Published in final edited form as:

Biometrics. 2012 March ; 68(1): 101–112. doi:10.1111/j.1541-0420.2011.01667.x.

Quantile Regression for Doubly Censored Data

Shuang Ji¹, Limin Peng^{1,*}, Yu Cheng², and HuiChuan Lai³

¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322, U.S.A.

²Departments of Statistics and Psychiatry, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.

³Departments of Nutritional Sciences and Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.

SUMMARY

Double censoring often occurs in registry studies when left censoring is present in addition to right censoring. In this work, we propose a new analysis strategy for such doubly censored data by adopting a quantile regression model. We develop computationally simple estimation and inference procedures by appropriately using the embedded martingale structure. Asymptotic properties, including the uniform consistency and weak convergence, are established for the resulting estimators. Moreover, we propose conditional inference to address the special identifiability issues attached to the doubly censoring setting. We further show that the proposed method can be readily adapted to handle left truncation. Simulation studies demonstrate good finite-sample performance of the new inferential procedures. The practical utility of our method is illustrated by an analysis of the onset of the most commonly investigated respiratory infection, *Pseudomonas aeruginosa*, in children with cystic fibrosis through the use of the US Cystic Fibrosis Registry.

Keywords

Conditional inference; Double censoring; Empirical process; Martingale; Regression quantile; Truncation

1. Introduction

Double censoring often occurs in biomedical research when an outcome of interest is subject to both left censoring and right censoring. For example, in patients with cystic fibrosis (CF), the onset of *Pseudomonas aeruginosa* (PA) is monitored as an important landmark event for lung disease progression and was shown to be associated with reduced survival (Kerem et al., 1992, Emerson et al., 2002). However, in the US Cystic Fibrosis Foundation Patient Registry (CFFPR), which documents the diagnosis and annual follow-ups of all known CF patients and is described in detail elsewhere (FitzSimmons, 1993), ages at the first PA infection were not known for patients who had been infected with PA infection and thus were positive at the first documented CFFPR visit, and those who had no detected or reported PA infection by the end of follow-up. This poses a double censoring scenario, where, with the time origin set as birth, time to first PA infection was subject to left

*lpeng@sph.emory.edu.

Supplementary Materials

Web Appendices A, B and C referenced in Sections 2–4 are available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

censoring by time to registry entry, which was always recorded, and also right censoring by time to last follow-up visit, which was not always known in advance due to the occurrence of random dropout. Such a double censoring setting frequently arises in observational studies and is the focus of this paper. Note, it is different from the double censoring problem studied by Sun (2006), which concerned the gap time between two related events subject to interval censoring.

The simultaneous presence of left censoring and right censoring can bring many complexities to the analysis of doubly censored data. For example, the estimator of the survival distribution function in the one-sample case is generally presented as the solution of self-consistency equations (Turnbull, 1974, Chang and Yang, 1987), and cannot be expressed in closed form. Gehan (1965) studied the two-sample problem by an extension of the Wilcoxon test. Ren (2008) proposed a weighted empirical likelihood-based semiparametric maximum likelihood estimator as a unified approach for the two-sample problem with various censoring schemes including double censoring. Overall, these approaches are more complicated than their counterparts for randomly right censored data, for example, Kaplan-Meier estimator (Kaplan and Meier, 1958) or log-rank test (Mantel, 1966).

This paper is concerned with the general regression setting. Among existing work, Zhang and Li (1996) proposed a Buckley-James-Ritov-type M-estimator. Their estimating equation is neither monotone nor continuous and thus may necessitate special efforts to address some computational issues. More recently, Ren and Gu (1997) and Ren (2003) proposed a parallel regression M-estimator. This approach requires the independence between censoring variables and covariates, and thus imposes a stronger random censoring assumption than the usual one. For scenarios where both left and right censoring times are always observed, Cai and Cheng (2004) studied semiparametric transformation models (Cheng et al., 1995), and Yan et al. (2009) adapted temporal process regression (Fine et al., 2004) to doubly censored data.

The primary goal of this paper is to develop a computationally simple regression approach that can accommodate the double censoring scenario as occurs in CFFPR while only requiring the standard random censoring assumption (i.e. the left and right censoring times are independent of the event time of interest given covariates). To this end, we adopt quantile regression (Koenker and Bassett, 1978), a regression strategy that has received increased attention in survival analysis. With right censored data, among the earliest breakthroughs, Powell (1984, 1986) extended the least absolute deviation (LAD) approach from traditional quantile regression to censored quantile regression, assuming the censoring variables are fixed or always observable. Later efforts have been made to accommodate non-fixed censoring which is not always known by requiring additional restrictions such as unconditional independent censoring (Ying et al., 1995, Honore et al., 2002, among others), or nearly independently and identically distributed (i.i.d.) errors (Yang, 1999). More recently, without imposing these constraints, Portnoy (2003) proposed a recursively reweighted estimator as a generalization of the Kaplan-Meier estimator, with subsequent work by Neocleous et al. (2006) and Portnoy and Lin (2010) devoted to further polish the algorithm and the asymptotic theory. Peng and Huang (2008) proposed an alternative estimation approach by utilizing the martingale structure of randomly right censored data. Their approach is well justified in theory and also has a convenient implementation.

As elaborated in Portnoy (2003) and Peng and Huang (2008), the use of quantile regression in survival analysis offers straightforward interpretation on event times as well as extra model flexibility to accommodate varying covariate effects, thereby providing a valuable alternative to the popular Cox proportional hazards model (Cox, 1972). For example,

through analyzing the CFFPR dataset with quantile regression, one would be able to assess how covariates impact various quantiles of time to first PA infection without restricting their effects to be constant as contrast to the traditional Cox regression. Such an analysis may also help detect population inhomogeneous risk patterns of the onset of PA infection, for example, some covariate effects that vary between patients who have high susceptibility to PA infection versus those who are less prone to this complication.

To the best of our knowledge, there has been little work on developing quantile regression methods tailored to the double censoring setting considered in this paper. To tackle the problem, our basic strategy is to identify an appropriate martingale process embedded with the present data structure and construct stochastic integral estimating equations by employing the technique of Peng and Huang (2008). Such an effort yields estimation and inference procedures which can be readily implemented by existing statistical software. The details are presented in Section 2 along with asymptotic justifications. In Section 3, we address an unique issue with double censoring, which is, both the lower and higher tails of the event time quantile function may suffer from non-identifiability. Unlike in the right censoring case, one may not simply deal with this identifiability problem by restricting the quantile inference range. In the case of lack of identifiability in the lower tail, we propose a novel solution based on a conditional version of quantile regression and present the corresponding estimation and inference.

In the rest of the paper, we first note that the new quantile regression approach for doubly censored data can be slightly modified to handle truncated data. The adaptation in the presence of left truncation is briefly described in Section 4. We then report in Section 5 the results from simulation studies conducted to evaluate the finite-sample performance of our proposals. An application to the CFFPR data is presented in Section 6 to illustrate the practical utility of the proposed method. In Section 7, we conclude this paper with a few remarks.

2. Quantile Regression Procedure with Doubly Censored Data

2.1 Data and Model

Let T denote the event time of interest, L be the left censoring time, U be the right censoring time, and $\tilde{\mathbf{Z}}$ be the $p \times 1$ vector of recorded covariates. Define $\mathbf{Z} = (1, \tilde{\mathbf{Z}})^T$, $X = \max\{L, \min(T, U)\}$, $R(t) = I(L < t \leq X)$, and $N(t) = I(X \leq t, \delta = 1)$. Here δ is the censoring indicator defined as

$$\delta = \begin{cases} 1, & \text{if } L < T \leq U; \\ 2, & \text{if } T \leq L; \\ 3, & \text{if } T > U; \end{cases}$$

and $N(t)$ and $R(t)$ may be viewed as the counting process and the at-risk process associated with T . We assume that $(L, U) \perp T$ given $\mathbf{Z}$. The observed data consists of n i.i.d. replicates of $(X, \delta, \mathbf{Z}, L)$, denoted by $\{(X_i, \delta_i, \mathbf{Z}_i, L_i)\}_{i=1}^n$.

Define the conditional τ -th quantile of T given $\mathbf{Z}$ by $Q_T(\tau|\mathbf{Z}) = \inf\{t : F_T(t|\mathbf{Z}) \geq \tau\}$, where $F_T(t|\mathbf{Z}) = \Pr(T \leq t|\mathbf{Z})$. We consider the quantile regression model taking the form,

$$Q_T(\tau|\mathbf{Z}) = g\{\mathbf{Z}^T \boldsymbol{\beta}_0(\tau)\}, \quad \tau \in (0, 1), \quad (1)$$

where $g(\cdot)$ is a known monotone link function, and $\boldsymbol{\beta}_0(\tau)$ is a vector of unknown coefficients representing covariate effects on $Q_T(\tau|\mathbf{Z})$. As noted in Peng and Huang (2008), model (1) is

a strict extension of the accelerated failure time (AFT) model (Buckley and James, 1979, Prentice, 1978, Louis, 1981, Wei and Gail, 1983, Tsiatis, 1990, Ritov, 1990, Wei et al., 1990, among others). The formulation of τ -varying coefficients enables the accommodation of population inhomogeneous covariate effects.

2.2 Estimation Procedure

To estimate $\beta_0(\tau)$ in model (1), our basic idea is to determine an appropriate martingale process which allows us to construct an unbiased stochastic integral estimating equation by using Peng and Huang (2008)'s technique.

Following this line, we consider $M(t) = N(t) - \int_0^t R(u) d\Lambda_T(u|\mathbf{Z})$, where $N(t) = I(X \leq t, \delta = 1)$, $R(t) = I(L < t \leq X)$, and $\Lambda_T(t|\mathbf{Z})$ denotes the cumulative hazard function of T given $\mathbf{Z}$. Let $N_i(t)$, $R_i(t)$, $\Lambda_T(t|\mathbf{Z}_i)$, and $M_i(t)$ be sample analogues of $N(t)$, $R(t)$, $\Lambda_T(t|\mathbf{Z})$, and $M(t)$. Denote the filtration $\sigma\{N_i(u), R_i(u+), \mathbf{Z}_i : i = 1, \dots, n; 0 \leq u \leq t\}$ by $\mathcal{F}_t$. Provided $(L_i, U_i) \perp T_i$ given $\mathbf{Z}_i$,

$$E[dN_i(t)|\mathcal{F}_{t-}] = Pr[t \leq T_i < t+dt, R_i(t)=1|\mathcal{F}_{t-}] = R_i(t)Pr[t \leq T_i < t+dt | T_i \geq t, L_i < t, U_i \geq t, \mathbf{Z}_i] = R_i(t)d\Lambda_T(t|\mathbf{Z}_i).$$

This shows that $M_i(t)$ is a martingale (Fleming and Harrington 1991) and

$$E\{M_i(t)|\mathbf{Z}_i\} = 0, \forall t \geq 0. \quad (2)$$

Under model (1), which implies $\Lambda_T(g\{\mathbf{Z}_i^T \beta_0(\tau)\}|\mathbf{Z}_i) = -\log(1 - \tau)$, it follows that

$$M_i[g\{\mathbf{Z}_i^T \beta_0(\tau)\}] = N_i[g\{\mathbf{Z}_i^T \beta_0(\tau)\}] - \int_0^\tau I[L_i < g\{\mathbf{Z}_i^T \beta_0(v)\} \leq X_i] dH(v)$$

from a use of variable transformation within the integral, where $H(x) = -\log(1 - x)$. This equality and (2) naturally lead to an estimating equation for $\beta_0(\cdot)$ given by

$$n^{1/2} S_n(\beta, \tau) = 0, \quad (3)$$

where

$$S_n(\beta, \tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \left(N_i[g\{\mathbf{Z}_i^T \beta(\tau)\}] - \int_0^\tau I[L_i < g\{\mathbf{Z}_i^T \beta(v)\} \leq X_i] dH(v) \right).$$

It is easy to see that $E\{S_n(\beta_0, \tau)\} = 0$. With all $L_i = 0$, equation (3) reduces to Peng and Huang (2008)'s estimating equation for randomly right censored data.

The stochastic integral representation of $S_n(\beta, \tau)$ suggests a grid-based procedure to obtain an estimator of $\beta_0(\cdot)$ based on equation (3). Specifically, define the estimator $\hat{\beta}(\tau)$ as a right-continuous step function that jumps only on a prespecified grid, $\mathcal{G}_{L_n} = \{0 = \tau_0 < \tau_1 < \dots < \tau_{L_n} = \tau_U < 1\}$, where τ_U is a prespecified constant subject to certain theoretical constraint. Because the definition of $Q_T(\cdot|\mathbf{Z}_0)$ and model (1) imply $g\{\mathbf{Z}^T \beta_0(0)\} = 0$, we always set $g\{\mathbf{Z}^T \hat{\beta}(0)\} = 0$. We propose to obtain $\hat{\beta}(\tau_j)$ ($j = 1, \dots, L_n$) by sequentially solving the following equation for $\beta(\tau_j)$:

$$n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{Z}_i (N_i [g\{\mathbf{Z}^T \boldsymbol{\beta}(\tau_j)\}] - \sum_{k=0}^{j-1} I[L_i < g\{\mathbf{Z}^T \widehat{\boldsymbol{\beta}}(\tau_k)\} \leq X_i] \times \{H(\tau_{k+1}) - H(\tau_k)\}) = 0. \quad (4)$$

Note that equation (4) is a monotone estimating equation Fygenon and Ritov (1994), and the left hand side of (4) equals $2n^{-\frac{1}{2}}$ times the gradient of the following L_1 -type convex function,

$$\begin{aligned} l_j(\mathbf{h}) &= \sum_{i=1}^n |I(\delta_i \\ &= 1)g^{-1}(X_i) \\ &\quad - \mathbf{h}^T I(\delta_i \\ &= 1)\mathbf{Z}_i| \\ &\quad + |R^* - \mathbf{h}^T \sum_{l=1}^n \{ \\ &\quad - I(\delta_l = 1)\mathbf{Z}_l\}| \\ &\quad + |R^* - \mathbf{h}^T \sum_{r=1}^n (2\mathbf{Z}_r \sum_{k=0}^{j-1} I[L_r < g\{\mathbf{Z}_r^T \widehat{\boldsymbol{\beta}}(\tau_k)\} \leq X_r] \times \{H(\tau_{k+1}) \\ &\quad - H(\tau_k)\})|, \end{aligned} \quad (5)$$

where R^* is a sufficiently large positive number which is expected to bound

$|\mathbf{h}^T \sum_{l=1}^n \{-I(\delta_l = 1)\mathbf{Z}_l\}|$ and $|\mathbf{h}^T \sum_{r=1}^n (2\mathbf{Z}_r \sum_{k=0}^{j-1} I[L_r < g\{\mathbf{Z}_r^T \widehat{\boldsymbol{\beta}}(\tau_k)\} \leq X_r] \times \{H(\tau_{k+1}) - H(\tau_k)\})|$ from above for all $\mathbf{h}$'s in the compact parameter space for $\boldsymbol{\beta}_0(\tau)$. In our numerical studies, we set $R^* = 10^8$. As a result, $\widehat{\boldsymbol{\beta}}(\tau_j)$ can be alternatively obtained as the minimizer of $l_j(\mathbf{h})$. This L_1 -minimization problem can be readily solved, for example, by using the Barrodale-Roberts algorithm Barrodale and Roberts (1974) implemented in standard statistical software, such as *l1 fit()* function in S-PLUS and *rq()* function in R. The same computational technique has been used in non-quantile regression settings, for example, the rank inference for the accelerated failure time model (Jin et al., 2003).

2.3 Asymptotic Results

Asymptotic studies of the proposed estimator are facilitated by the stochastic integral representation of our estimating function. We introduce necessary notation before stating the regularity conditions and theorems.

Let $F_X(\cdot|\mathbf{Z})$ and $F_X(\cdot|\mathbf{Z})$ be the distribution function and survival function of X given $\mathbf{Z}$, respectively. Define $F_{X,\delta}(t|\mathbf{Z}) = \Pr(X \leq t, \delta = 1|\mathbf{Z})$ and $F_{X,L}(t|\mathbf{Z}) = \Pr(X \geq t, L \geq t|\mathbf{Z})$. Let $f_T(\cdot|\mathbf{Z})$, $f_X(\cdot|\mathbf{Z})$, $\tilde{f}_X(\cdot|\mathbf{Z})$, $\tilde{f}_{X,\delta}(\cdot|\mathbf{Z})$, $\tilde{f}_{X,L}(\cdot|\mathbf{Z})$ and $g'(\cdot)$ denote the first order derivatives of $F_T(\cdot|\mathbf{Z})$, $F_X(\cdot|\mathbf{Z})$, $F_X(\cdot|\mathbf{Z})$, $F_{X,\delta}(\cdot|\mathbf{Z})$, $F_{X,L}(\cdot|\mathbf{Z})$ and $g(\cdot)$, respectively. Define

$$\begin{aligned}
\mu(\mathbf{b}) &= E[\mathbf{Z}N\{g(\mathbf{Z}^T \mathbf{b})\}], \\
\mathbf{B}(\mathbf{b}) &= E[\mathbf{Z}^{\otimes 2} \tilde{f}_{X,\delta}\{g(\mathbf{Z}^T \mathbf{b})|\mathbf{Z}\}g'(\mathbf{Z}^T \mathbf{b})], \mathbf{v}_n(\mathbf{b}) = n^{-1} \sum_{i=1}^n \mathbf{Z}_i N_i\{g(\mathbf{Z}_i^T \mathbf{b})\} - \mu(\mathbf{b}), \tilde{\mu}(\mathbf{b}) = E[\mathbf{Z}I\{L < g(\mathbf{Z}^T \mathbf{b}) \leq X\}], \\
\mathbf{J}(\mathbf{b}) &= E[\mathbf{Z}^{\otimes 2} (\tilde{f}_X\{g(\mathbf{Z}^T \mathbf{b})|\mathbf{Z}\} \\
&\quad - \tilde{f}_{X,L}\{g(\mathbf{Z}^T \mathbf{b})|\mathbf{Z}\})g'(\mathbf{Z}^T \mathbf{b})], \\
\tilde{\mathbf{v}}_n(\mathbf{b}) &= n^{-1} \sum_{i=1}^n \mathbf{Z}_i I\{L_i < g(\mathbf{Z}_i^T \mathbf{b}) \leq X_i\} - \tilde{\mu}(\mathbf{b}),
\end{aligned}$$

and $\|\mathcal{G}_{L_n}\| = \max\{|\tau_j - \tau_{j-1}|, j = 1, \dots, L_n\}$.

The regularity conditions include:

- C1.** The covariate space $\mathbf{Z}$ is bounded, i.e., $\sup_i \|\mathbf{Z}_i\| < \infty$.
- C2.** (a) Each component of $E(\mathbf{Z}N[g(\mathbf{Z}^T \boldsymbol{\beta}_0(\tau))])$ is a Lipschitz function of τ ; (b) $\tilde{f}_{X,\delta}(t|z)$ and $f_X(t|z)$ are bounded above uniformly in t and z .
- C3.** (a) $\tilde{f}_{X,\delta}\{g(\mathbf{Z}^T \mathbf{b})|\mathbf{Z}\} > 0$ for all $\mathbf{b} \in \mathcal{B}(d_0)$, (b) $E(\mathbf{Z}^{\otimes 2}) > 0$; (c) each component of $\mathbf{J}(\mathbf{b})\mathbf{B}(\mathbf{b})^{-1}$ is uniformly bounded in $\mathbf{b} \in \mathcal{B}(d_0)$, where $\mathcal{B}(d_0)$ is a neighborhood containing $\{\boldsymbol{\beta}_0(\tau), \tau \in (0, \tau_U)\}$, defined as $\mathcal{B}(d_0) = \{\mathbf{b} \in \mathbb{R}^p : \inf_{\tau \in (0, \tau_U)} \|\boldsymbol{\mu}(\mathbf{b}) - \boldsymbol{\mu}(\boldsymbol{\beta}_0(\tau))\| \leq d_0\}$.
- C4.** $\inf_{\tau \in [v, \tau_U]} \text{eigmin} \mathbf{B}\{\boldsymbol{\beta}_0(\tau)\} > 0$ for any $v \in (0, \tau_U)$, where $\text{eigmin}(\cdot)$ denotes the minimal eigenvalue of a matrix.

We establish the uniform consistency and weak convergence of $\hat{\boldsymbol{\beta}}(\tau)$ stated in the following theorems.

THEOREM 1: Assuming conditions C1–C4 hold and $\lim_{n \rightarrow \infty} \|\mathcal{G}_{L_n}\| = 0$, then $\sup_{\tau \in [v, \tau_U]} \|\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\| \rightarrow_p 0$, where $0 < v < \tau_U$.

THEOREM 2: Assuming conditions C1–C4 hold and $\lim_{n \rightarrow \infty} n^{1/2} \|\mathcal{G}_{L_n}\| = 0$, then $n^{1/2} \{\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\}$ converges weakly to a Gaussian process for $\tau \in [v, \tau_U]$, where $0 < v < \tau_U$.

The regularity conditions and the proofs of Theorems 1–2 bear similarity with those in Peng and Huang (2008). Specifically, condition C1 requires covariates to be bounded and this is often met in practice. The smoothness of $\boldsymbol{\beta}_0(\cdot)$ and the boundedness of $\tilde{f}_{X,\delta}(t|z)$ and $f_X(t|z)$ assumed in condition C2 are also reasonable in practical settings. It is rather common in quantile regression literature to require positive densities and the positive definiteness of $E(\mathbf{Z}^{\otimes 2})$ as in C3(a) and C3(b). Given bounded density functions and covariates, coupled with the condition C4 commented later, C3(c) is satisfied with $\mathbf{b} = \boldsymbol{\beta}_0(\tau)$. By the continuity of $\mathbf{J}(\mathbf{b})\mathbf{B}(\mathbf{b})^{-1}$, it is thus natural to expect C3(c) holds in a neighborhood of $\boldsymbol{\beta}_0$, $\mathcal{B}(d_0)$.

We would like to point out that the regularity condition C4 is the crucial constraint that ensures the identifiability of $\{\boldsymbol{\beta}_0(\tau), \tau \in (0, \tau_U)\}$. In the simple one-sample case, this condition is equivalent to

$$f_T[g\{\beta_0(\tau)\}] \cdot \Pr[L < g\{\beta_0(\tau)\} \leq U] > 0, \forall \tau \in (v, \tau_U],$$

for any $v \in (0, \tau_U)$. Assuming $f_T(\cdot)$ is bounded away from 0, this condition further reduces to $\Pr\{L < Q_T(\tau) \leq U\} > 0, \forall \tau \in (v, \tau_U]$, implying $\tau_U \leq F_T(U^+)$ and $v \geq F_T(L^-)$. Here and hereafter, for a random variable Y , we use Y^+ and Y^- to denote the upper bound and the lower bound of its support respectively. This requirement concurs with Chang and Yang (1987)'s identifiability condition proposed for estimating the distribution function of doubly censored data in the one-sample case. Since $v > 0$ can be chosen arbitrarily, it implies $L^- \leq T^-$. In the general regression setting, however, C4 only renders implicit conditions on L and τ_U to guarantee the identifiability of $\beta_0(\cdot)$. In practice, an infeasible minimization of $l_j(\mathbf{h})$ with some $j < L_n$ may suggest the violation of the identifiability condition C4. In this case, one would need to reduce the τ -range for inference by resetting τ_U to a value less than τ_j .

Under regularity conditions C1–4, providing $E\{S_n(\beta_0, \tau)\} = 0$ and $\mu(\beta(0)) = 0$ shown in Section 2.1, we can establish the uniform consistency of the proposed estimator by straightforwardly adapting the arguments in Peng and Huang (2008, Appendices). The proof of Theorem 1 is thus omitted. We sketch the proof of Theorem 2 in Web Appendix A, which provides more concrete information on the asymptotic distribution of $n^{1/2}\{\hat{\beta}(\cdot) - \beta_0(\cdot)\}$.

2.4 Inference

For inference on $\beta_0(\cdot)$, we propose resampling-based approaches, given the complexity in the asymptotic distribution of $\hat{\beta}(\cdot)$ shown in the proof of Theorem 2. The resampling-based inference procedures stated above can be justified by following the lines of Peng and Huang (2008), and thus are omitted here.

More specifically, we perturb the objective function (5) by $\xi_1, \dots, \xi_n$, a set of i.i.d. variates from a nonnegative known distribution with mean 1 and variance 1, for example, $Exp(1)$. The resulting objective function is

$$\begin{aligned} \tilde{l}_j(\mathbf{h}) &= \sum_{i=1}^n |\xi_i I(\delta_i) \\ &= 1) g^{-1}(X_i) \\ &\quad - \mathbf{h}^T \xi_i I(\delta_i) \\ &= 1) Z_i| \\ &\quad + |R^* - \mathbf{h}^T \sum_{l=1}^n \{ \\ &\quad - \xi_l I(\delta_l = 1) Z_l \}| \\ &\quad + |R^* - \mathbf{h}^T \sum_{r=1}^n (2\xi_r Z_r \sum_{k=1}^{j-1} I[L_r \leq g\{Z_r^T \beta^*(\tau_k)\} \leq X_r] \times \{H(\tau_{k+1}) - H(\tau_k)\})| \end{aligned} \quad (6)$$

for $j = 1, \dots, L_n$, where $\beta^*(\tau_j)$ is defined as the minimizer of $\tilde{l}_j(\mathbf{h})$ and can be obtained sequentially using the same procedure as that taken to compute $\hat{\beta}(\tau_j)$. For a fixed τ_* , we can approximate the variance of $\hat{\beta}(\tau_*)$ by repeatedly generating the variates set $\{\xi_1, \dots, \xi_n\}$ for B times and calculating the variance of the resulting $\{\beta_k^*(\tau_*)\}_{k=1}^B$. A 95% confidence interval for $\beta_0(\tau_*)$ can be constructed based on a normal approximation. Along the lines of Peng and Huang (2007) and Huang and Peng (2009), a 95% confidence band for $\{\beta_0(\cdot) : \tau \in [l, u]\}$

may be given by $\{\hat{\beta}(\tau) \pm \rho_{0.95} \hat{\sigma}(\tau), \tau \in [l, u]\}$, where $\rho_{0.95}$ is the 95% empirical percentile of $\sup_{\tau \in [l, u]} |\hat{\beta}^*(\tau) - \hat{\beta}(\tau)| / \hat{\sigma}(\tau)$ with $\hat{\sigma}(\tau)$ being the empirical standard deviation of $\hat{\beta}^*(\tau)$.

Second-stage inference can also be conducted in a similar fashion to that of Peng and Huang (2008). First, we consider the general hypothesis $H_0 : \psi\{\beta_0(\tau)\} = r_0(\tau), \tau \in [l, u]$, where $\psi(\cdot)$ is a known function and $r_0(\tau)$ is a hypothesized value for $\psi\{\beta_0(\tau)\}$. Let $\psi(\mathbf{x}) = \mathbf{x}^{(q)}$ and $r_0(\tau) = 0$, where $u^{(l)}$ denotes the l -th component of vector $\mathbf{u}$ and $2 \leq q \leq p + 1$. Testing H_0 is equivalent to assess whether the effect of $Z^{(q)}$ is significant for $\tau \in [l, u]$. One natural test may take the form, $\Gamma = n^{1/2} \int_l^u \{\psi\{\widehat{\beta}(v)\} - r_0(v)\} \Theta(v) dv$, where $\Theta(\cdot)$ is a nonnegative weight function. The distribution of Γ may be approximated by the empirical distribution of $\Gamma^* = n^{1/2} \int_l^u \{\psi\{\beta^*(v)\} - \psi\{\widehat{\beta}(v)\}\} \Theta(v) dv$ given the observed data.

Another second-stage hypothesis of interest is given by $\tilde{H}_0 : \tilde{\psi}\{\beta_0(\tau)\} = \eta_0, \tau \in [l, u]$, where $\tilde{\psi}(\cdot)$ is a known function and η_0 is an unspecified constant. With $\psi(\mathbf{x}) = \mathbf{x}^{(q)}$, H_0 depicts the scenario where the effect of $Z^{(q)}$ is constant over $\tau \in [l, u]$. To test $\tilde{H}_0$, one may adopt the test statistic $\tilde{\Gamma} = n^{1/2} \int_l^u \{\tilde{\psi}\{\widehat{\beta}(v)\} - \tilde{\rho}\} \tilde{\Theta}(v) dv$, where $\tilde{\Theta}(\cdot)$ is a nonconstant weight function and $\tilde{\rho} = \int_l^u \tilde{\psi}\{\widehat{\beta}(v)\} dv / (u - l)$. One may reject $\tilde{H}_0$ when $\tilde{\Gamma}$ is greater than the $(1 - \alpha/2)$ th quantile or less than the $(\alpha/2)$ th empirical quantile of $\tilde{\Gamma}^* = n^{1/2} \int_l^u \{\tilde{\psi}\{\beta^*(v)\} - \tilde{\psi}\{\widehat{\beta}(v)\}\} - (\rho^* - \tilde{\rho}) \tilde{\Theta}(v) dv$. As a useful by-product of the hypothesis testing for H_0 , $\hat{\rho}$ in the special case with $\psi(\mathbf{x}) = \mathbf{x}^{(q)}$ and $\tilde{\Theta}(v) = 1$, denoted by $\hat{\eta}$, provides a consistent estimate for $\eta_0 = \int_l^u \beta_0^{(q)}(v) dv$, the average quantile effect of $Z^{(q)}$ across $\tau \in [l, u]$. The inference on η_0 can be made similarly upon the resampling quantity $\eta^* = \int_l^u \{\beta^{*(q)}(v) - \widehat{\beta}^{(q)}(v)\} dv$.

3. A Conditional Version of Quantile Regression

As pointed out in Section 2.3, certain conditions are required for the identifiability of $\beta_0(\tau)$ and may not be satisfied in some real datasets, for example, when $L^- > T^-$. It is of practical interest to propose some remedies when data fail to identify some part of $\beta_0(\cdot)$, most likely to be $\beta_0(\tau)$ with small or large τ . In the presence of only random right censoring, one practical solution is to adaptively imposing an upper bound, τ_U , on the τ -range in which $\beta_0(\tau)$ is estimated (Peng and Huang, 2008). This action has little impact on the estimation because the estimating equations for $\{\beta_0(\tau), 0 < \tau < \tau_U\}$ stand alone without involving estimates for $\beta_0(\tau)$ with $\tau > \tau_U$.

Dealing with the identifiability issue in the double censoring case is more challenging because non-identifiability can occur on both tails of regression quantiles. One notable difficulty in view of equation (3) is that the sequential procedure presented in Section 2.2 would always require estimates at the lower tail, which however may not be identifiable due to left censoring.

When the identifiability of the lower tail of $\beta_0(\cdot)$ is precluded by doubly censored data, we propose a conditional version of model (1), which takes the form,

$$Q_T(\tau | Z, T > t_0) = g\{Z^T \alpha_0(\tau)\} + t_0, \quad \tau \in (0, 1), \quad (7)$$

where $Q_T(\tau | Z, T > t_0) = \inf\{t \geq t_0 : \Pr(T \leq t | Z, T > t_0) \geq \tau\}$, and $t_0 > 0$ is a prespecified constant subject to certain theoretical and practical constraints. A more detailed discussion on how to choose t_0 is relegated to the end of this section. In model (7), the unknown coefficients in $\alpha_0(\tau)$ represent the effects of covariates on the τ -th conditional quantile of T provided $T > t_0$.

The reasoning for adopting this conditional version of quantile regression is similar to that for estimating a conditional survival function of left truncated and right censored data when the unconditional one is not identifiable (Tsai et al., 1987). Model (7) essentially imposes a lower bound of t_0 for regression quantiles. As elaborated later, doing so helps circumvent the difficulty associated with non-identifiable lower tail of $\beta_0(\tau)$ at the cost of estimating a quantity which may slightly deviate from the primary interest.

Model (7) necessitates a different estimation procedure from that of the unconditional model (2). Estimating equation (3) can not be directly borrowed without modification. The critical step in the adaptation of equation (3) is to identify an analogue of $M(t)$ when model (7) is assumed instead of model (2). Along this line, we propose a natural substitute of $M(t)$, given by $\check{M}(t) = \check{N}(t) - \int_0^t \check{R}(u) d\Lambda_T(u | \mathbf{Z}, T > t_0)$, where $\check{N}(t) = I(t_0 < X \leq t, \delta = 1)$, $\check{R}(t) = I(t_0 \vee L < t \leq X)$, and $\Lambda_T(u | \mathbf{Z}, T > t_0)$ is the cumulative hazard function of T conditional on $\mathbf{Z}$ and $T > t_0$, namely, $-\log \Pr(T > t | \mathbf{Z}, T > t_0)$. Here $\vee$ is the maximum operator. Note that $\check{M}(t)$ resembles the standard martingale for T truncated by $t_0 \vee L$ except for the conditional hazard function involved. The use of the conditional hazard $d\Lambda_T(u | \mathbf{Z}, T > t_0)$ in place of $d\Lambda_T(u | \mathbf{Z})$ in $\check{M}(t)$ is in tune with the assumed conditional model (7).

There are two key facts that need to be verified before we adapt the estimation procedure in Section 2. First, we need to show $E\{\check{M}(t) | \mathbf{Z}\} = 0$ for $\forall t \geq t_0$. To this end, standard arguments based on martingale however may not be directly applicable, and we instead prove this by examining the connection between $M(t)$ and $\check{M}(t)$. A detailed proof is provided in Web Appendix B. Secondly, we need to have $g\{\mathbf{Z}^T \mathbf{a}_0(0)\} = 0$ as the boundary condition. Like in the unconditional case, this easily follows from the definition of $Q_T(\tau | \mathbf{Z}, T > t_0)$, and model assumption (7).

The estimating equation, motivated by $E\{\check{M}(g\{\mathbf{Z}^T \mathbf{a}(\tau)\} + t_0) | \mathbf{Z}\} = 0$, is then given by

$$n^{1/2} \check{S}_n(\alpha, \tau) = 0, \quad (8)$$

where

$$\check{S}_n(\alpha, \tau) = \sum_{i=1}^n \mathbf{Z}_i \left(\check{N}_i[g\{\mathbf{Z}_i^T \alpha(\tau)\} + t_0] - \int_0^\tau \check{R}_i[g\{\mathbf{Z}_i^T \alpha(v)\} + t_0] dH(v) \right),$$

where $\check{N}_i(\cdot)$ and $\check{R}_i(\cdot)$ are sample analogues of $\check{N}(\cdot)$ and $\check{R}(\cdot)$.

An estimator of $\mathbf{a}_0(\tau)$, denoted by $\hat{\mathbf{a}}(\cdot)$, can be easily obtained based on equation (8) by slightly modifying the algorithm presented in Section 2.2. The key is to note that $\check{S}_n(\mathbf{a}, \tau)$ can be rewritten as

$$\check{S}_n(\alpha, \tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i^* \left\{ N_i^*[g\{\mathbf{Z}_i^{*T} \alpha(\tau)\}] - \int_0^\tau I(L_i^* < g\{\mathbf{Z}_i^{*T} \alpha(v)\} \leq X_i^*) dH(v) \right\},$$

where $\mathbf{Z}_i^* = \mathbf{Z}_i I(X_i > t_0)$, $X_i^* = X_i - t_0$, $L_i^* = L_i - t_0$, and $N_i^*(t) = I(X_i^* \leq t, \delta = 1)$. This shows that we can simply replace X_i , L_i , $\mathbf{Z}_i$ in the proposed estimation procedure for $\beta_0(\tau)$ by X_i^* , L_i^* , and $\mathbf{Z}_i^*$ respectively to compute $\hat{\mathbf{a}}(\tau)$. The same strategy can be used to adapt the resampling-based procedures in Section 2.4 to make inference on $\mathbf{a}_0(\tau)$.

The analogy in estimating equation also suggests the similarity in asymptotic studies between the unconditional and conditional cases. Specifically, under regularity conditions $C1' - C4'$, which are $C1 - C4$ defined based on transformed data $(X^*, \delta, \mathbf{Z}^*, L^*)$, we have the following theorems:

THEOREM 3: Assuming conditions $C1' - C4'$ hold and $\lim_{n \rightarrow \infty} \|\mathcal{G}_{L_n}\| = 0$, then

$$\sup_{\tau \in [v, \tau_U]} \|\hat{\alpha}(\tau) - \alpha_0(\tau)\| \rightarrow_p 0,$$

where $0 < v < \tau_U$.

THEOREM 4: Assuming conditions $C1' - C4'$ hold and $\lim_{n \rightarrow \infty} \|n^{1/2} \mathcal{G}_{L_n}\| = 0$, then $n^{1/2} \{\hat{\alpha}(\tau) - \alpha_0(\tau)\}$ converges weakly to a Gaussian process for $\tau \in [v, \tau_U]$, where $0 < v < \tau_U$.

Similar to the unconditional case, the regularity condition $C4'$ is concerned with the identifiability of $\{\alpha_0(\tau) : \tau \in (0, \tau_U)\}$ and implicitly impose the theoretical requirement on t_0 and τ_U . In the one-sample case, it becomes

$$f_T[g\{\beta_0(\tau)\} + t_0] \cdot \Pr[L < g\{\beta_0(\tau)\} + t_0 \leq U] > 0, \forall \tau \in (v, \tau_U)$$

for $v \in (0, \tau_U)$. It is easy to see that, with t_0 chosen to be greater than L^- , $L < g\{\beta_0(\tau)\} + t_0$ would hold with a positive probability and thus the above condition would only impose constraints on τ_U . This finding in the one-sample case is suggestive of the diminished identifiability issue with the lower tail of $\alpha_0(\cdot)$ when the conditional version of quantile regression model is adopted.

In practice, to perform the proposed conditional quantile regression, one may choose t_0 as a constant that is greater than the observed lower bound of L and also produces converged estimates for model (7) with τ close to 0. The final selection of t_0 may be further adjusted according to the scientific interpretation of $Q_T(\tau | \mathbf{Z}, T > t_0)$, thereby yielding more meaningful conditional inference.

4. Extension to Handle Left Truncation

In this section, we present an extension to scenarios where left truncation is present, as frequently occurs in observational studies. Following the notation in Section 2, when X is subject to left truncation by event time A , the observed data include n i.i.d. replicates of $(X', L', A', \delta', \mathbf{Z})$, denoted by $\{(X'_i, L'_i, A'_i, \delta'_i, \mathbf{Z}_i)\}_{i=1}^n$, where $\{X', L', A', \delta', \mathbf{Z}'\}$ follows the conditional distribution of $\{X, L, A, \delta, \mathbf{Z}\}$ given $X \geq A$. It is assumed that (L, U, A) is independent of T given $\mathbf{Z}$. Such data can be referred to as doubly censored data with left truncation. With all $L_i = 0$, the data reduce to the usual left truncated right censored data.

Adopting the same idea for doubly censored data studied in Section 2, we construct an estimating equation for model (1) by utilizing the martingale structure associated with the observed truncated data described above. In the current setting, we define $N'(t) = I(X' \leq t, \delta' = 1)$, and the at-risk process as $R'(t) = I(L' \vee A' < t \leq X')$. We show in Web Appendix C that $M'(t) = N'(t) - \int_0^t R'(s) d\Lambda_\tau(s)$ is a martingale process. By this fact, only minor changes to equation (3) may be needed in order to accommodate the presence of left truncation. The proposed estimating equation for $\beta_0(\cdot)$ is given by

$$n^{1/2} \mathbf{K}_n(\beta, \tau) = 0, \quad (9)$$

where

$$\mathbf{K}_n(\beta, \tau) = n^{-1} \sum_{i=1}^n \mathbf{Z}_i \left(\dot{N}_i' [g\{\mathbf{Z}_i^T \beta(\tau)\}] - \int_0^\tau I[L_i' \vee A_i' < g\{\mathbf{Z}_i^T \beta(v)\}] \leq X_i' dH(v) \right).$$

The estimation and inference procedures can be developed based on equation (9) similarly to those described in Section 2. In the view of the analogy between equation (9) and equation (3), we can establish Theorems 1–2 with $(X, \delta, \mathbf{Z}, L)$ in C1–C4 replaced by $(X', \delta', \mathbf{Z}', L' \vee A')$ by adopting essentially the same arguments.

Following the same reasoning, one may estimate $\mathbf{a}_0(\cdot)$ in model (7) by solving the equation,

$$n^{1/2} \check{\mathbf{K}}_n(\mathbf{a}, \tau) = 0, \text{ with } \check{\mathbf{K}}_n(\mathbf{a}, \tau) = \sum_{i=1}^n \mathbf{Z}_i \left(\dot{N}_i' [g\{\mathbf{Z}_i^T \mathbf{a}(\tau)\} + t_0] - \int_0^\tau \dot{R}_i' [g\{\mathbf{Z}_i^T \mathbf{a}(v)\} + t_0] dH(v) \right),$$

where $\dot{N}_i'(t) = I(t_0 < X_i' \leq t, \delta' = 1)$, and $\dot{R}_i'(t) = I(L_i' \vee A_i' \vee t_0 < t \leq X')$. Theorem 3–4 can be generalized to the resulting estimator of $\mathbf{a}_0(\tau)$.

5. Simulation Studies

We studied the finite-sample performance of the proposed methods through Monte-Carlo simulations. We first generated event times from an AFT model with i.i.d. errors:

$$\log T = b_1 Z_1 + b_2 Z_2 + \varepsilon,$$

where ε followed the extreme value distribution. The covariates Z_1 and Z_2 were generated from *Unif*(0, 1) and *Bernoulli*(0.5) respectively. We obtained the right censoring time U and left censoring time L by generating U and L respectively from *Unif*(0.1 · $I(Z_2 = 1)$, c_u) and *Unif*(0, c_l) · W until $L \leq U$ with Z_2 fixed, where W was a *Bernoulli*(1 – p_0) variate. It is easy to show that under this set-up, both model (1) and model (7) hold with $g(\cdot) = \exp(\cdot)$, $\beta_0(\tau) = \mathbf{a}_0(\tau) = \{Q_\varepsilon(\tau), b_1, b_2\}^T$. By setting $p_0 = 0.2$, L had a probability mass of 0.2 at zero and rendered a scenario that the lower tail of $\beta_0(\cdot)$ was identifiable. In this case, the model (1) was considered. Choosing $b_1 = 0$, $b_2 = -0.5$, $c_l = 0.5$, $c_u = 3.8$ resulted in 20% right censoring and 20% left censoring. With $p_0 = 0$, we studied model (7) with $t_0 = 0.16$. We set $p_0 = 0$, $b_1 = 0$, $b_2 = -1.0$, $c_l = 0.3$, $c_u = 4.5$. The resulting proportions of right censoring and left censoring are 15% and 20% respectively. Under each configuration, we generated 1000 data sets of sample size $n = 200$. We set $B = 200$ in the resampling procedures with $\{\zeta_i\}_{i=1}^B$ generated from *Exp*(1). An equally spaced grid on $\tau \in (0, 1)$ with size 0.01 was adopted when estimating $\beta_0(\cdot)$ or $\mathbf{a}_0(\cdot)$. We also carried out tests on the overall significance and the constant effect hypotheses for each covariate. In the latter test we adopted the weight function $\Theta(v) = I\{v \geq (l + u)/2\}$. We set $l = 0.1$ and $u = 0.7$.

Table 1 presents the results from estimating model (1) and model (7) in the AFT model setting. We report absolute values of biases (Bias), empirical standard deviations (EmpSD), and average estimated resampling-based standard deviations (AvgSD) of $\hat{\beta}(\tau)$ and $\hat{\mathbf{a}}(\tau)$, and coverage rates of 95% Wald confidence intervals of $\beta_0(\tau)$ and $\mathbf{a}_0(\tau)$ with $\tau = 0.1, 0.3, 0.5$ and 0.7. It is observed from Table 1 that in either unconditional or conditional case, biases are small, the resampling-based standard deviation estimates agree well with the empirical ones, and the coverage rates are in general close to the nominal level.

Table 2 presents the hypothesis testing results. The empirical rejection rates (ERR) for both tests at level 0.05 are reported, together with the estimated average effects (AvgEst), empirical standard deviations of the average effects, and average resampling-based standard deviation estimates of the average effects. We see that the type I errors are close to the nominal level 0.05. The estimated average covariate effects of Z_1 and Z_2 are close to the true values. The resampling-based standard deviation estimates for the average covariate effect estimates agree well with the empirical standard deviations.

In addition to the AFT setting, we considered a log linear model with heteroscedastic errors. That is, event times were generated from the model:

$$\log T = b_1 Z_1 + b_2 Z_2 \xi + \varepsilon,$$

where Z_1 followed $Unif(0, 1)$, Z_2 followed $Bernoulli(0.5)$, and ξ followed $Exp(1)$. The right censoring time U and left censoring time L were generated in the same way as in the AFT setting. We considered two different configurations: (a) ε was a $N(0, 1)$ variate, $p_0 = 0.2$, $b_1 = 0$, $b_2 = -1.5$, $c_l = 0.7$, $c_u = 4.8$, (b) ε was an extreme value variate, $p_0 = 0$, $t_0 = 0.03$, $b_1 = 0$, $b_2 = -4.5$, $c_l = 0.05$, $c_u = 4.0$. One can verify that model (1) holds under configuration (a) with $g(\cdot) = \exp(\cdot)$ and $\beta_0(\tau) = \{\beta_0(\tau), \beta_1(\tau), \beta_2(\tau)\}^T$, where $\beta_0(\tau) = Q_\varepsilon(\tau)$, $\beta_1(\tau) = 0$, and $\beta_2(\tau) = Q_{b_2 \xi + \varepsilon}(\tau) - \beta_0(\tau)$. Under configuration (b), model (7) was satisfied with $g(\cdot) = \exp(\cdot)$, $\alpha_0(\tau) = Q_\varepsilon(\tau) = \log\{-\log(1 - \tau)\}$, $\alpha_1(\tau) = 0$, and $\alpha_2(\tau) = \log[Q_{\exp(b_2 \xi + \varepsilon)}(1 - \Pr\{\exp(b_2 \xi + \varepsilon) > t_0\} (1 - \tau)) - t_0] - \alpha_0(\tau)$. Here, for a random variable Y , $Q_Y(\tau)$ denotes its τ -th quantile. Note that unlike in the AFT settings, the effects of Z_2 were not constant and took a complicated analytic form. We approximated the true coefficients for Z_2 by using bootstrapping. In (a), there were 25% right censoring and 20% left censoring. In (b), the rates of right censoring and left censoring were 15% and 25% respectively.

Tables 3–4 present estimation results and hypothesis testing results as Tables 1–2. We can see that in the presence of varying covariate effects, our proposed method also performs well. In both unconditional and conditional cases, the regression quantile estimates are virtually unbiased with standard deviations accurately estimated by the proposed resampling method. The estimates for average effects also have small biases with estimated standard deviations agreeing well with empirical ones. The proposed test for the overall significance and the test for the constancy of covariate effect appear to have the right sizes and also reasonable power.

We conducted simulations with a larger sample size, $n = 400$. A comparison of the results corresponding to $n = 200$ and $n = 400$ (Web Appendix C) suggests that the performance of the proposed estimators and tests further improves as sample size increases.

Simulations were also performed to compare our approach with a naive approach that simply discards all left-censored subjects. Data were generated from the same configurations as the unconditional cases in Table 1 and Table 2. We evaluated the estimation of model (1). Figure 1 displays the mean estimated coefficients from the proposed approach and those from the naive approach along with the true coefficients. This figure shows that the proposed estimators are virtually unbiased while the naive approach can produce substantial biases particularly in the estimation of non-zero coefficients. Different bias patterns with the naive method are observed in some other set-ups (with results not shown), for example, a large deviation from zero of the naive estimates for the zero-coefficients. These findings show that it is important to account for double censoring arising from practical situations.

We also conducted simulations for cases with left truncation. Results reported in Web Appendix E show satisfactory performance of our proposals in Section 4.

6. CFFPR Data Example

We apply the proposed quantile regression method to the CFFPR data discussed in Section 1. Cystic Fibrosis (CF) is one of the most common and life-shortening genetic disorders affecting the lungs and digestive systems of about 30,000 children and adults in the United States and 70,000 worldwide (Cystic Fibrosis Foundation 2010). *Pseudomonas aeruginosa* (PA), the predominant bacterial pathogen infecting 80% of CF patients under age 18, accelerates decline in lung function (Kosorok et al. 2001) and serves as an important predictor of mortality in CF (Retsch-Bogart et al. 2008). In our analysis, we used the CFFPR data collected during 1986–2005 to investigate the association between onset ages of the first detected PA infection and several risk factors in CF patients diagnosed by age 10. Similar data were analyzed by Lai et al. (2004) under the Cox model and Yan et al. (2009) based on temporal process regression.

In the CFFPR, a patient's age at the first PA infection, which is the event time of interest T , is subject to both left and right censoring. Among 12,818 CF patients diagnosed between 1986 and 2000, 3,343 (26.1%) patients had PA infection at study entry (i.e. T is left censored by the patient's age at the first CFFPR record L) and 2,213 (17.3%) patients had no PA infection documented by December 2005 (i.e. T is right censored by age at the last follow-up before the cut-off date U). To avoid the complication with delayed entry, we restricted the study population to subjects who were diagnosed before age 10 years during 1986–2000 and alive at age 10 years. The first restriction was imposed because the first 10 years were known to have greatest potential to take advantage of early diagnosis (Campbell and White, 2005, Grosse et al., 2006). The second restriction was imposed to avoid left truncation due to mortality prior to CF diagnosis. Since the mortality rate before age 10 years was very low, about 1.5% (Grosse et al., 2006), we expect excluding patients who died before age 10 years would only result in a small deviation from the general young CF population. The restricted sample contains 11,179 patients with 23.7% left censoring and 16.2% right censoring.

We applied the proposed quantile regression method to this doubly-censored restricted CFFPR sample. Since the support of left censoring time L in this dataset appears to have a lower bound approaching 0, suggesting the lower tail identifiability, we fit the data with the unconditional model (1), choosing $g(\cdot)$ as the identity function. The same set of covariates examined in Yan et al. (2009) was considered, including gender (1 for females and 0 for males), diagnosis mode (denoted by “factor”) and diagnosis year (denoted by “dx”). Diagnosis mode was defined according to common clinical practices that identify CF prior to 2005, which includes four categories: diagnosis at birth due to meconium ileus (MI), diagnosis shortly after birth by neonatal/prenatal screening (SCR), diagnosis at variable ages because of family history (FH), and diagnosis at variable ages from various symptoms (SYMP) other than MI (Lai et al., 2004, Yan et al., 2009). Diagnosis year was classified into three periods, i.e., 1986–1989 (dx86), 1990–1993 (dx90), and 1994–2000 (dx94), that coincided with the major therapeutic breakthroughs in CF, which are, Pulmozyme in 1994 (Fuchs et al. 1994, Ramsey and Dorkin 1994) and TOBI in 1999 (Ramsey et al. 1999). Boy patients who were diagnosed between 1986 and 1989 (dx86) by symptoms other than MI (SYMP) was chosen as the reference comparison group. We adopted the same τ -grid as that used in simulations.

Figure 2 displays the proposed coefficient estimates for $\beta_0(\tau)$ in bold solid lines along with their 95% pointwise Wald confidence intervals in bold dashed lines for $\tau \in [l, u]$ with $l=0.10$ and $u=0.65$. The naive estimates obtained from the subset excluding left-censored observations and the corresponding 95% confidence intervals are plotted in dot-dash lines and dotted lines respectively. The estimates for $\beta_0(\tau)$ with τ close to 0 (not shown), though exhibiting rather large variability, are all converged solutions to the L_1 -minimization

problem (5). This further suggests that the lower tail identifiability of regression quantiles may be of little concern in this example. As shown in Figure 2 (panel A), with all observations included in the analysis, the estimated intercept indicates that about 10% of male patients with CF diagnosed during 1986–1989 by SYMP acquired their first PA infection by age 2 years and approximately 65% of them had their first PA infection by age 9 years.

With regard to the gender effect (Figure 2, panel B), the regression estimates for most of the quantile range, $\tau \in (0.2, 0.65)$, are negative, indicating that girls acquired their first PA infection earlier than boys. More importantly, the gender difference is more pronounced at larger τ 's, which correspond to patients who acquired first PA infection at older ages. Similarly, many other covariates show non-constant effects across the quantile range of 0.1 to 0.65. For example, the effect of "FH diagnosis" (Figure 2, panel D) increases with τ , while the coefficient estimates for both "diagnosis 1990–1993" and "diagnosis 1994–2000" (Figure 2, panel F and G) decrease with τ . Such varying effects would not have been identified by traditional Cox regression or classic linear regression that only models the mean.

Our observation of the earlier onset of PA infection in girls compared to boys is consistent to previous literature (Lai et al., 2004, Yan et al., 2009). In contrast, our result on the beneficial effect of "FH diagnosis" (i.e., later first PA infection compared to the "SYMP diagnosis") appears to be new. This effect would have been masked by an opposite effect if left-censored cases were excluded from the analysis. Panel D of Figure 2 shows that even the pointwise 95% confidence intervals do not overlap across the entire range of τ , leading to an erroneous conclusion on the effect of "FH diagnosis". Substantial discrepancy is also noted on the magnitude of the regression estimates of covariates "MI diagnosis" (panel C), "diagnosis 1990–1993" (panel F) and "diagnosis 1994–2000" (panel G) between the proposed method and the naive method. These systematic differences are also observed in our simulation studies, and provide strong evidence to support the importance of appropriately accommodating left censored observations when investigating PA infections in CF.

Formal tests on the significance of covariate effects were performed based on the average quantile effects across τ ranging from 0.10 to 0.65. The delaying effect of FH diagnosis and the accelerating effects of more recent diagnosis cohorts (1990–1993 and 1994–2000) on first PA infections observed in Figure 2 (panels D, F, and G) are confirmed by very significant p-values, which are all < 0.001 . The effect of gender is marginally significant with a p-value of 0.06. In the view of rather monotone patterns of the coefficients for these covariates, we conducted the proposed constancy tests using the weight function $\Theta(t) = I[t < (l + u)/2]$. These analyses show that the effects of gender, FH, dx90 and dx94 may vary across τ , and the corresponding p-values are 0.002, 0.006, < 0.001 , and < 0.001 respectively, confirming the visual trends illustrated in Figure 2. These results have some interesting clinical implications. The earlier acquisition of first PA infections in females with CF reported repeatedly in the literature (Lai et al., 2004, Yan et al., 2009) is not uniform across the entire female population, this gender effect is smaller among CF patients who are subject to high risk versus low risk of PA infection. The association between more recent diagnosis cohorts and shorter times to first PA infection may be explained by the increased culture frequency in these patients, which may shorten the time to detecting PA infection in patients with late onsets of PA infection in a greater extent, as compared to those who experienced first PA infections at young ages. This may be because patients in the latter group tended to have frequent sick visits in early life, which may offset the benefit of frequent cultures in the detection of PA infection.

7. Remarks

In this paper we propose a quantile regression method for doubly censored data. The stochastic integration presentation of the proposed estimating equation facilitates asymptotic studies and entails computationally simple implementations. A useful solution to handle the unique identifiability issue with doubly censored regression quantiles is proposed based on conditional inference. We also present an adaptation of our method to settings where left truncation is present.

As the key methodological idea of this work, we utilize the martingale structure embedded with the observed doubly censored data to construct an estimating function which takes the form of a stochastic integral of $\beta(\cdot)$. The advantages of using such an estimating function include: (1) it has expectation zero with all $\tau \in (0, 1)$ under the standard random censoring mechanism (i.e. $(L, U) \perp T | Z$), (2) it enables a simple sequential algorithm to approximate the solution of the resulting estimating equation. In principle, other estimating equations that possess similar features may be used as alternative approaches.

Note that the proposed estimation procedure does not guarantee the monotonicity of $Z^T \beta(\tau)$ in τ . Nevertheless, we do not expect this would cause serious practical issues. Given the uniform consistency of $\beta_0(\tau)$ and bounded covariates, one may use $\sup_{u \in (0, \tau)} \{Z^T \beta(u)\}$ instead of $Z^T \beta(\tau)$ to produce a monotone and yet consistent prediction of quantile functions

The double censoring mechanism, $(L, U) \perp C$ given Z , is adopted in this work and thus both L and U are allowed to depend on Z . In addition, we require L be always observed, which is often true in registry study settings, as exemplified by the CFFPR data. Such additional information on left censoring time contributes to identifying an appropriate martingale for constructing estimating equations. How to relax the assumption on known L nevertheless merits future research.

Based on our numerical experience, the proposed estimator is very robust to the choice of grid points. A grid-free estimation procedure may be developed by adapting the technique in Huang (2010). Provided the uniform consistency of $\beta(\tau)$, one may apply linear or

polynomial interpolations on $\{\widehat{\beta}(\tau_j)\}_{j=1}^L$ to obtain a consistent estimator of $\beta_0(\tau)$ that is smooth in τ .

As suggested by one referee, a smooth B-spline estimator of $\beta_0(\tau)$ may be derived directly based on equation (3). The computational procedure required to obtain the smooth B-spline estimator is more complex than for our step function estimator, because the representation (4) is no longer available. In addition, while we anticipate the large sample behavior of the smooth B-spline estimator will be the same as that of our $\hat{\beta}(\tau)$, the asymptotic arguments we used for our $\hat{\beta}(\tau)$ cannot be readily adapted, due to the more complex structure of the B-spline estimator. Development of computational techniques and rigorous asymptotic theory for the smooth B-spline estimator merits future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Dr. Zhumin Zhang for her useful comments on the manuscript. This work was partially supported by National Science foundation grants DMS-0706985 and DMS-1007660 (to Peng) and National Institutes of Health grant DK072126 (to Lai).

References

- Barrodale I, Roberts F. Solution of an overdetermined system of equations in the l_1 norm. *Communications of the ACM*. 1974; 17:319–320.
- Buckley J, James I. Linear regression with censored data. *Biometrika*. 1979; 66:429–436.
- Cai T, Cheng S. Semiparametric regression analysis for doubly censored data. *Biometrika*. 2004; 91:277–290.
- Campbell, Pr; White, T. Newborn screening for cystic fibrosis: an opportunity to improve care and outcomes. *The Journal of Pediatrics*. 2005; 147:S2–S5. [PubMed: 16202776]
- Chang MN, Yang GL. Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *The Annals of Statistics*. 1987; 15:1536–1547.
- Cheng S, Wei L, Ying Z. Analysis of transformation models with censored data. *Biometrika*. 1995; 82:835–845.
- Cox D. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*. 1972; 34:187–220.
- Emerson J, Rosenfeld M, McNamara S, Ramsey B, Gibson R. *Pseudomonas aeruginosa* and other predictors of mortality and morbidity in young children with cystic fibrosis. *Pediatric Pulmonology*. 2002; 34:91–100. [PubMed: 12112774]
- Fine JP, Yan J, Kosorok MR. Temporal process regression. *Biometrika*. 2004; 91:683–703.
- FitzSimmons S. The changing epidemiology of cystic fibrosis. *The Journal of Pediatrics*. 1993; 122:1–9. [PubMed: 8419592]
- Fyngenson M, Ritov Y. Monotone estimating equations for censored data. *The Annals of Statistics*. 1994; 22:732–746.
- Gehan EA. A generalized two-sample wilcoxon test for doubly censored data. *Biometrika*. 1965; 52:650–653. [PubMed: 5858975]
- Grosse S, Rosenfeld M, Devine O, Lai H, PM F. Potential impact of newborn screening for cystic fibrosis on child survival: a systematic review and analysis. *The Journal of Pediatrics*. 2006; 149:362–366. [PubMed: 16939748]
- Honore B, Khan S, Powell J. Quantile regression under random censoring. *Journal of Econometrics*. 2002; 109:67–105.
- Huang Y. Quantile Calculus and Censored Regression. *The Annals of Statistics*. 2010; 38:1607–1637.
- Huang Y, Peng L. Accelerated recurrent time models. *Scandinavian Journal of Statistics*. 2009; 36:636–648.
- Jin Z, Lin D, Wei L, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003; 90:341–353.
- Kaplan E, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Kerem E, Reisman J, Corey M, Canny G, Levison H. Prediction of mortality in patients with cystic fibrosis. *New England Journal of Medicine*. 1992; 326:1187–1191. [PubMed: 1285737]
- Koenker R, Bassett G. Regression quantiles. *Econometrica*. 1978; 46:33–50.
- Lai H, Cheng Y, Cho H, Kosorok M, Farrell P. Association between initial disease presentation, lung disease outcomes, and survival in patients with cystic fibrosis. *American Journal of Epidemiology*. 2004; 159:537–546. [PubMed: 15003957]
- Louis T. Nonparametric analysis of an accelerated failure time model. *Biometrika*. 1981; 68:381–390.
- Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*. 1966; 50:163–170. [PubMed: 5910392]
- Neocleous T, Branden KV, Portnoy S. Correction to censored regression quantiles by S. Portnoy. *Journal of the American Statistical Association*. 2006; 101:860–861.
- Peng L, Huang Y. Survival analysis with temporal covariate effects. *Biometrika*. 2007; 94:719–733.
- Peng L, Huang Y. Survival analysis with quantile regression models. *Journal of the American Statistical Association*. 2008; 103:637–649.
- Portnoy S. Censored regression quantiles. *Journal of the American Statistical Association*. 2003; 98:1001–1012.

- Portnoy S, Lin G. Asymptotics for censored regression quantiles. *Journal of Nonparametric Statistics*. 2010; 22:115–130.
- Powell J. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*. 1984; 25:303–325.
- Prentice R. Linear rank tests with right-censored data. *Biometrika*. 1978; 65:167–179.
- Ren J-J. Regression m-estimators with non-i.i.d doubly censored data. *The Annals of Statistics*. 2003; 31:1186–1219.
- Ren J-J. Weighted empirical likelihood in some two-sample semiparametric models with various types of censored data. *The Annals of Statistics*. 2008; 36:147–166.
- Ren J-J, Gu M. Regression m-estimators with doubly censored data. *The Annals of Statistics*. 1997; 25:2638–2664.
- Ritov Y. Estimation in a linear regression model with censored data. *The Annals of Statistics*. 1990; 18:303–328.
- Sun, J. *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer; 2006.
- Tsai W, Jewell N, Wang M. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*. 1987; 74:883–886.
- Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*. 1990; 18:354–372.
- Turnbull BW. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*. 1974; 69:169–173.
- Wei L, Gail M. Nonparametric estimation for a scale-change with censored observations. *Journal of the American Statistical Association*. 1983; 78:382–388.
- Wei L, Ying Z, Lin D. Linear regression analysis of censored survival data based on rank tests. *Biometrika*. 1990; 77:845–851.
- Yan J, Cheng Y, Fine JP, Lai HJ. Uncovering symptom progression history from disease registry data with application to young cystic fibrosis patients. *Biometrics*. 2009; 66:594–602. [PubMed: 19522871]
- Yang S. Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association*. 1999; 94:137–145.
- Ying Z, Jung S, Wei L. Survival analysis with median regression models. *Journal of the American Statistical Association*. 1995; 90:178–184.
- Zhang C-H, Li X. Linear regression with doubly censored data. *The Annals of Statistics*. 1996; 24:2720–2743.

Figure 1.
Comparison among True coefficients (Bold Solid Lines), Mean Estimated Coefficients from the Proposed Method (Solid Lines), and Mean Estimated Coefficients from the Naive Approach (Dotted Lines).

Figure 2. Coefficient Estimates (Bold Solid Lines) and 95% Pointwise Confidence Intervals (Bold Dashed Lines) from the Proposed Method, in Contrast with Coefficient Estimates (Dot-Dash Lines) and 95% Pointwise Confidence Intervals (Dotted Lines) from the Naive Method.

Table 1

Simulation Results under AFT Models. Bias: absolute biases; AvgSD: average estimated resampling-based standard deviations; EmpSD: empirical standard deviations; Cov95: coverage rates of 95% Wald confidence intervals.

Unconditional Case						Conditional Case					
$b_1=0, b_2=-0.5$						$b_1=0, b_2=-1$					
τ	Bias	AvgSD	EmpSD	Cov95		Bias	AvgSD	EmpSD	Cov95		
0.1	β_0	0.07	0.93	0.96	$\hat{\alpha}_0$	0.02	0.77	0.72	0.90		
	β_1	0.01	1.36	1.49	$\hat{\alpha}_1$	0.06	1.18	1.15	0.92		
	β_2	0.00	0.92	0.86	$\hat{\alpha}_2$	0.01	0.75	0.66	0.95		
0.3	β_0	0.04	0.55	0.49	$\hat{\alpha}_0$	0.02	0.41	0.38	0.94		
	β_1	0.03	0.82	0.77	$\hat{\alpha}_1$	0.04	0.66	0.61	0.95		
	β_2	0.02	0.51	0.43	$\hat{\alpha}_2$	0.02	0.39	0.36	0.96		
0.5	β_0	0.01	0.35	0.31	$\hat{\alpha}_0$	0.01	0.31	0.28	0.95		
	β_1	0.03	0.55	0.50	$\hat{\alpha}_1$	0.02	0.49	0.45	0.96		
	β_2	0.01	0.32	0.29	$\hat{\alpha}_2$	0.01	0.29	0.27	0.95		
0.7	β_0	0.01	0.28	0.24	$\hat{\alpha}_0$	0.03	0.28	0.26	0.95		
	β_1	0.02	0.43	0.39	$\hat{\alpha}_1$	0.02	0.43	0.39	0.96		
	β_2	0.00	0.25	0.23	$\hat{\alpha}_2$	0.01	0.25	0.23	0.96		

Simulation Results on Hypothesis Testing and Second-Stage Inference under AFT Models. ERR: empirical rejection rates; AvgEst: estimated average effects; AvgSD: average estimated resampling-based standard deviations; EmpSD: empirical standard deviations.

Table 2

Unconditional Case				
$b_1=0, b_2=-0.5$				
$H_0: \beta(\tau) = 0, l \leq \tau \leq u$		$H_0: \beta(\tau) = \eta_0, l \leq \tau \leq u$		
ERR	AvgEst	AvgSD	EmpSD	ERR
β_0	-0.82	0.40	0.39	0.89
β_1	-0.03	0.60	0.63	0.05
β_2	-0.49	0.37	0.34	0.04
Conditional Case				
$b_1=0, b_2=-1$				
$H_0: u(\tau) = 0, l \leq \tau \leq u$		$H_0: u(\tau) = \eta_0, l \leq \tau \leq u$		
ERR	AvgEst	AvgSD	EmpSD	ERR
$\hat{a}_0$	-0.79	0.31	0.30	0.96
$\hat{a}_1$	-0.03	0.49	0.49	0.05
$\hat{a}_2$	-1.01	0.30	0.29	0.03

Simulation Results under Log-Linear Models with Heteroscedastic Errors. Bias: absolute biases; AvgSD: average estimated resampling-based standard deviations; EmpSD: empirical standard deviations; Cov95: coverage rates of 95% Wald confidence intervals.

Table 3

Unconditional Case						Conditional Case					
$b_1=0, b_2=-1.5$						$b_1=0, b_2=-4.5$					
τ	Bias	AvgSD	EmpSD	Cov95		Bias	AvgSD	EmpSD	Cov95		
0.1	β_0	0.06	0.64	0.60	0.91	$\hat{\alpha}_0$	0.05	0.71	0.62	0.94	
	β_1	0.05	1.11	1.09	0.91	$\hat{\alpha}_1$	0.01	1.15	1.04	0.94	
	β_2	0.09	1.09	1.02	0.85	$\hat{\alpha}_2$	0.02	0.82	0.75	0.94	
0.3	β_0	0.04	0.41	0.37	0.94	$\hat{\alpha}_0$	0.01	0.40	0.36	0.95	
	β_1	0.02	0.70	0.63	0.95	$\hat{\alpha}_1$	0.01	0.67	0.62	0.95	
	β_2	0.00	0.65	0.53	0.95	$\hat{\alpha}_2$	0.01	0.49	0.44	0.95	
0.5	β_0	0.01	0.33	0.31	0.94	$\hat{\alpha}_0$	0.01	0.32	0.28	0.96	
	β_1	0.00	0.56	0.52	0.95	$\hat{\alpha}_1$	0.03	0.54	0.49	0.95	
	β_2	0.01	0.42	0.37	0.95	$\hat{\alpha}_2$	0.00	0.39	0.36	0.95	
0.7	β_0	0.02	0.32	0.29	0.96	$\hat{\alpha}_0$	0.03	0.30	0.27	0.95	
	β_1	0.00	0.53	0.49	0.95	$\hat{\alpha}_1$	0.02	0.51	0.46	0.96	
	β_2	0.02	0.35	0.31	0.96	$\hat{\alpha}_2$	0.01	0.36	0.33	0.95	

Simulation Results on Hypothesis Testing and Second-Stage Inference under Log-Linear Models with Heteroscedastic Errors. ERR: empirical rejection rates; AvgEst: estimated average effects; AvgSD: average estimated resampling-based standard deviations; EmpSD: empirical standard deviations.

Table 4

Unconditional Case					
$b_1=0, b_2=-1.5$					
$H_0: \beta(\tau) = 0, I \leq \tau \leq u$			$H_0: \beta(\tau) = \eta_0, I \leq \tau \leq u$		
ERR	AvgEst	AvgSD	EmpSD	ERR	
β_0	-0.32	0.32	0.31	0.90	
β_1	0.00	0.54	0.53	0.03	
β_2	-1.50	0.45	0.42	0.23	
Conditional Case					
$b_1=0, b_2=-4.5$					
$H_0: u(\tau) = 0, I \leq \tau \leq u$			$H_0: u(\tau) = \eta_0, I \leq \tau \leq u$		
ERR	AvgEst	AvgSD	EmpSD	ERR	
$\hat{a}_0$	-0.79	0.30	0.28	0.98	
$\hat{a}_1$	-0.02	0.50	0.48	0.04	
$\hat{a}_2$	-1.66	0.37	0.36	0.14	