Journal of Statistical Planning and Inference 160 (2015) 89–106

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Varying-coefficient mean–covariance regression analysis for longitudinal data

Shu Liu [a,*], Guodong Li [b]

[a] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, PR China

[b] Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form 19 September 2014
Accepted 3 November 2014
Available online 26 December 2014

MSC:
primary 62H12
secondary 62A10

Keywords:
Within-subject correlation
Efficiency
Local polynomial regression
Modified Cholesky decomposition
Longitudinal data
Varying-coefficient

ABSTRACT

By considering within-subject correlation among repeated measures over time, we propose a new and efficient estimation of varying-coefficient models for longitudinal data. Based on a modified Cholesky decomposition, the within-subject covariance matrix is decomposed into a unit triangular matrix involving generalized autoregressive coefficients and a diagonal matrix involving innovation variances. Local polynomial smoothing method is used to estimate the unknown varying coefficient functions of marginal mean and innovation variances. A method is also developed to estimate the autoregressive coefficients. All the resulting estimators are shown to be consistent and asymptotically normal. The proposed estimator of varying coefficient functions are asymptotically more efficient than the ones which ignore the within-subject correlation structure. Simulations are conducted to demonstrate finite sample behaviors of the proposed estimators, and a real example is given to illustrate the value of the proposed methodology.

1. Introduction

In biomedical studies, subjects are often measured repeatedly over a given time period, so that the measurements of each subject are possibly correlated with each other but different subjects can be assumed to be independent. Data of this type are frequently referred to as longitudinal samples. The methodology for parametric longitudinal data analysis is quite mature. In recent years, there has been substantial interest in developing nonparametric and semiparametric regression methods for longitudinal data. Of importance are the varying-coefficient models. Since their introduction by Hastie and Tibshirani (1993), varying coefficient models have become an increasingly popular option for dimension reduction in nonparametric regression with multiple predictors. Due to their flexibility to explore the dynamic features which may exist in the data and their easy interpretation, the varying coefficient models have been widely applied in many scientific areas, such as in economics, psychology, sociology and many other fields of natural and social sciences. The theory and methodology have also experienced rapid developments; see Fan and Zhang (2008) for a comprehensive review of various statistical procedures proposed for the varying-coefficient models.

The varying-coefficient models are particularly appealing in longitudinal studies where they allow one to explore the extent to which covariates affect responses changing over time. Wu et al. (1998) used the local polynomial kernel method

* Corresponding author.
E-mail address: liu2008shu@126.com (S. Liu).

http://dx.doi.org/10.1016/j.jspi.2014.11.008

to obtain estimators by minimizing a local least-squares criterion and developed approximated pointwise and simultaneous confidence regions for the unknown coefficient functions. Lin and Ying (2001) studied a least-squares procedure that leads to a consistent estimator of the unknown coefficient functions. Huang et al. (2004) proposed a class of global estimation methods for the varying-coefficient models based on basis approximations. Xue and Zhu (2007) proposed a naive, mean-corrected and residual-adjusted empirical likelihood ratio test for the unknown coefficient functions. Tang and Cheng (2008) discussed M-estimation of the unknown coefficient functions by using B-spline series approximation. By linearizing the gSCAD penalty, Noh and Park (2010) proposed a one-step estimation that has the oracle property in variable selection. Yang et al. (2014) investigated the empirical likelihood inference of varying-coefficient errors-in-variables models. To name just a few.

An important feature of longitudinal data is the presence of serial correlation within repeated measurements on a given subject. Almost all papers mentioned above have not seriously investigated the issue of incorporating the correlation in estimation. It is well known that misspecification of the correlation may result in a great loss of efficiency. In addition, the correlation structure itself is of interest. Therefore, it is also essential to model the covariance structure. Since there are usually much more parameters in the covariance matrix and the positive definiteness of the covariance matrix has to be assured, modeling the correlation matrix is more challenging than modeling the mean. Pourahmadi (1999, 2000) proposed a modified Cholesky decomposition to decompose the covariance matrix, which is attractive due to the fact that it leads automatically to positive definite covariance matrices, and the parameters in it are related to well-founded statistical concepts. As a result, the regression techniques can be used to model the parameters in this decomposition and model based inference for the parameters in the mean and the covariances could be made. See Leng et al. (2010) for a detailed discussion.

For the merits of modified Cholesky decomposition, many researchers have used it to investigate various statistical models. For example, within the framework of generalized linear models, Ye and Pan (2006) proposed to use generalized estimating equations (GEE) to model the parameters in the decomposition by several sets of parametric estimating equations. To relax the parametric assumption, Pan et al. (2009) proposed a nonparametric local kernel weighted likelihood based approach to model the mean and covariance structures based on modified Cholesky decomposition. Since a fully nonparametric approach is hampered by some serious drawbacks, such as the curse of dimensionality, difficulty of interpretation, and lack of extrapolation capability, Leng et al. (2010) proposed a data-driven approach based on semiparametric regression models for the mean and the covariance simultaneously. More recently, Yao and Li (2013) developed a new estimation of nonparametric regression functions for clustered or longitudinal data based on Cholesky decomposition. Unlike Leng et al. (2010) using GEE method, they proposed to use the profile least squares techniques to estimate the correlation structure and regression function simultaneously. Motivated by these articles, in this paper, we investigate the statistical inference of varying-coefficient regression models for longitudinal data which do not require using the GEE methods to obtain the estimated variance when the longitudinal data are unbalanced like Yao and Li (2013) and allowing the regressive coefficients to depend on the time like Leng et al. (2010). Obviously, our approach is more flexible and appealing than Leng et al. (2010) and Yao and Li (2013) in both the mean and the covariance. It is noteworthy that many other efforts have been made to model the correlation structure of the longitudinal data, and several methods have been proposed. They include generalized estimating equations methods (Liang and Zeger, 1986), the nonparametric method (e.g. Li, 2011), the semiparametric method (e.g. Fan et al., 2007), the basis matrix method (e.g. Zhou and Qu, 2012), and so on.

The rest of paper is organized as follows. Section 2 describes the initial local linear estimator of the coefficient functions which ignores the within-subject correlation. Fitting generalized autoregressive coefficients and innovation variances is conducted in Section 3. Section 4 develops a new two-stage local polynomial estimation procedure to estimate the coefficient functions. In Section 5, we present Monte Carlo simulation results. We further illustrate the proposed procedure via analyzing a panel hormone study in Section 6. Section 7 ends the article with a discussion. Regularity conditions and technical proofs are presented in the Appendix.

2. Model setup

For n randomly selected subjects with each subject being repeatedly measured over time, the longitudinal sample of $\{Y(t), t, \mathbf{Z}(t)\}$ is denoted by $\{(Y_{ij}, t_{ij}, \mathbf{Z}_{ij});\ i = 1, \ldots, n,\ j = 1, \ldots, m_i\}$, where t_{ij} is the jth measurement time of the ith subject, $Y_{ij} = Y_i(t_{ij})$ and $\mathbf{Z}_{ij} = (Z_{1ij}(t_{ij}), \ldots, Z_{pij}(t_{ij}))^\top$ are the observed outcome and covariate vector, respectively, of the ith subject at time t_{ij}, and m_i is the ith subject number of repeated measurements. Here $^\top$ denotes the transpose of a vector or matrix. The total number of measurements for the study is $N = \sum_{i=1}^{n} m_i$. We also assume in our asymptotic study that m_i is bounded, but the number of subjects n goes to infinity. In contrast to the independent identically distributed (i.i.d.) sample, the measurement within each subject is possibly correlated, although the inter-subject measurements are assumed to be independent. Moreover, without loss of generality, we further assume that all t_{ij} are scaled into the interval $\mathcal{T} = [a, b]$.

A varying-coefficient model for such longitudinal sample has the form

$$Y_{ij} = \mathbf{Z}_{ij}^\top \boldsymbol{\alpha}(t_{ij}) + \varepsilon_i(t_{ij}), \quad i = 1, \ldots, n,\ j = 1, \ldots, m_i, \tag{2.1}$$

where $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_p(t))^\top$ is a vector of smoothing functions of t, and $\{\varepsilon_i(t),\ i = 1, \ldots, n\}$ are zero-mean stochastic processes independent of $\mathbf{Z}_{ij}$.