Computational Statistics and Data Analysis 117 (2018) 1–18

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Quantile regression for partially linear varying-coefficient model with censoring indicators missing at random

Yu Shen, Han-Ying Liang *

School of Mathematical Science, Tongji University, Shanghai 200092, PR China

ARTICLE INFO

Article history:
Received 21 February 2016
Received in revised form 29 April 2017
Accepted 20 July 2017
Available online 8 August 2017

MSC 2010:
62N01
62E20

Keywords:
Quantile regression
Partially linear varying-coefficient model
Variable selection
Random censorship
Missing at random

ABSTRACT

In this paper, we focus on the partially linear varying-coefficient quantile regression model when the data are right censored and the censoring indicator is missing at random. Based on the calibration and imputation methods, a three-stage approach is proposed to construct the estimators of the linear part and the nonparametric varying-coefficient function for this model . At the same time, we discuss the variable selection of the covariates in the linear part by adopting adaptive LASSO penalty. Under appropriate assumptions, the asymptotic normality of the proposed estimators is established, and the penalized estimators are proven to have the oracle property. Simulation study and a real data analysis are conducted to evaluate the performance of the proposed estimators.

1. Introduction

Flexible and refined statistical models are widely sought in a large array of statistical problems. Partially linear varying-coefficient (PLVC) model, as a combination of the partially linear model and varying-coefficient model, has drawn much attention of many researchers, and various methods have been proposed to estimate its parametric part and nonparametric function. See Zhang et al. (2002), Ahmad et al. (2005), Lam and Fan (2008), among others.

Most existing estimation methods for the PLVC model focus on mean regression by least squares or likelihood method. However they are sensitive to outliers and may be inefficient for many non-normal errors. The quantile regression (QR) introduced by Koenker and Bassett (1978) is more robust in exploring the underlying relationship between the covariates and the response. We refer to Koenker (2005) for a comprehensive overview of the QR. Some authors have discussed the PLVC QR model. Wang et al. (2009) used B-spline basis functions to define estimator of the QR with longitudinal data; Kai et al. (2011) proposed a novel three-stage estimation procedure and studied variable selection; Cai and Xiao (2012) discussed the QR in dynamic models with time series.

In many applications, especially in survival analysis and biomedical studies, the response cannot be completely observed due to possible right censoring, such as withdrawal of patients from the study, or death from a cause unrelated to the disease of being studied. Based on the linear model, the censored QR was first studied by Powell (1986) for the case of fixed censoring that assumes known censoring time for all observations. For the random right-censored model, Ying et al. (1995) proposed a semi-parametric procedure for median regression model, Portnoy (2003) developed a recursively reweighted estimation of the QR by the classical Kaplan–Meier estimator, Wang and Wang (2009) proposed a locally weighted censored

* Corresponding author.
E-mail address: hyliang@tongji.edu.cn (H.-Y. Liang).

http://dx.doi.org/10.1016/j.csda.2017.07.006
0167-9473/

QR approach that adopts the redistribution of mass idea and employs a local reweighting scheme, and Leng and Tong (2013) proposed a QR estimator through an unbiased estimating equation. Based on the varying-coefficient model, Yin et al. (2014) proposed a QR model and estimated regression parameters by generalized method of moments, and Xie et al. (2015) adopted a weighted inverse probability approach to develop the QR estimation. Furthermore, the censoring indicator may not be observed completely in practice. For instance, as described by Wang and Shen (2008), in clinical trials, individuals may fail from one of multiple causes, one of which is of interest. The time to death from the cause of interest may be censored by a death from a different cause. However, cause of death may sometimes be unavailable, for example: whether the death is attributable to the cause of interest may require information that is not collected or lost, or it may be difficult to determine the cause for some patients. In such cases, some censoring indicators are missing. In this paper, we assume that the censoring indicator is missing at random (MAR). The MAR assumption is common in statistical analysis involving missing data and is reasonable in many practical situations. See Little and Rubin (1987). Some authors have discussed the QR with random missing data. Yi and He (2009) investigated the linear QR model with missing response, and Sherwood et al. (2013) and Yang and Liu (2014) studied the linear QR model with missing covariates. For the varying-coefficient model with missing covariates, Sun and Sun (2015) studied the QR, Tang and Zhou (2015) studied the composite QR. For the right-censored data with missing censoring indicator, Subramanian (2004) and Wang and Ng (2008) suggested different estimators of the survival function, Song et al. (2010) and Qiu et al. (2015) discussed additive hazards regression model, and Wang and Dinse (2011) and Li and Wang (2012) proposed weighted least squared estimators for linear mean regression model. However there is no literature studying the PLVC QR model with censoring indicator MAR so far.

In addition, it is well known that variable selection for predictors has drawn the attention of many researchers. Various powerful penalization methods have been developed for the variable selection, such as LASSO (Tibshiranit, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006), and so on. Recently, the effective variable selection procedure has been developed for the PLVC model, for instance, under the complete data setting, Li and Liang (2008) proposed the nonconcave penalized quasi-likelihood method for variable selection in the PLVC mean regression model, and Kai et al. (2011) discussed the variable selection for the PLVC composite QR model. However there is little research for the variable selection of the PLVC QR model related to the right censored data with censoring indicator MAR.

In this paper, we focus on the PLVC QR model when the data are right censored and the censoring indicator is MAR. Using calibration and imputation methods, we suggest a three-stage approach to estimate both parameters of the linear part and the functions of the nonparametric varying-coefficient part in the model. Furthermore, when the covariates in the linear part are high-dimensional, we construct the penalized estimators by applying adaptive LASSO penalty, and study the variable selection of the model. A simulation study and a real data analysis are conducted to evaluate the performance of the proposed estimators.

The rest of the paper is organized as follows. In Section 2, a three-stage approach is proposed to construct the estimators in the PLVC QR model with the censoring indicator MAR, and the penalized estimators of the parameter in the linear part by applying adaptive LASSO penalty are also discussed. Main results are described in Section 3. Numerical study is presented in Section 4. The proofs of main results are given in Section 5.

2. Methodology

Consider the following partially linear varying-coefficient QR model

$$T_i = X_i^T\beta_\tau + Z_i^T\alpha_\tau(U_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where T_i is the response variable, $X_i \in \mathbb{R}^p$, $Z_i \in \mathbb{R}^q$ and $U_i \in \mathbb{R}$ are the explanatory variables, $\beta_\tau = (b_{\tau1}, b_{\tau2}, \ldots, b_{\tau p})^T$ is an unknown p-dimensional parameter vector, $\alpha_\tau(\cdot) = (\alpha_{\tau1}(\cdot), \alpha_{\tau2}(\cdot), \ldots, \alpha_{\tau q}(\cdot))^T$ is an unknown q-dimensional coefficient function vector, and ε_i is the random error whose τth quantile on $W_i = (X_i, Z_i, U_i)$ equals zero for $\tau \in (0, 1)$. Let $Q_{T_i}(\tau|W_i)$ be the condition quantile of T_i given W_i, then

$$Q_{T_i}(\tau|W_i) = \underset{a}{\operatorname{argmin}}\, E\big\{\rho_\tau(T_i - a)|W_i\big\},$$

where $\rho_\tau(t) = t[\tau - I(t < 0)]$ is called the check loss function. Model (2.1) indicates that

$$Q_{T_i}(\tau|W_i) = X_i^T\beta_\tau + Z_i^T\alpha_\tau(U_i). \tag{2.2}$$

In practice, T_i may be right censored due to some reasons. Let C_i be the censoring time with distribution function $G(\cdot)$, and $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \le C_i)$. Define a missing indicator ξ_i, which is 0 if δ_i is missing and is 1, otherwise. Then one can observe an i.i.d. sample $\{Y_i, W_i, \delta_i\xi_i, \xi_i, 1 \le i \le n\}$ from $(Y, W, \delta\xi, \xi)$. Suppose that T_i is independent of C_i, and that δ_i is MAR, which implies that ξ_i and δ_i are conditionally independent given (Y_i, W_i), i.e., $P(\xi_i = 1|Y_i, W_i, \delta_i) = P(\xi_i = 1|Y_i, W_i) := \Delta(Y_i, W_i)$.

Under the presence of censoring, since T_i is independent of C_i, we have

$$E\left\{\frac{\delta_i}{1 - G(Y_i)}\rho_\tau(Y_i - a)|W_i\right\} = E\big\{\rho_\tau(T_i - a)|W_i\big\}.$$