneib (2010) studied the behavior of cAIC in detail and suggested that ignoring the uncertainty in the estimation of variance component parameters induces a built-in-bias such that the resulting cAIC will favor a more complex model. Greven and Kneib (2010) suggested using the corrected cAIC (Liang et al., 2008) to fix this deficiency and proposed an analytic expression to avoid heavy computational burden. However, such a corrected cAIC is obtained based on normality assumptions and is not directly extendable to GLMMs. For HGLM, Lee et al. (2006) considered an information criterion based on conditional log-likelihood. Ha et al. (2007), proposed some improved information criteria based on extensions of conditional likelihood deviance and studied their relative merit.

The purpose of this paper is to develop the information based model identification criterion based on the conditional Akaike information for GLMMs. In particular, when taking the unknown variance component parameters into consideration, an asymptotically unbiased estimator for cAI (denoted as cAICC) is derived and employed as model selection criterion. The behavior of the criterion is investigated in both theoretical and numerical aspects. Furthermore, an analogous definition of generalized degree of freedom (GDF) for GLMMs is derived and the relationship between the bias correction and the GDF is investigated.

The rest of the paper is arranged as follows. In the next section, the formulation and estimation procedure for GLMMs is briefly described. In Section 3, the model identification criterion for GLMMs is studied and major results are presented. In Section 4 the performance of the model selection criterion is assessed via simulation study under different settings. Two sets of real data are then used to demonstrate the proposed model selection method. The final section gives some concluding remarks.

2. Model formulation and estimation

Consider an n-dimensional observable response vector w, which has a distribution depending on the vector quantity η defined by

$$\eta = X\beta + Zu, \tag{1}$$

where X is an $n \times k$ matrix of rank k, $Z = (Z_1, \ldots, Z_l)$ is an $n \times v$ matrix of rank v, β is the k-dimensional fixed effect vector and the v-dimensional random effects $u = (u_1^{\mathrm{T}}, \ldots, u_l^{\mathrm{T}})^{\mathrm{T}}$. Z and u are partitioned conformally and u_j are independent random effects distributed as v_j-dimensional $N\{0, \tau_j^2 A_j(\varphi)\}$, $\sum_{j=1}^{l} v_j = v$. τ_j^2 and φ (p-dimensional vector) are the unknown variance component parameters. For convenience, denote $A = \mathrm{diag}\{\tau_j^2 A_j(\varphi)\}$, $\lambda = (\varphi^{\mathrm{T}}, \tau_j^2)^{\mathrm{T}}$, $j = 1, \ldots, l$.

Let $l_1(w \mid \beta, u)$ be the log-likelihood of w given u, $l_2(u \mid \varphi, \tau_j^2)$ be the logarithm of the probability density function of u. As noted by McGilchrist (1994), the BLUP-type log-likelihood

$$h = l_1 + l_2$$

carries over the spirit of best linear unbiased predictors (BLUP) into a non-normal framework. The estimators $\hat{\beta}_w$ and $\hat{u}_w$ are found by maximizing h and these are called penalized likelihood estimators (PLE). Equivalent estimators can also be derived in the hierarchical modeling framework (Lee and Nelder, 1996). To find the maximum, take the derivative in both sides of h with respect to β and u, it follows that

$$\frac{\partial h(w \mid \beta, u)}{\partial \beta} = X^{\mathrm{T}} \frac{\partial l_1(w \mid \beta, u)}{\partial \eta} \tag{2}$$