Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

ELSEVIER

Q1 Subject-wise empirical likelihood inference in partial linear models for longitudinal data

Q2 Lianfen Qian [a,*,1], Suojin Wang [b]

[a] Department of Mathematical Sciences, Florida Atlantic University, Boca Raton, FL 33431, United States

[b] Department of Statistics, Texas A&M University, College Station, TX 77843, United States

A R T I C L E I N F O

Article history:
Received 24 June 2016
Received in revised form 2 December 2016
Accepted 2 February 2017
Available online xxxx

Keywords:
Confidence region
Empirical likelihood
Longitudinal data
Maximum empirical likelihood estimator

A B S T R A C T

In analyzing longitudinal data, within-subject correlations are a major factor that affects statistical efficiency. Working with a partially linear model for longitudinal data, a subject-wise empirical likelihood based method that takes the within-subject correlations into consideration is proposed to estimate the model parameters. A nonparametric version of the Wilks Theorem for the limiting distribution of the empirical likelihood ratio, which relies on a kernel regression smoothing method to properly centered data, is derived. The estimation of the nonparametric baseline function is also considered. A simulation study and an application are reported to investigate the finite sample properties of the proposed method. The numerical results demonstrate the usefulness of the proposed method.

1. Introduction

In a longitudinal study, subjects are repeatedly measured over time. Two major features of the data resulted from a longitudinal study are the aging effect and the within-subject correlation structure. To estimate the aging effect, it is well known that a fully nonparametric regression model suffers from the "curse of dimensionality". To avoid the difficulty, a partially linear model is popularized. That is, one proposes some regression structure on covariate effects but leave the time effect nonparametric.

On the other hand, the within-subject correlation structure is a much more difficult problem. While the method of generalized estimating equations (GEEs) by Liang and Zeger (1986) is very useful, it has been shown that ignoring the within-subject correlation structure may result in substantial loss of efficiency in estimation and inference; see, for example, Albert and McShane (1995), Fitzmaurice (1995), Hall and Severini (1998), Wang and Carey (2003), and Wang et al. (2010).

In this paper, we plan to tackle the latter problem in a partial linear model via an empirical likelihood method. To be more precise, let $Y(t)$ be the response variable and $X(t)$ be the q dimensional covariate vector at time t. Assume the baseline function of the response variable is $g(t)$ at time t. We consider the following partial linear model for longitudinal data:

$$Y(t) = \beta^T X(t) + g(t) + \epsilon(t), \quad (1)$$

where β is the q dimensional vector of the regression coefficient parameters and $\epsilon(t)$ is a stochastic error process with $E\{\epsilon(t)\} = 0$. We wish to estimate the model parameters β and the baseline function $g(t)$ simultaneously. Assume that t is bounded in asymptotic considerations.

* Corresponding author.
E-mail address: Lqian@fau.edu (L. Qian).

1 A special appointed Professor of Wenzhou University, Zhejiang, China.

http://dx.doi.org/10.1016/j.csda.2017.02.001
0167-9473/

Please cite this article in press as: Qian, L., Wang, S., Subject-wise empirical likelihood inference in partial linear models for longitudinal data. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.02.001

Taking expectation on both sides of (1) we have

$$E[Y(t)] = \beta^T E[X(t)] + g(t). \quad (2)$$

Subtracting (2) from (1) we obtain

$$Y(t) - E[Y(t)] = \beta^T \{X(t) - E[X(t)]\} + \epsilon(t).$$

Suppose that we have a data set that contains n subjects. For each subject i, the measurements are made at times $t_{i1}, \ldots, t_{im_i}$. Thus the model under consideration is

$$Y_i(t_{ij}) = \beta^T X_i(t_{ij}) + g(t_{ij}) + \epsilon_i(t_{ij}), \quad i = 1, \ldots, n, j = 1, \ldots, m_i,$$

with the centered sample model as follows:

$$Y_i(t_{ij}) - E[Y_i(t_{ij})] = \beta^T \left\{X_i(t_{ij}) - E[X_i(t_{ij})]\right\} + \epsilon_i(t_{ij}). \quad (3)$$

For simplicity, we denote $Y_{ij} = Y_i(t_{ij})$, $X_{ij} = X_i(t_{ij})$ and $t_i = (t_{i1}, \ldots, t_{im_i})^T$. We introduce the vector notation: $Y_i = Y_i(t_i) = (Y_{i1}, \ldots, Y_{im_i})^T$, $X_i = X_i(t_i) = (X_{i1}, \ldots, X_{im_i})^T$ and $\boldsymbol{\epsilon}_i = \epsilon_i(t_i) = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^T$. Let $\widetilde{Y}_i = Y_i - E(Y_i)$, $\widetilde{X}_i = X_i - E(X_i)$ be the mean centered measurements. Then the matrix representation of (3) is

$$\widetilde{Y}_i = \widetilde{X}_i \beta + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n. \quad (4)$$

Since the error process is unknown, Owen's empirical likelihood method (1988) is suitable. Xue and Zhu (2007) studied the estimation problem using empirical likelihood method without considering the within-subject correlation structure. Fan and Li (2004) studied two estimation methods: difference-based estimator and profile least squares approach. Neither of the papers took the within-subject correlation into consideration. Wang et al. (2010) tackled this problem with two empirical likelihood based methods in estimating the regression coefficients and obtained promising results. In this paper, we adopt one of the methods, the subject-wise empirical likelihood approach, to analyze longitudinal data in partial linear models. In particular, we will make use of the centered sample model (3) in our procedure.

The rest of the paper is organized as follows. In Section 2, we propose a subject-wise empirical likelihood estimator for partial linear models with the aim of achieving better efficiency. Some analytic justifications are provided for our proposed approach. While our main focus is on how to estimate parameters β more efficiently in model (1) with a proper confidence region, we briefly address the issue of estimating the baseline function. In Sections 3 and 4 we report some results of our empirical studies on both simulated and real data sets. Some concluding remarks are given in Section 5.

2. Subject-wise empirical likelihood estimation

2.1. Estimation of the coefficient parameters

We propose a new method to estimate the regression coefficients with better efficiency via empirical likelihood in the partial linear model (1). Recall that in Eq. (4), the components within each $\boldsymbol{\epsilon}_i$ are allowed to be correlated.

Let $\Sigma_i = \mathrm{cov}(Y_i|X_i) = \mathrm{cov}(\widetilde{Y}_i|\widetilde{X}_i)$ and $\widehat{\Sigma}_{in}$ be an estimator of Σ_i. Furthermore, we assume that $\widehat{\Sigma}_{in}$ converges to Σ_i^*, an $m_i \times m_i$ positive definite matrix. Let

$$\widetilde{Z}_i(\beta) = \left\{\widetilde{z}_{i1}(\beta), \ldots, \widetilde{z}_{im_i}(\beta)\right\}^T = \Sigma_i^{*-1}(\widetilde{Y}_i - \widetilde{X}_i\beta). \quad (5)$$

If $\Sigma_i^* = \Sigma_i$ for all $i = 1, \ldots, n$, then $\widetilde{Z}_i(\beta)$ is the usual generalized least squares estimating function. If $\Sigma_i^* = \sigma^2 I_{m_i}$ for some $\sigma^2 > 0$, where I_{m_i} is the m_i dimensional identity matrix, then $\widetilde{Z}_i(\beta)$ is the estimating function with the working independence assumption. Notice that $E[\widetilde{Z}_i(\beta)] = 0$ when β is the true parameter. Define the auxiliary variable

$$\widetilde{\Psi}_i(\beta) = \sum_{i=1}^{m_i} \widetilde{z}_{ij}(\beta)\widetilde{X}_{ij} = \widetilde{X}_i^T \widetilde{Z}_i(\beta).$$

Let $m_Y(t) = E[Y(t)]$ and $m_X(t) = E[X(t)]$ be the mean processes of $Y(t)$ and $X(t)$, respectively. Notice that $m_Y(t)$ and $m_X(t)$ are unknown. Hence, $\widetilde{Z}_i(\beta)$ contains unknown functions. Thus we need to estimate the two mean functions against time t first. One way to handle this is to use kernel smoothing method. Let $\widehat{m}_Y(t)$ and $\widehat{m}_X(t)$ be the kernel estimators of $m_Y(t)$ and $m_X(t)$ with kernel density function K. To be precise, that is

$$\widehat{m}_Y(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} Y_{ij} K\left(\frac{t_{ij}-t}{h}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} K\left(\frac{t_{ij}-t}{h}\right)},$$