Journal of Statistical Planning and Inference 187 (2017) 78–102

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

GEE analysis for longitudinal single-index quantile regression

Weihua Zhao [a], Heng Lian [b,*], Hua Liang [c]

[a] School of Science, Nantong University, Nantong 226007, China

[b] Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong

[c] Department of Statistics, The George Washington University, Washington, D.C. 20052, USA

ARTICLE INFO

Article history:
Received 13 September 2016
Received in revised form 29 January 2017
Accepted 27 February 2017
Available online 9 March 2017

Keywords:
Asymptotic normality
B-splines
Estimating equations
Quantile regression
Single-index models

ABSTRACT

We consider a single-index quantile regression model for longitudinal data. Based on generalized estimating equations, an estimation procedure is proposed by taking into account the correlation within subject. Under mild assumptions, we derive the convergence rate of the estimator of the unknown link function and the asymptotic normality of estimator of the index parameter using the "projection" technique. Since the estimating equations are non-continuous, we further adopt the smoothing approach and show that estimators obtained from the smoothed estimating equations are asymptotically equivalent to that from the unsmoothed estimating equations. It is also shown that the estimator is more efficient when the correlation is correctly specified. Finally, we present numerical examples including simulations and analysis of a lung function data.

1. Introduction

Consider the longitudinal single-index model

$$Y_{ij} = g(\mathbf{X}_{ij}^T \boldsymbol{\beta}_0) + e_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, m_i. \tag{1}$$

Here Y_{ij} is the response for measurement j on subject i, $\mathbf{X}_{ij} \in R^p$ is the associated p-dimensional covariate, g is an unknown link function and $\boldsymbol{\beta}_0$ is the index parameter. The data are assumed to be independent for different subject i, but dependent within a subject. If the interest is on the conditional mean of the response, one assumes $E[e_{ij}|\mathbf{X}_i] = E[e_{ij}|\mathbf{X}_{ij}] = 0$, where $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i})$. We are however concerned with quantile estimates and thus we assume $P(e_{ij} \leq 0|\mathbf{X}_i) = P(e_{ij} \leq 0|\mathbf{X}_{ij}) = \tau, \tau \in (0, 1)$. Note the quantities $g, \boldsymbol{\beta}_0, e_{ij}$ implicitly depend on the quantile level τ we consider, but we omit these dependences in our notations for simplicity, as well as for some quantities in the rest of the paper.

Single-index models were proposed in Ichimura (1993) and Härdle et al. (1993) to incorporate link functions that relax the stringent linearity assumption in the parametric model while largely retaining its easy interpretability. It has received a lot of attention in both the statistics and the econometrics literature (Newey and Stoker, 1993; Newey, 1994; Yu and Ruppert, 2002; Antoniadis et al., 2004; Wang et al., 2010; Chen et al., 2013a,b). In the early literature, single-index models are usually estimated by the kernel method. Yu and Ruppert (2002) and Wang and Yang (2009) used penalized spline and regression spline for estimation, respectively. Carroll et al. (1997) considered generalized single-index model with an additional linear part, Xia et al. (1999) proposed an extended partially linear single-index model. Further variants and extensions of single-index models include Xia and Li (1999) and Fan et al. (2003).

* Corresponding author.
E-mail address: henglian@cityu.edu.hk (H. Lian).

http://dx.doi.org/10.1016/j.jspi.2017.02.011

Clustered data often arise in biological and biomedical research, where the measurements within the same cluster are correlated, while the measurements from different clusters are independent. For example, in longitudinal studies, the subjects are measured repeatedly over a given period of time. The measurements from the same subject are often correlated and thus form a cluster. A popular approach for clustered data analysis is generalized estimating equation (GEE) (Liang and Zeger, 1986), in which both the within-cluster and between-cluster variations are considered. A remarkable property of the GEE estimator is that it is consistent and asymptotically normal even with a misspecified covariance matrix. Furthermore, the GEE estimator is efficient when the covariance matrix is correctly specified. GEE and its extensions have been thoroughly studied for various parametric and semiparametric models and are broadly applied in diverse disciplines, see Diggle et al. (2002) for a comprehensive survey. Single-index models for longitudinal data are considered in Bai et al. (2009), Lai et al. (2013) and Chen et al. (2015).

The above works focused on mean regression, which only provides limited information about the distribution of the response and is known to be very sensitive to heavy-tailed error distributions. The parametric quantile regression introduced by Koenker and Bassett (1978) has been well developed in the econometrics and statistics literature. When the distribution of the errors in the model is heavy-tailed or the data contain some outliers, it is well known that median regression, a special case of quantile regression, is more robust than mean regression. More importantly, it can be used to obtain a large collection of conditional quantiles to characterize the entire conditional distribution. To construct a richer class of regression models and capture the relationships between the covariates and the response distribution, nonparametric quantile estimation has been studied in Hendricks and Koenker (1992) and Yu and Jones (1998). For varying coefficient models, Kim (2007) studied quantile regression for independent data using splines, and Cai and Xu (2008) used local polynomial estimation method for time series data. Further extensions to partially linear varying-coefficient models were considered by Wang et al. (2009) and Cai and Xiao (2012). Due to the reason that the index parameter is nested within the unknown link function, single-index models are generally harder to analyze theoretically than other semiparametric models such as varying-coefficient models or additive models. Wu et al. (2010), Kong and Xia (2012) and Ma and He (2016) are the first to analyze quantile single-index models.

In this paper, we are interested in quantile single-index models for longitudinal data (1) based on GEE. This model can capture nonlinear effects of the predictors in the conditional quantile, and allows to appropriately deal with within cluster correlations to improve estimation efficiency. Our estimation of the nonparametric link function is based on polynomial spline estimation, which is an alternative to local polynomial estimation method. The comparative advantages of spline methods were carefully documented in Li (2000), among which the most notable is the computational convenience, although it is not our main intention here to promote splines. The disadvantage is that the exact bias term for the nonparametric function is harder to obtain, making demonstration of asymptotic normality for the estimator of the nonparametric functions difficult. Theoretically, our study is nontrivial due to the combined complexity of dealing with single-index models with parameter nested within an unknown link functions, taking into account dependence using a working correlation matrix and discontinuous estimating equations. We make several contributions in this study. Firstly, we establish asymptotic normality for the estimator of the index parameter using an appropriately defined projection onto the space of single-index functions. Unlike the projection for independent and identically distributed case, the correct definition of projection here involves the working correlation matrix. Unlike semiparametric GEE analysis in Chen et al. (2015), empirical process theory is heavily used in our theoretical analysis due to the unsmooth nature of the estimating equations. Secondly, we explicitly show that the asymptotic variance of the parameters is the smallest when the correlation is correctly specified. To our knowledge, our work represents the first effort in demonstrating the advantage of taking dependence into account in GEE for semiparametric quantile regression. Thirdly, and most importantly, although theoretically satisfactory, the discontinuous estimating equations are hard to use in practice and not amenable to implementation directly. Thus we propose to smooth the indicator function similar to the kernel smoothing used in Horowitz (1998). An important theoretical contribution of our work is to establish that the estimators obtained from the smoothed estimating equations are asymptotically equivalent to that obtained from the original unsmoothed ones.

The rest of the article is organized as follows. In Section 2.1, we propose the estimating equations that jointly estimate the nonparametric and parametric components. The theoretical properties are established in Section 2.2. In Section 2.3 we then propose to smooth the GEE for computational convenience and show that the estimators attain the same asymptotic properties as before. Section 3 is devoted to numerical experiments including simulations and a real data analysis. Finally we conclude with possible future extensions in Section 4. The technical details are relegated to the Appendix.

2. Methodology

2.1. Estimation with generalized estimating equations

The data are assumed to be independent for different subject i, and we assume $m_i \leq m$ where m is a fixed number. For theoretical convenience we actually assume $m_i \equiv m$, although in practice there is not much extra difficulty in dealing with unbalanced data as long as we can estimate the working correlation matrix. In view of this, we still keep using the notation m_i as much as possible to allow for this generality. The theory can also be modified for the case m_i are i.i.d. from some distribution.