Statistics and Probability Letters 119 (2016) 84–90

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Automatic variable selection for varying coefficient models with longitudinal data

Ruiqin Tian [a,*], Liugen Xue [b], Dengke Xu [a]

[a] Department of Statistics, Zhejiang Agriculture and Forestry University, Hangzhou, 311300, China

[b] College of Applied Sciences, Beijing University of Technology, Beijing, 100124, China

ARTICLE INFO

Article history:
Received 14 March 2016
Received in revised form 11 July 2016
Accepted 14 July 2016
Available online 25 July 2016

MSC:
62G05
62G20

Keywords:
Varying coefficient models
Variable selection
Longitudinal data
Generalized estimating equations
Quadratic inference function

ABSTRACT

We propose a novel variable selection for varying coefficient models with longitudinal data. The theoretical properties of the resulting estimators are established. In addition, simulation studies and a real data set are conducted to evaluate the proposed method.

1. Introduction

The varying coefficient model is an important generalization of the linear regression model and has gained a lot of popularity during the past decade. Due to its flexibility to explore the dynamic features which may exist in the data and its easy interpretation, the varying coefficient models have been widely used in various disciplines, such as finance, economics, medicine, ecology, and biology. It has also experienced rapid developments in both theory and methodology. For example, Hastie and Tibshirani (1993) studied the varying coefficient model; the other discussions and extensions about varying coefficient model can also be seen in Fan and Zhang (1999), Chiang et al. (2001), Huang et al. (2002), Lai et al. (2016) and so on.

Longitudinal data are often collected from experimental studies or clinical trials. The subjects under study are measured repeatedly during the period of the study. A major aspect of longitudinal data is the within-subject correlation among the repeated measurement. Ignoring this within-subject correlation causes a loss of efficiency in general problems. The commonly used regression method for analysis are generalized estimating equations (GEE). GEE using a working correlation matrix with nuisance parameters estimate regression parameters consistently even when the correlation structure is misspecified. However, under such misspecification, the estimator can be inefficient. For this reason, Qu et al. (2000) proposed a method of quadratic inference function (QIF). It avoids estimating the nuisance correlation structure parameters by assuming that the inverse of working correlation matrix can be approximated by a linear combination of several known basis matrices. The QIF can efficiently take the within-cluster correlation into account and is more efficient than the GEE

* Corresponding author.
E-mail address: tianruiqin@hotmail.com (R. Tian).

http://dx.doi.org/10.1016/j.spl.2016.07.012
0167-7152/

approach when the working correlation is misspecified. The QIF estimator is also more robust against contamination when there are outlying observations (Qu and Song, 2004).

Recently, variable selection is an important topic in all regression analysis and many procedures have been developed for this. Generally speaking, most of the variable selection procedures are based on penalized estimation using penalty functions. Such as, L_q penalty (Frank and Friedman, 1993), LASSO penalty (Tibshirani, 1996), SCAD penalty (Fan and Li, 2001), and so on. These variable selection procedures are based on penalized estimation using penalty functions, which have a singularity at zero. Therefore, these variable selection procedures require convex optimization, which will incur a computational burden. To overcome this problem, Ueki (2009) developed a new variable selection procedure called the smooth-threshold estimating equations that can automatically eliminate irrelevant parameters by setting them as zero. Recently, Lai et al. (2012) explored GEE estimation and smooth-threshold GEE variable selection for single-index models with clustered data. Li et al. (2013) considered variable selection for the generalized linear models with longitudinal data using smooth-threshold generalized estimating equations.

In this paper we use the smooth-threshold generalized estimating equations based on quadratic inference function (SGEE-QIF) to the varying coefficient model with longitudinal data. The proposed procedure automatically eliminates the irrelevant coefficient functions by setting them as zero, and simultaneously estimates the nonzero regression coefficient functions by solving the SGEE-QIF. Compared with the shrinkage methods and the existing research findings reviewed above, our method offers the following improvements. Firstly, our approach can be easily implemented without solving any convex optimization problems, and possess the oracle property. Secondly, compared with Ueki (2009) and Li et al. (2013), we extend the smooth-threshold estimating equations approach to the varying coefficient model for longitudinal data, which have become a favored tool for modeling longitudinal data. Thirdly, the proposed method is easy to deal with the longitudinal correlation structure.

The rest of this paper is organized as follows. In Section 2 we first propose a variable selection procedure for varying coefficient model with longitudinal data. Asymptotic properties of the resulting estimators are considered in Section 3. In Section 4 we give the choice of the tuning parameters. In Section 5 we carry out simulation studies to assess the finite sample performance of the method. We further illustrate the proposed methodology via a real data analysis in Section 6. Section 7 contains some concluding remarks. The technical proofs of all asymptotic results are provided in the Supplementary Material (see Appendix A).

2. Methodology

2.1. Quadratic inference functions

In a longitudinal study, n subjects and m_i observations over time for the ith subject ($i = 1, \ldots, n$) for a total of $N = \sum_{i=1}^{n} m_i$ observation. Each observation consists of a response variable Y_{ij} and a covariate vector $X_{ij} \in R^p$, $j = 1, \ldots, m_i$. We assume that the varying coefficient model satisfies the following mean structure:

$$Y_{ij} = X_{ij}^T \theta(U_{ij}) + \varepsilon_{ij}, \tag{2.1}$$

where $\theta(\cdot)$ is a p-vector of unknown functions, ε_{ij} is random error with $E(\varepsilon_{ij}|X_{ij}) = 0$. In addition, we give assumptions on the first two moments of the observations $\{Y_{ij}\}$, that is, $E(Y_{ij}) = \mu_{ij} = X_{ij}^T \theta(U_{ij})$ and $\mathrm{Var}(Y_{ij}) = v(\mu_{ij})$, where $v(\cdot)$ is a known variance function. Here, we are interested in the longitudinal data setting where data are correlated within the same subject, while the observations from different subjects are independent.

Following He et al. (2002), we replace $\theta(\cdot)$ by its basis function approximations. More specifically, let $B(u) = (B_1(u), \cdots, B_L(u))^T$ be B-spline basis functions with the order of M, where $L = K + M$, and K is the number of interior knots. We use the B-spline basis functions because they have bounded support and are numerically stable (Schumaker, 1981). Selection of knots is generally an important aspect of spline smoothing. In this paper, similar to He et al. (2005), the number of internal knots is taken to be the integer part of $N^{1/5}$. Then, $\theta_k(u)$ can be approximated by

$$\theta_k(u) \approx B(u)^T \beta_k, \quad k = 1, \ldots, p. \tag{2.2}$$

Let $W_{ij} = I_p \otimes B(U_{ij}) \cdot X_{ij}$, $\beta = (\beta_1^T, \ldots, \beta_p^T)^T$, $Y_i = (Y_{i1}, \ldots, Y_{im_i})^T$, and write X_i, W_i, ε_i in a similar fashion. According to the GEE method from Liang and Zeger (1986), we can define the following generalized estimating function for β

$$U(\theta, \alpha) = \sum_{i=1}^{n} W_i^T V_i^{-1} (Y_i - W_i \beta), \tag{2.3}$$

where V_i is the covariance matrix of Y_i. Following Liang and Zeger (1986), we simplify the covariance of the ith subject V_i by taking $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$, where $A_i = \mathrm{diag}(\mathrm{Var}(Y_{i1}), \ldots, \mathrm{Var}(Y_{im_i}))$, $R(\alpha)$ is a common working correlation with a nuisance parameters α. If the working correlation $R(\alpha)$ is misspecified, the estimator of the regression parameter is still consistent, but is not efficient.