Computational Statistics and Data Analysis 55 (2011) 703–714

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Detecting random-effects model misspecification via coarsened data

Xianzheng Huang*

Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

ARTICLE INFO

Article history:
Received 29 June 2009
Received in revised form 10 June 2010
Accepted 11 June 2010
Available online 20 June 2010

Keywords:
Generalized linear mixed models
Kullback–Leibler divergence
Nonlinear mixed models

ABSTRACT

Mixed effects models provide a suitable framework for statistical inference in a wide range of applications. The validity of likelihood inference for this class of models usually depends on the assumptions on random effects. We develop diagnostic tools for detecting random-effects model misspecification in a rich class of mixed effects models. These methods are illustrated via simulation and application to soybean growth data.

1. Introduction

Mixed effects models are widely used in statistical applications in biology, agriculture, sociology, and environmental science, where clustered data are often collected. Under the framework of mixed effects models, it is straightforward to make prediction at the inter-cluster level as well as the intra-cluster level. Davidian and Giltinan (2003) and McCulloch et al. (2008) provide a comprehensive survey of mixed effects models. A major concern in this realm lies in random-effects assumptions. Hartford and Davidian (2000), Heagerty and Kurland (2001), Agresti et al. (2004) and Litière et al. (2007, 2008) showed various adverse effects of erroneous random-effects assumptions on inference.

Semiparametric and nonparametric methods have been developed to relax the parametric assumptions on random effects. Davidian and Gallant (1993) used a flexible, smooth density to characterize the distribution of random effects. Fattinger et al. (1995) used spline functions to transform the random effects, initially assumed to be normal, so that the resulting random effects can follow an arbitrary distribution dictated by data. Lai and Shih (2003) took a nonparametric approach to estimate the distribution of random effects. The price one pays to avoid suspicious parametric assumptions by using these methods is often heavy computation. Some of these methods only provide discrete estimation for the random-effects distribution. This can be unsatisfactory when the distributional characteristics of random effects are of interest. Because a parsimonious parametric model can be appealing for its simplicity and potential gain in efficiency, techniques to check parametric assumptions on random effects in mixed effects models are valuable.

There are three main lines of development thus far for assessing the validity of model specification. The first line of work is based on the information matrix equivalence under a correct model (White, 1981, 1982). Focusing on the variance–covariance structure for the random effects in nonlinear mixed models (NLMM), Vonesh et al. (1996) compared two variance–covariance estimators that are equivalent under the correct model. Litière (2007) developed a series of tests motivated by the information matrix equivalence to diagnose random-effects misspecification in linear mixed models (LMM) and generalized linear mixed models (GLMM). The second line of development aims at obtaining an empirical distribution of the random effects. For example, Lange and Ryan (1989) estimated the cumulative distribution of random effects using the empirical Bayes estimates of individual random effects. Ritz (2004) extended their results and derived a

* Tel.: +1 803 777 8772; fax: +1 803 777 4048.
E-mail address: huang@stat.sc.edu.

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.06.012

weighted empirical process for the random effects. Waagepetersen (2006) simulated random effects from the conditional distribution of random effects given the observed data in GLMM. The third line of research has some similarity with the first line but is relatively underdeveloped. Instead of comparing two information matrices, one compares two estimators for the fixed effects, with one robust to model misspecification and the other sensitive to it. This idea was once prevalent in the econometrics community (Hausman, 1978) and was recently used by Tchetgen and Coull (2006) to construct a test for a specific form of GLMM.

These existing diagnostic methods are either too general to provide guidance for correction when misspecification is detected or too specific to allow immediate extension to complex models or to detect departures other than normality assumption. Admittedly, as shown by Verbeke and Molenberghs (submitted for publication), without assuming other component models in a hierarchical mixed effects model correctly specified, random-effects assumptions are unverifiable. With multiple uncertain component models, one can test the sensitivity of inference to multiple model assumptions but may not be able to single out the assumptions on random effects. If one has more confidence in the other component models than in the random effects, then it is possible to verify random-effects assumptions by studying the robustness of inference under these assumptions. This is the underlying philosophy of the methods discussed in this article. As a starting point, Huang (2009) provided empirical evidence that MLEs in a GLMM for binary response are not robust to data grouping in the presence of random-effects model misspecification. Motivated by this finding, she constructed diagnostic tests based on the discrepancy between the observed-data MLEs and the MLEs from the induced grouped data. Compared to the aforementioned existing methods, her tests can be more informative because the outcomes of the tests can depend on the source of model misspecification besides its existence.

In this article, we shed further light on the idea motivating Huang's method, which is the content of Section 2. In Section 3, this idea is extended to a richer class of mixed effects models where the response is not binary. New test statistics that are computationally more efficient than those in Huang (2009) are defined in Section 4, where simulation studies are also presented. In Section 5, the diagnostic techniques are applied to a real data example. Concluding remarks and future research topics are given in Section 6.

2. Diagnostic method for GLMM

2.1. Test statistics

Denote by $\boldsymbol{\Omega}$ the $p \times 1$ vector of unknown parameters in the model, by $\widetilde{\boldsymbol{\Omega}}$ and $\widetilde{\boldsymbol{\Omega}}_c$ the limiting MLE based on the observed data and that based on the grouped data, respectively, where the limit is taken as the number of clusters, m, tends to infinity and the size of the cluster is bounded. Under a correct model, it is expected that $\widetilde{\boldsymbol{\Omega}} = \widetilde{\boldsymbol{\Omega}}_c$, and erroneous model assumptions can result in $\widetilde{\boldsymbol{\Omega}} \neq \widetilde{\boldsymbol{\Omega}}_c$. Following this claim, Huang (2009) considered testing $H_0 : \widetilde{\boldsymbol{\Omega}} = \widetilde{\boldsymbol{\Omega}}_c$, and proposed test statistics that compare the finite-sample counterparts of $\widetilde{\boldsymbol{\Omega}}$ and $\widetilde{\boldsymbol{\Omega}}_c$, denoted by $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Omega}}_c$. The follow test statistics implement elementwise comparison,

$$\mathbf{t}_1 = (\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{\Omega}}_c)\#\text{vecdiag}(\hat{\mathbf{V}}_1^{-1}), \tag{1}$$

where $\hat{\mathbf{V}}_1$ is an estimator for the variance–covariance matrix of $\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{\Omega}}_c$, $\text{vecdiag}(\hat{\mathbf{V}}_1^{-1})$ denotes the column vector consisting of the diagonal elements of $\hat{\mathbf{V}}_1^{-1}$, and "#" is the elementwise multiplication operator. Detailed derivation of $\hat{\mathbf{V}}_1$ is given in Huang (2009), where it is shown that, under H_0, an element in $\mathbf{t}_1$ associated with a parameter γ, $t_1(\gamma)$, follows a Student's t distribution with $m - p$ degrees of freedom. Another test statistic compares the entire vector of MLEs as follows,

$$F_1 = \frac{m-p}{p(m-1)}(\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{\Omega}}_c)^{\mathrm{T}}\hat{\mathbf{V}}_1^{-1}(\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{\Omega}}_c), \tag{2}$$

which follows an $F(p, m-p)$ distribution under H_0. A significantly large value of F_1 casts doubt about the veracity of H_0. Individual values of $t_1(\gamma)$ assess the sensitivity of individual parameter estimate to data grouping, which can relate to how a model is misspecified. These tests do not directly test the validity of random-effects assumptions. Rather, they are designed to test if the inference is robust to data grouping under these model assumptions, and being robust is no guarantee for correct model assumptions, but is merely some reassurance that the inference, even if it is inconsistent, may not deteriorate (in terms of consistency but not efficiency) when grouped data are used. In the next subsection, we use a concrete example to elaborate the fundamental idea behind these tests and, by so doing, give one some confidence in extending this idea to other mixed effects models.

2.2. Limiting maximum likelihood estimators and coarsened data

Grouped data is a special case of coarsened data (Heitjan and Rubin, 1991; Tsiatis, 2006). Other examples of coarsened data common in practice include censored data, incomplete data due to missingness, rounded data, etc. Before focusing on grouped data, it is instructive to first introduce the notion of coarsened data generically. Let $(\boldsymbol{\Delta}_i, \mathbf{Y}_i^*)$ be the ith datum in the coarsened data, where $\boldsymbol{\Delta}_i$ is the coarsening variable (if needed), and $\mathbf{Y}_i^* = C_{\boldsymbol{\Delta}_i}(\mathbf{Y}_i)$ is the coarsened response, $C_{\boldsymbol{\Delta}_i}(\mathbf{Y}_i)$ is