Journal of the Korean Statistical Society 42 (2013) 215–225

Contents lists available at SciVerse ScienceDirect

Journal of the Korean Statistical Society

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

Empirical likelihood inference in mixtures of semiparametric varying coefficient EV models for longitudinal data with nonignorable dropout

Xing-cai Zhou [a,b,*], Jin-Guan Lin [a]

[a] Department of Mathematics, Southeast University, Nanjing 210096, China

[b] Department of Mathematics and Computer Science, Tongling University, Tongling, Anhui, 244000, China

ARTICLE INFO

Article history:
Received 2 November 2011
Accepted 24 August 2012
Available online 13 September 2012

AMS 2000 subject classifications:
primary 62H12
secondary 62A10

Keywords:
Empirical likelihood
Varying coefficient
Errors-in-variables
Longitudinal data
Nonignorable dropout
Confidence regions

ABSTRACT

In this paper, empirical likelihood inference in mixtures of semiparametric varying coefficient errors-in-variables (EV) models for longitudinal data with nonignorable dropout is investigated. The empirical log-likelihood ratio statistic for the fixed-effects parameters and the mean parameters of random effects are proposed. The proposed statistic at the true parameters is proven to be asymptotically χ^2_{q+r}, where q and r are the dimensions of the fixed and random effects respectively, and the corresponding confidence regions for the parameters of interest are then constructed. We also obtain the maximum empirical likelihood estimator of the parameters, and prove that it is asymptotically normal under some suitable conditions. Simulation studies are undertaken to assess the finite sample performance of the proposed method.

1. Introduction

Longitudinal data arises frequently in biomedical and economic applications, which can be easily found in medical follow-up studies for monitoring disease progression. Longitudinal data is characterized by the correlation among the observations with the same subject, and the independence among different subjects. In the long-term longitudinal studies, dropout and other types of missing data are common; in many cases, dropout induces a missingness process that is nonignorable in the sense that missingness depends probabilistically on unobserved outcomes, ever after conditioning on observable information (Hogan, Lin, & Herman, 2004). The model-based approaches are often used to handle informative dropout in longitudinal data, for example, likelihood-based approaches including selection models (Diggle & Kenward, 1994; Ten Have, Kunselman, Pulkstenis, & Landis, 1998) and mixture models (Hogan & Laird, 1997b; Hogan et al., 2004; Little, 1994), and moment-based methods under the selection modeling framework (Rotnitzky, Robins, & Scharfstein, 1998; Scharfstein, Robins, & Rotnitzky, 1999). Hogan et al. (2004) proposed mixtures of varying coefficient models, which can be viewed as an extension of pattern-mixture models (Little, 1994) and conditional linear models (Hogan & Laird, 1997a), to analyze longitudinal data where dropout might be at continuous times and potentially nonignorable. Li and Xue (2010) extended the model of Hogan et al. (2004) to the mixtures of semiparametric varying coefficient models, where the responses follow a semiparametric varying coefficient random effects model conditional on dropout time, and some of the regression coefficients depend on dropout time through unspecified nonparametric functions. Suppose an experiment with n subjects

* Corresponding author at: Department of Mathematics, Southeast University, Nanjing 210096, China.
E-mail address: xczhou@nuaa.edu.cn (X.-c. Zhou).

1226-3192/$ – see front matter
doi:10.1016/j.jkss.2012.08.002

and m_i observations in the ith subject ($i = 1, \ldots, n$) for a total of $N = \sum_{i=1}^{n} m_i$. For the ith subject, let $\boldsymbol{Y}_i$ be an $m_i \times 1$ observed outcome vector, $\boldsymbol{X}_{ij}$ and $\boldsymbol{W}_{ij}$ respectively be $p \times 1$ and $q \times 1$ covariates associated with fixed effects, $\boldsymbol{Z}_{ij}$ be $r \times 1$ covariate associated with random effects, and U_i be the dropout time, and denote $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im_i})^T$, $\boldsymbol{W}_i = (\boldsymbol{W}_{i1}, \ldots, \boldsymbol{W}_{im_i})^T$ and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i})^T$. The mixtures of semiparametric varying coefficient models for longitudinal data with nonignorable dropout have the form

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\alpha}(U_i) + \boldsymbol{W}_i \boldsymbol{\beta}_1 + \boldsymbol{Z}_i \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{1.1}$$

where $\boldsymbol{\alpha}(u) = (\alpha_1(u), \ldots, \alpha_p(u))^T$ is a $p \times 1$ vector of unknown regression coefficient functions of the dropout time u; $\boldsymbol{\beta}_1$ is a $q \times 1$ vector of unknown regression coefficient; $\boldsymbol{\gamma}_i$ is a $r \times 1$ vector of random effects with $E(\boldsymbol{\gamma}_i | \boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i) = \boldsymbol{\gamma}$, which is unknown; ϵ_i is an $m_i \times 1$ vector of residuals with $E(\boldsymbol{\epsilon}_i | \boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i) = 0$, and $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\gamma}_i$ are mutually independent. When $\boldsymbol{\beta}_1 = 0$, i.e. the parametric component is removed, the model (1.1) reduces to the model of Hogan et al. (2004), who estimated the shapes of the function $\boldsymbol{\alpha}(u)$ using the natural cubic smoothing splines. For estimating the parameters of the model (1.1), Li and Xue (2010) proposed the local linear version of the profile-kernel method, and showed the asymptotic normality of the estimator of $\boldsymbol{\beta}_1$ under some regular conditions.

In many applications, however, there often exist covariates with measurement errors, which may cause difficulties and complications in conducting statistical analysis. In the last two decades, a lot of effort has been dedicated to the error-in-variables (EV) model in the literature. Comprehensive reviews on the EV model can be found in Fuller (1987) and Carroll, Ruppert, and Stefanski (1995), and the references therein. Some work has been done in the estimation of the involved parameter in the partially linear EV model, e.g., Cui and Li (1998), Li and Xue (2008), Liang (2000), Liang, Häardle, and Carroll (1999) and Zhu and Cui (2003). For the purely varying coefficient partially linear EV model, You, Zhou, and Chen (2006) studied the estimation of parametric component, and focused on how to change the usual local polynomial technique to find consistent estimators of the regression coefficient functions. For the semiparametric varying coefficient partially linear EV model, You and Chen (2006) proposed a modified profile least squares estimator for the parametric component and a local polynomial estimator for the nonparametric component by correcting the attenuation, and showed that the modified profile least squares estimator is consistent and asymptotically normal; and Wang, Li, and Lin (2011) further studied the model by the empirical likelihood method. When $\boldsymbol{\alpha}(U_i) = \alpha$ where α is a constant vector and $m_i = 1$, the model (1.1) becomes the mixed effects model. The mixed effects EV model has also been considered by Chen, Zhong, and Cui (2009), Cui, Ng, Kai, and Zhu (2004) and Zhong, Fung, and Wei (2002). Cui (2004) gave the parameter estimators in the partial linear EV model with replicated observations and studied their asymptotic properties. Recently, Zhao and Xue (2009) investigated the semiparametric varying coefficient partially linear EV model with longitudinal data. However, a statistical analysis of the mixtures of the semiparametric varying coefficient EV model for longitudinal data with nonignorable dropout still seems to be missing.

In this paper, we consider model (1.1) in the case where the covariates $\boldsymbol{W}_{ij}$ and $\boldsymbol{Z}_{ij}$ are measured with additive errors, and $\boldsymbol{X}_{ij}$ and U_i are error free. That is, we cannot observe $\boldsymbol{W}_{ij}$ and $\boldsymbol{Z}_{ij}$, but can observe $\boldsymbol{W}^*_{ij}$ and $\boldsymbol{Z}^*_{ij}$ with

$$\begin{cases} \boldsymbol{W}^*_{ij} = \boldsymbol{W}_{ij} + \boldsymbol{\mu}_{ij}, \\ \boldsymbol{Z}^*_{ij} = \boldsymbol{Z}_{ij} + \boldsymbol{\nu}_{ij}, \end{cases} \quad i = 1, \ldots, n, j = 1, \ldots, m_i, \tag{1.2}$$

where $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\nu}_{ij}$ are the measurement errors with $E(\boldsymbol{\mu}_{ij}) = E(\boldsymbol{\nu}_{ij}) = 0$, and are assumed to have known positive definite covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$, respectively, as in You and Chen (2006), Zhao and Xue (2009) and Zhu and Cui (2003) among others, and $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\nu}_{ij}$ are independent. Denote $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{im_i})^T$ and $\boldsymbol{\nu}_i = (\boldsymbol{\nu}_{i1}, \ldots, \boldsymbol{\nu}_{im_i})^T$. We further assume that $\{\boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{Z}_i, U_i, \boldsymbol{\gamma}_i, \boldsymbol{\mu}_i, \boldsymbol{\nu}_i, \boldsymbol{\epsilon}_i\}$ are independent. When $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$ are unknown, we can estimate them by repeatedly measuring $\boldsymbol{W}^*_{ij}$ and $\boldsymbol{Z}^*_{ij}$ respectively; see Liang et al. (1999) for details.

Constructing confidence regions for the fixed effects parameter $\boldsymbol{\beta}_1$ and the mean parameter $\boldsymbol{\gamma}$ of random effects in model (1.1)–(1.2) are of great interest to many practitioners. In this paper, the empirical likelihood method is adopted to investigate the mixtures of semiparametric varying coefficient EV models (1.1)–(1.2) for constructing the confidence regions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}$. One motivation is that empirical likelihood inference does not involve the asymptotic covariance of the estimators for parameter components, which is a rather complex structure for the EV model under longitudinal data. Another motivation is that empirical likelihood has many advantages, for example, the shape of an empirical-likelihood-based confidence region does not symmetry predetermine so that it can be determined automatically by the data, and the region is range preserving and transformation respecting; see Owen (2001). Recently, the empirical likelihood method has been successfully applied to a large class of EV models; e.g., linear EV models (Cui & Chen, 2003), nonlinear EV models (Stute, Xue, & Zhu, 2007), linear mixed effects EV models (Chen et al., 2009), partially linear EV models (Li & Xue, 2008), varying coefficient partially linear EV models (Wang et al., 2011), and varying coefficient partially linear EV models with longitudinal data (Zhao & Xue, 2009).

The rest of the paper is organized as follows. The methodology and the main results are introduced in Section 2. Section 3 reports the results of some simulation studies. In Section 4, we present a brief discussion of the results and methods. Some assumptions and the proofs of the main results are stated in the Appendix.