Journal of Statistical Planning and Inference 178 (2016) 112–127

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Quantile regression models for current status data

Fang-Shu Ou, Donglin Zeng *, Jianwen Cai

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States

ARTICLE INFO

Article history:
Received 5 December 2014
Received in revised form 2 June 2016
Accepted 7 June 2016
Available online 4 July 2016

Keywords:
Current status data
Concave–convex procedure
M-estimation
Quantile regression
Subsampling

ABSTRACT

Current status data arise frequently in demography, epidemiology, and econometrics where the exact failure time cannot be determined but is only known to have occurred before or after a known observation time. We propose a quantile regression model to analyze current status data, because it does not require distributional assumptions and the coefficients can be interpreted as direct regression effects on the distribution of failure time in the original time scale. Our model assumes that the conditional quantile of failure time is a linear function of covariates. We assume conditional independence between the failure time and observation time. An M-estimator is developed for parameter estimation which is computed using the concave–convex procedure and its confidence intervals are constructed using a subsampling method. Asymptotic properties for the estimator are derived and proven using modern empirical process theory. The small sample performance of the proposed method is demonstrated via simulation studies. Finally, we apply the proposed method to analyze data from the Mayo Clinic Study of Aging.

1. Introduction

Quantile regression (Koenker and Bassett, 1978) is a robust estimation method for regression models which offers a powerful and natural approach to examine how covariates influence the location, scale, and shape of a response distribution. Unlike linear regression analysis, which focuses on the relationship between the conditional mean of the response variable and explanatory variables, quantile regression specifies changes in the conditional quantile as a parametric function of the explanatory variables. It has been applied in a wide range of fields including ecology, biology, economics, finance, and public health (Cade and Noon, 2003; Koenker and Hallock, 2001). Quantile regression for censored data was first introduced by Powell (Powell, 1984, 1986), where the censored values for the dependent variable were assumed to be known for all observations (also known as the "Tobit" model). While this approach established an ingenious way to correct for censoring, the objective function was not convex over parameter values making global minimization difficult. Several methods have been proposed to mitigate related computational issues (Buchinsky and Hahn, 1998; Chernozhukov and Hong, 2002).

In most survival analysis, however, censoring time is not always observed. To accommodate a random censoring time, several methods were proposed over the past few decades. Early methods (Ying et al., 1995; Yang, 1999; Honore et al., 2002) required stringent assumptions on the censoring time, i.e. the censoring time must be independent of covariates. Under the conditional independence assumption where failure time and censoring time are independent conditional on covariates, Portnoy (2003) proposed a recursively reweighted estimator. Unfortunately, the quantile cannot be computed until the entire lower quantile regression process was computed first. The recursive scheme also complicated asymptotic inference. To overcome inferential difficulties, Peng and Huang (2008) and Peng (2012) developed a quantile regression

* Corresponding author.
E-mail addresses: ou.fang-shu@mayo.edu (F.-S. Ou), dzeng@email.unc.edu (D. Zeng), cai@bios.unc.edu (J. Cai).

http://dx.doi.org/10.1016/j.jspi.2016.06.001
0378-3758/

method for survival data subject to conditionally independent censoring and used a martingale-based procedure which made asymptotic inference more tractable. However, the method developed by Peng and Huang (2008) still has the same drawback as in Portnoy (2003), namely, the entire lower quantile regression process must be computed first. Huang (2010) developed a new concept of quantile calculus while allowing for zero-density intervals and discontinuities in a distribution. The grid-free estimation procedure introduced by Huang (2010) circumvented grid dependency as in Portnoy (2003) and Peng and Huang (2008). To avoid the necessity of assuming that all lower quantiles were linear, Wang and Wang (2009) proposed a locally weighted method. Their approach assumed linearity at one prespecified quantile level of interest and thus relaxed the assumption of Portnoy (2003); however, their method suffered the curse of dimensionality and hence can only handle a small number of covariates.

Current status data arise extensively in epidemiological studies and clinical trials, especially in large-scale longitudinal studies where the event of interest, such as disease contraction, is not observed exactly but is only known to happen before or after an examination time. Many likelihood-based methods have been developed for current status data, such as proportional hazards models, proportional odds models, and additive hazards models (see Sun, 2007 for a survey of different methods). Despite the fact that the development for censored quantile regression flourishes, the aforementioned methods were developed for right-censoring and are not suitable for current status data. To the best of our knowledge, the only method available for quantile regression models on interval-censored data was proposed by Kim et al. (2010) which was a generalization of the method proposed by McKeague et al. (2001). The proposed method can only be applied when the covariates took on a finite number of values since the method required estimation of the survival function conditional on covariates. The proposed method performed well in simulation studies, yet no theoretical justifications were offered. In this paper, we develop a new method for the conditional quantile regression model for current status data while allowing the censoring time to depend on the covariates.

The remaining paper is organized as follows. In Section 2, the proposed model is introduced and we establish estimation and inference procedures. Consistency and the asymptotic distribution are established in Section 3 with technical details deferred to Appendix. In Section 4, the small-sample performance is demonstrated via simulation studies and the application to data from the Mayo Clinic Study of Aging (MCSA) is given. Section 5 summarizes the method presented herein and avenues of further research.

2. The method

2.1. Model and data

Let T denote failure time and X a $k \times 1$ covariate vector with the first component set to one. We consider a quantile regression model for the failure time,

$$Q_T(\tau \mid X) = X'\beta(\tau), \quad \tau \in (0, 1), \tag{1}$$

where $Q_T(\tau \mid X)$ is the conditional quantile defined as $Q_T(\tau \mid X) = \inf\{t : \mathrm{pr}(T \le t \mid X) \ge \tau\}$ and the vector of unknown regression coefficients, $\beta(\tau)$, represents the covariate effects on the τth quantile of T which may depend on τ. Each element of $\beta(\tau)$ can be interpreted as an estimated difference in τth quantile by one unit change of the corresponding covariate while other variables in the model are held constant. Our interest lies in the estimation and inference on $\beta(\tau)$.

Let C denote the observation time and define $\delta \equiv I(T \le C)$ where $I(\cdot)$ is the indicator function. For current status data, T is not observed and the observed data consist of n independent replicates of (C, X, δ), denoted by $\{(C_i, X_i, \delta_i)_{i=1,\dots,n}\}$. It is assumed that T is conditionally independent of C given X. Since T is unobserved, we cannot directly estimate the conditional quantile function $Q_T(\tau \mid X)$ in Eq. (1) making a standard quantile regression unsuitable for our problem.

The τth conditional quantile of a random variable Y conditional on X can be characterized as the solution to the expected loss minimization problem,

$$Z(\beta) = E\{E[\rho_\tau(T - X'\beta(\tau)) \mid X]\}, \tag{2}$$

where $\rho_\tau(u) = u[\tau - I(u < 0)]$. Quantiles possess "equivariance to monotone transformations" (Koenker, 2005) which means that we may analyze a transformation $h(T)$ since the conditional quantile of $h(T)$ is $h(X'\beta(\tau))$ if $h(\cdot)$ is nondecreasing (Powell, 1994). In current status data, we observe realizations of the transformed variable $\delta \equiv I(T \le C)$ or, equivalently, $(1 - \delta) \equiv I(T > C)$ where the transformation is $h(T \mid C) = I(T > C)$ which is nondecreasing. We apply the same transformation to the conditional quantile, $X'\beta(\tau)$, and use the transformed conditional quantile, $I(X'\beta(\tau) > C)$, in the subsequent analysis. Since the objective function in (2) is well-defined and is sufficient to identify the parameters of interests (Powell, 1994), we can substitute $(1 - \delta)$ and $I(X'\beta > C)$ in Eq. (2) to get

$$Z(\beta) = E[E\{\rho_\tau[(1 - \delta) - I(X'\beta(\tau) > C)] \mid X, C\}]. \tag{3}$$

Eq. (3) may now be used to identify $\beta(\tau)$ since it contains only the observable variables (C, X, δ). We can show that the derivative of $Z(\beta)$ with respect to β is zero at the true β (see Appendix for details). Due to censoring, it is possible that not all $\beta(\tau)$ can be estimated using the observed data. We provide a sufficient condition to guarantee the identifiability for a fixed quantile in Section 3.1.