Journal of Multivariate Analysis 144 (2016) 176–188

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Quantile regression of longitudinal data with informative observation times

Xuerong Chen [a,*], Niansheng Tang [d], Yong Zhou [b,c]

[a] School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China
[b] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[c] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[d] Department of Statistics, Yunnan University, Kunming, Yunnan, China

ARTICLE INFO

Article history:
Received 15 April 2013
Available online 2 December 2015

AMS subject classifications:
62J99
62E20
62N99
62G35

Keywords:
Estimating equation
Informative observation times
Longitudinal data
Quantile regression
Resampling method

ABSTRACT

Longitudinal data are frequently encountered in medical follow-up studies and economic research. Conditional mean regression and conditional quantile regression are often used to fit longitudinal data. Many methods focused on the cases where the observation times are independent of the response variables or conditionally independent of them given the covariates. Few papers have considered the case where the response variables depend on the observation times or observation times are random variables associated with a counting process. In this paper, we propose a marginally conditional quantile regression approach for modeling longitudinal data with random observing times and informative observation times. Estimators of the parameters in the proposed conditional quantile regression are derived by constructing non-smooth estimating equations when the observation times follow a counting process. Consistency and asymptotic normality for these estimators are established. Asymptotic variance is estimated based on a resampling method. A simulation study is conducted and suggests that the finite sample performance of the proposed approach is very good, and an illustrative approach is provided.

1. Introduction

Longitudinal data arise frequently in many types of studies, for example, medical follow-up studies and observational investigations. In these longitudinal studies, observations from an individual are collected repeatedly over time. Various methods including generalized estimating equation and random effects model have been developed to analyze longitudinal data, see, [2,11]. Recently, non-parametric and semiparametric models for longitudinal data have attracted much attention. For nonparametric methods see [5,32,23,35,33,27]; for semiparametric approaches see [19,34,12]. One major difficulty in analyzing longitudinal data is that the observation times are often different across subjects. In [16–18,14], these authors considered time-varying coefficient regression models for longitudinal data, based on modeling the observation times by counting process. Under this framework, the observation times were allowed to have arbitrary pattern and to depend on covariates. These seminal work provided ways for modeling the time-dependent observations, survival data and recurrent event data in a unified framework.

Most existing methods assumed that the response variable is independent of observation times completely or conditionally independent given the covariates. This assumption may be unrealistic in applications. Informative observation times

* Corresponding author.
E-mail address: chenxr522@foxmail.com (X. Chen).

http://dx.doi.org/10.1016/j.jmva.2015.11.007
0047-259X/

often occur when they are subject or response variable-dependent. For example, consider the bladder cancer study conducted by Veterans Administration Cooperative Urological Research Group (see [26]). In the beginning of this study, all subjects who participated in the study had superficial bladder tumors and these tumors were removed. During the study, many patients suffered from multiple recurrences of tumor, and the recurrent tumors were removed during clinical visits. The clinical visit times and the number of tumors occurred between clinical visits were collected. One aim of this study was to compare the recurrence rates of the tumors of patients in different treatment groups. It is worth noting that, some subjects had significantly more clinical visits than others, which suggests that the number of clinical visits may contain some information about the tumor recurrence rates and the clinic visit times may depend on subject or covariate. It is important to make use of these information for inference on the recurrence rate of tumor. This motivated several authors to consider to incorporating the informative observation times in longitudinal data analysis. For example, Sun et al. [25] considered a semiparametric regression approach by using the estimating equation approach when the response variable depends on the observation times. They proposed a marginal model for the response variable process conditional on the covariates and the observation times, and the observation times were assumed to follow a counting process. Their model is a generalization of the marginal model proposed by Lin and Ying [14]. Almost all papers on longitudinal data whose observation times follow a counting process were conducted by conditional mean regression method. Besides the traditional conditional mean regression method, conditional quantile regression is another important approach used in longitudinal data analysis. When data contain some outliers or the error distribution is skewed or has heavy tails, the latter method is more robust and efficient than the former one.

Quantile regression method has been widely applied to the analysis of longitudinal data. He et al. [4] reviewed and compared three estimators of median regression in linear models for longitudinal data. Motivated by the penalized least squares for random effects models, Koenker [9] proposed a penalized quantile regression method when there were a large number of individual fixed effects that can significantly inflate the variability of the estimates of the main covariate effects. Karlsson [8] considered the nonlinear quantile regression model for longitudinal data. Fenske et al. [3] detected the risk factor for obesity in early childhood by using quantile regression methods for longitudinal data. Mu and Wei [20] studied the dynamic quantile regression transformation model for longitudinal data. Liu and Bottai [15] studied the mixed-effects models with longitudinal data by employing the quantile regression method. Wang and Fygenson [29] developed quantile regression inference procedures for longitudinal data when some of the measurements were censored by fixed constants. Wang et al. [31] developed a quantile estimation method for partially linear varying coefficient models using splines. Wang and Zhu [30] considered a quantile regression approach for longitudinal data by empirical likelihood method.

However, the existing literature of quantile regression for longitudinal data did not consider the case where observation times are informative. Furthermore, the observations times in the estimating methods are assumed to be independent of the covariates. To relax these limitations, in this paper, we study the quantile regression method for longitudinal data when the response variable depends on the observation times which depend on covariates by following a counting process. To make inference to parameters, estimating equations are constructed. The main difficulties are the Taylor expansion cannot be used to derive the asymptotic distribution of the estimators and Newton algorithm can no longer be used to compute the estimators, because the involved estimating equations are non-smooth. In this paper, the key results of empirical process theory, namely the uniform law of large number and the stochastic equicontinuity, are used to derive the asymptotic properties of estimators. This method has been used in the literature on non-smooth estimating equations, see for example, [22,1]. To overcome the computational difficulty, an iterative method based on MM algorithm of quantile regression (see [6]) is proposed. Because it is not easy to estimate the asymptotic variances of quantile regression estimators directly, these asymptotic variances are estimated by using the resampling method proposed by Jin et al. [7] in this paper.

This paper is organized as follows. In Section 2, we introduce some notations and describe the models we consider in this paper. In Section 3, the inference procedure and the MM algorithm based-iterative method are provided. The consistency, asymptotic normality of the proposed estimators and the asymptotic variance estimate are given in Section 4. Section 5 reports the simulation results and a real data example is given in Section 6. The conditions and the proofs are presented in the Appendix.

2. Notation and statistical models

Suppose that a longitudinal study consists of a random sample of n subjects. For subject i, let $Y_i(t)$ be the response variable and $X_i(t)$ be a p-dimensional vector of possibly time-dependent covariates, $i = 1, \ldots, n$. The observations of $Y_i(t)$ are taken at time points $t_{i1} < \cdots < t_{in_i}$, where n_i is the total number of observations on the ith subject. The number of observations of the ith subject by time t is $N_i(t) = \sum_{j=1}^{n_i} I(t_{ij} \le t) = N_i^*(\min(t, C_i))$, where C_i is the follow-up time or censoring time for the ith subject and $N_i^*(t)$ is the underlying counting process of sampling times for subject i. We assume that the covariate history $\{X_i(t) : 0 \le t \le C_i\}$ is observed for each individual.

For inference about the response process $Y_i(t)$, if it is completely or conditionally independent of $N_i^*(t)$, then the marginal approach is usually used [14]. Otherwise, as described in [25], there are three choices: modeling them jointly, modeling $Y_i(t)$ marginally and then $N_i^*(t)$ conditional on it, or modeling $N_i^*(t)$ marginally and then $Y_i(t)$ conditional on it. Our main interest is on the longitudinal process, rather than the observation times. The evaluation of the covariate effects on the longitudinal process is also of interest. Hence, we adopt the third choice, that is, we model $N_i^*(t)$ marginally and then model $Y_i(t)$ conditionally on it.