

Survey Methodology

Weighted censored quantile regression

by Chithran Vasudevan, Asokan Mulayath Variyath
and Zhaozhi Fan

Release date: May 7, 2019



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- | | |
|---|----------------|
| • Statistical Information Service | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line | 1-514-283-9350 |

Depository Services Program

- | | |
|------------------|----------------|
| • Inquiries line | 1-800-635-7943 |
| • Fax line | 1-800-565-7757 |

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "[Standards of service to the public](#)."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© Her Majesty the Queen in Right of Canada as represented by the Minister of Industry, 2019

All rights reserved. Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

An HTML version is also available.

Cette publication est aussi disponible en français.

Weighted censored quantile regression

Chithran Vasudevan, Asokan Mulayath Variyath and Zhaozhi Fan¹

Abstract

In this paper, we make use of auxiliary information to improve the efficiency of the estimates of the censored quantile regression parameters. Utilizing the information available from previous studies, we computed empirical likelihood probabilities as weights and proposed weighted censored quantile regression. Theoretical properties of the proposed method are derived. Our simulation studies shown that our proposed method has advantages compared to standard censored quantile regression.

Key Words: Empirical Likelihood; Right censoring; Kaplan-Meier Estimator.

1 Introduction

In quantile regression (Koenker, 2005), the conditional quantiles of the response variable for a given set of predictor variables are modelled. The regression parameters are estimated by minimizing a check loss function at a specific quantile, τ , instead of the square loss function as in the standard linear regression. A quantile regression model based on properly selected quantiles could provide a global assessment of the covariate effects on the response, which is often ignored by the standard linear regression model. Recently, censored quantile regression has been studied extensively. Powell (1984) introduced the least absolute deviation (LAD) estimator, also called the median regression model for the left censored survival data, using the censored Tobit model (Tobin, 1958). Powell (1986) generalized the LAD estimation to any quantile.

Portnoy (2003) introduced a censored quantile regression model under random censoring as a generalization of the Kaplan-Meier estimator recursively using the Kaplan-Meier estimator (Kaplan and Meier, 1958). Peng and Huang (2008) developed a censored quantile regression model based on the Nelson-Aalen estimator using counting processes and martingale theory. In survival analysis setup, for the i^{th} ($i = 1, 2, \dots, n$) subject, let T_i be the logarithm of the failure time, C_i the logarithm of right censoring time, \mathbf{X}_i the p -vector covariate and let $Y_i = \min(T_i, C_i)$ be the logarithm of the survival time. For a given quantile, τ , the regression coefficients, $\boldsymbol{\beta}(\tau)$, can be estimated as

$$\hat{\boldsymbol{\beta}}(\tau) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_{\tau}(Y_i - \min\{C_i, \mathbf{X}_i^T \boldsymbol{\beta}\}), \quad (1.1)$$

where $\rho_{\tau}(u) = u[\tau - \mathbb{I}(u < 0)]$, is the check loss function.

In many studies, we may have some information about the target population from previous studies. This is common in survey sampling since surveys are carried out repeatedly with similar objectives. For example, in survey sampling, information about the population mean and variance could be available from previous surveys or records. The information of the parameters as well as type of relationship, distributional

1. Chithran Vasudevan, Department of Mathematics and Statistics, Memorial University, St.John's, NL A1C 5S7. E-mail: chithran@mun.ca; Asokan Mulayath Variyath, Department of Mathematics and Statistics, Memorial University, St.John's, NL A1C 5S7. E-mail: varyath@mun.ca; Zhaozhi Fan, Department of Mathematics and Statistics, Memorial University, St.John's, NL A1C 5S7. E-mail: zhaozhi@mun.ca.

assumptions, etc. also could be considered as auxiliary information available for analysis. The auxiliary information could be effectively used to improve the efficiency of the statistical inference (Kuk and Mak, 1989; Rao, Kovar and Mantel, 1990; Chen and Qin, 1993). The idea used in this paper can be easily extendable in survey sampling to arrive efficient parameter estimates by making use of the information available from previous surveys.

Consider a known relationship between the survival time, Y (or the failure time, T) and a set of covariates \mathbf{X} , as $Y = f(\mathbf{X}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter of interest. The knowledge about this relationship can be treated as auxiliary information. In a more general case, the auxiliary information can be expressed as $E\{g(\mathbf{Z}; \boldsymbol{\theta})\} = 0$ for some d -dimensional parameter, $\boldsymbol{\theta} \in R^d$, where \mathbf{Z} is the observed data from the present study and $g(\mathbf{Z}; \boldsymbol{\theta}) \in R^q$, some function with $q \geq d$. The parameter, $\boldsymbol{\theta}$ could be unknown, but can be estimated using the information available from previous studies.

Chen and Qin (1993) introduced the use of auxiliary information to improve the efficiency of estimators in the context of survey sampling using empirical likelihood (Owen, 1988, 2001). Li and Wang (2003) accommodated the auxiliary information to the censored linear regression model using empirical likelihood by defining a synthetic variable (Koul, Susarla and Ryzin, 1981). Fang, Li, Lu and Qin (2013) proposed the effective use of auxiliary information in the linear regression model with right censored data using empirical likelihood, by utilizing the Buckley-James (Buckley and James, 1979) estimating equation. Tang and Leng (2012) introduced an empirical likelihood based linear quantile regression model using auxiliary information.

In this paper, we propose an empirical likelihood (EL) based approach to accommodate auxiliary information to the censored quantile regression. EL is a non-parametric likelihood approach proposed by Owen (1988, 2001), which has similar properties of parametric likelihood. We utilize the EL based data driven probabilities as the weights by using the estimating function, $g(\mathbf{Z}; \boldsymbol{\theta})$ and incorporate those weights into the censored quantile regression model. The resulted weighted censored quantile regression parameter $\beta(\tau)$ can be estimated as

$$\hat{\beta}(\tau) = \arg \min_{\beta \in \mathcal{R}^p} \sum_{i=1}^n \omega_i \rho_{\tau}(Y_i - \min\{C_i, \mathbf{X}_i^T \beta\}), \quad (1.2)$$

where ω_i 's are the weights. We propose to use the EL based data driven probabilities as the weights. Our simulation results show that the EL based weighted censored quantile regression performs more efficiently than the standard linear censored quantile regression.

The rest of the paper is organized as follows. In Section 2, we present the estimation procedure of the EL based data driven probabilities. In Section 3, we introduce the EL based weighted censored quantile regression and investigate the asymptotic properties of the estimators. In Section 4, performance analysis of the proposed method is conducted using the simulations. The application to the north central cancer treatment lung cancer data is also presented as an illustration. Our conclusions are given in Section 5.

2 Estimation of weights using empirical likelihood

We develop a method that converts the auxiliary information to the EL based data driven probabilities, which are further used in the weighted censored quantile regression as the weights. Qin and Lawless (1994) developed the EL approach based on a set of estimating equations. Let $\{\mathbf{Z}_i\}_{i=1}^n$ be the observed data and the available auxiliary information is represented by the estimating function $g(\mathbf{Z}_i; \boldsymbol{\theta})$ with parameter, $\boldsymbol{\theta}$ which is known. Then, the maximum empirical likelihood is given by

$$L_{\text{EL}}(\boldsymbol{\theta}) = \sup \left\{ \prod_{i=1}^n P_i : P_i \geq 0, \sum_{i=1}^n P_i = 1, \sum_{i=1}^n P_i g(\mathbf{Z}_i; \boldsymbol{\theta}) = 0 \right\}, \quad (2.1)$$

where $P_i = \Pr(Z_i = z_i)$ and $\boldsymbol{\theta}$ is the parameter in the auxiliary information which can be assumed to be known. The parameter, $\boldsymbol{\theta}$ could be any parametric information available from the previous studies which has an influence on the model parameter, $\boldsymbol{\beta}(\tau)$. For a given $g(\mathbf{Z}_i; \boldsymbol{\theta})$, $\boldsymbol{\theta}$ should satisfy $E\{g(\mathbf{Z}_i; \boldsymbol{\theta})\} = 0$ to avoid the non-existence of solutions due to convex hull issues. This is the scenario for when zero is not an inner point of the convex hull of the $g(\mathbf{Z}_i; \boldsymbol{\theta})$, $i = 1, 2, \dots, n$, which will fail to provide positive P_i 's. For a given value of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, using the Lagrange multiplier method, $L_{\text{EL}}(\boldsymbol{\theta}_0)$ attains its maximum at

$$P_i(\boldsymbol{\theta}_0) = \frac{1}{n \{1 + \lambda_{\boldsymbol{\theta}_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)\}}, \quad i = 1, 2, \dots, n. \quad (2.2)$$

The Lagrange multiplier, $\lambda_{\boldsymbol{\theta}_0}$ is the solution to the equation

$$\sum_{i=1}^n \frac{g(\mathbf{Z}_i; \boldsymbol{\theta}_0)}{n \{1 + \lambda_{\boldsymbol{\theta}_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)\}} = 0.$$

The estimated $P_i(\cdot)$'s are used as the weights (ω_i) in (1.2) for the weighted censored quantile regression. In some cases, $\boldsymbol{\theta}$ may not be available and in such situations, we can use an estimate of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}_A$ obtained from previous studies. Chen and Qin (1993) showed that for a random sample, Y_i , and $P_i(\cdot)$'s are estimated using (2.2), $\tilde{F}_n(y) = \sum_{i=1}^n P_i \mathbb{I}(Y_i \leq y)$ has smaller variance than the empirical distribution function, $\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(Y_i \leq y)$. As a result, with Bahadur representation (Bahadur, 1966), for a given τ ($0 < \tau < 1$), the quantile estimate, $\tilde{F}_n^{-1}(\tau)$ has smaller variance than $\hat{F}_n^{-1}(\tau)$ (See Kuk and Mak, 1989; Rao et al., 1990). Hence our proposed method is expected to be more efficient than the ordinary censored quantile regression.

3 Estimation of weighted censored quantile regression parameters

Define the distribution function of T_i conditional on the p -vector covariate, \mathbf{X}_i as $F_{T_i}(t | \mathbf{X}_i) = \Pr(T_i \leq t | \mathbf{X}_i)$. Let $\Lambda_{T_i}(t | \mathbf{X}_i) = -\log\{1 - \Pr(T_i \leq t | \mathbf{X}_i)\}$, $N_i(t) = \mathbb{I}(Y_i \leq t, \delta_i = 1)$, and $M_i(t) = N_i(t) - \Lambda_{T_i}(t \Delta Y_i | \mathbf{X}_i)$. Here $\Lambda_{T_i}(\cdot | \mathbf{X}_i)$ is the cumulative hazard function conditional on \mathbf{X}_i , $N_i(t)$ is the counting process and $M_i(t)$ is the martingale process associated with $N_i(t)$ (Fleming and Harrington, 2011). We consider an extension of censored quantile regression estimation procedure proposed by Peng and Huang (2008), incorporating the P_i 's as known weights arrived based on the auxiliary information

available through the known parameter θ . Note that θ and $\beta(\tau)$ are distinct parameters and estimating function $g(z; \theta)$ used for computing P_i 's are different from the estimating functions used for quantile regression parameters in (1.1). Since P_i 's are independent of $\beta(\tau)$, $E\{P_i M_i(t) | \mathbf{X}_i\} = \mathbf{0}$ (by the martingale property) for $t \geq 0$, we have

$$E \left\{ \sqrt{n} \sum_{i=1}^n P_i \mathbf{X}_i (N_i(e^{\mathbf{X}_i^T \beta_0(\tau)}) - \Lambda_T[e^{\mathbf{X}_i^T \beta_0(\tau)} \wedge Y_i | \mathbf{X}_i]) \right\} = \mathbf{0}, \quad (3.1)$$

where $\beta_0(\tau)$ denotes the true $\beta(\tau)$, in (1.2) for a given quantile, τ .

Since $\Lambda_{T_i}(\cdot | \mathbf{X}_i)$, $i = 1, 2, \dots, n$ are unknown functions, Peng and Huang (2008) derived the relationship between $\Lambda_T[e^{\mathbf{X}_i^T \beta_0(\tau)} \wedge Y_i | \mathbf{X}_i]$ and $\beta_0(\tau)$ to use (3.1) to estimate $\beta_0(\tau)$. Using the fact that $F_{\beta_0} [e^{\mathbf{X}_i^T \beta_0(u)} | \mathbf{X}_i] = \tau$ and utilizing the monotonicity of $\mathbf{X}_i^T \beta_0(\tau)$ in τ , they showed that $\Lambda_T[e^{\mathbf{X}_i^T \beta_0(\tau)} \wedge Y_i | \mathbf{X}_i] = \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^T \beta(u)}] dH(u)$, where $H(u) = -\log(1-u)$ for $0 \leq u < 1$.

So, our weighted censored quantile regression estimating equation is

$$\sqrt{n} S_n(\beta, \tau) = \mathbf{0}, \quad (3.2)$$

where

$$S_n(\beta, \tau) = \sum_{i=1}^n P_i \mathbf{X}_i \left\{ N_i(e^{\mathbf{X}_i^T \beta(\tau)}) - \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^T \beta(u)}] dH(u) \right\}.$$

Here P_i 's are defined in (2.2). Let $s(\beta, \tau) = E\{S_n(\beta, \tau)\}$ and the martingale property of $\mathbb{M}(\cdot)$ gives $s(\beta_0, \tau) = \mathbf{0}$. For a particular quantile, τ_k and an estimator of $\beta_0(\tau_k)$, $\hat{\beta}(\tau_k)$ is a right-continuous step function which jumps only on a grid, $\mathbb{S}_L = \{0 = \tau_0 < \tau_1 < \dots < \tau_L = \tau_U < 1\}$. Here L depends on the sample size, n . The size of \mathbb{S}_L is defined as $\|\mathbb{S}_L\| = \max_k (\tau_k - \tau_{k-1})$.

For different quantiles, $\tau_0, \tau_1, \dots, \tau_L$ ($0 = \tau_0 < \tau_1 < \dots < \tau_L < 1$), based on (3.2), we can obtain $\hat{\beta}(\tau_k)$ ($k = 1, 2, \dots, L$) by recursively solving the following monotone estimating equation for $\beta(\tau_k)$:

$$\sqrt{n} \sum_{i=1}^n P_i \mathbf{X}_i \left\{ N_i(e^{\mathbf{X}_i^T \beta(\tau_k)}) - \sum_{r=0}^{k-1} \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^T \hat{\beta}(\tau_r)}] \{H(\tau_{r+1}) - H(\tau_r)\} \right\} = \mathbf{0}. \quad (3.3)$$

We define the estimators, $\hat{\beta}(\tau_k)$ as the generalized solutions (Fygenon and Ritov, 1994) because equation (3.3) is not continuous and the solution may not be unique.

3.1 Asymptotic theory

We derived the asymptotic properties of the EL based weighted censored quantile regression estimators using the approach of Peng and Huang (2008). Now we prove the uniform consistency and weak Gaussian convergence of the proposed weighted censored quantile regression estimator, $\hat{\beta}(\cdot)$.

Define $F(t | \mathbf{X}) = \Pr(Y \leq t | \mathbf{X})$, $\bar{F}(t | \mathbf{X}) = \Pr(Y > t | \mathbf{X})$, $\tilde{F}(t | \mathbf{X}) = \Pr(Y \leq t, \delta = 1 | \mathbf{X})$, $\bar{f}(y | \mathbf{X}) = -f(y | \mathbf{X}) = -dF(y | \mathbf{X})/dy$ and $\tilde{f}(y | \mathbf{X}) = d\tilde{F}(y | \mathbf{X})/dy$. (For a vector h , $h^{\otimes 2} = hh^T$, $h^{(l)} = l^{\text{th}}$ component of h , $\|h\|$ is the Euclidean norm of h .)

Define $\mathbf{W}_i = \lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0) \mathbf{X}_i$, $i = 1, 2, \dots, n$ as a p -vector.

Regularity conditions:

- R.1: The observations, \mathbf{Z}_i , $i = 1, 2, \dots, n$ are iid observations from some distribution. Without loss of generality, we assume that $(Y_i, \delta_i, \mathbf{X}_i^\top)^\top \subset \mathbf{Z}_i$ for all $i = 1, 2, \dots, n$.
- R.2: There exists $\boldsymbol{\theta}_0$ such that $E\{g(\mathbf{Z}_i; \boldsymbol{\theta}_0)\} = 0$, the matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = E\{g(\mathbf{Z}_i; \boldsymbol{\theta}_0) g(\mathbf{Z}_i; \boldsymbol{\theta}_0)^\top\}$ is positive definite, $\frac{\partial g(\mathbf{z}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is continuous in the neighborhood of $\boldsymbol{\theta}_0$. The matrix $E\left\{\frac{\partial g(\mathbf{Z}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}$ is of full rank.
- R.3: There exist functions $H_{lj}(\mathbf{z})$ such that for $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_0$, $\left|\frac{\partial g_l(\mathbf{z}; \boldsymbol{\theta})}{\partial \theta_j}\right| \leq H_{lj}(\mathbf{z})$, where for a constant C , $E\{H_{lj}^2(\mathbf{Z})\} \leq C < \infty$ for $l = 1, \dots, q$ and $j = 1, \dots, d$.
- R.4: $\sup_i \|\mathbf{X}_i\| < \infty$ and $\sup_i \|\mathbf{X}_i \mathbf{Y}_i\| < \infty$.
- R.5: (a) Each component of $E[\mathbf{XN}(e^{\mathbf{X}^\top \boldsymbol{\beta}_0(\tau)})]$ is a Lipschitz function of τ .
 (b) $\tilde{f}(t|\mathbf{x})$ and $f(t|\mathbf{x})$ are bounded above uniformly in t and \mathbf{x} .
- R.6: (a) $\tilde{f}(e^{\mathbf{X}^\top \mathbf{b}}|\mathbf{X}) > 0$ for all $\mathbf{b} \in \mathfrak{B}(d_0)$, where $\mathfrak{B}(d) = \left\{\mathbf{b} \in \mathfrak{R}^p: \inf_{\tau \in (0, \tau_U)} \|\boldsymbol{\mu}(\mathbf{b}) - \boldsymbol{\mu}\{\boldsymbol{\beta}_0(\tau)\}\| \leq d\right\}$ for $d > 0$, and $\boldsymbol{\mu}(\mathbf{b}) = E[\mathbf{XN}(e^{\mathbf{X}^\top \mathbf{b}})]$, is a neighbourhood containing $\{\boldsymbol{\beta}_0(\tau), \tau \in (0, \tau_U)\}$.
 (b) To have the positive definiteness, $E\{\mathbf{X}^{\otimes 2}\} > 0$.
 (c) Each component of $E[\mathbf{X}^{\otimes 2} \tilde{f}(e^{\mathbf{X}^\top \mathbf{b}}|\mathbf{X}) e^{\mathbf{X}^\top \mathbf{b}}] \times (E[\mathbf{X}^{\otimes 2} \tilde{f}(e^{\mathbf{X}^\top \mathbf{b}}|\mathbf{X}) e^{\mathbf{X}^\top \mathbf{b}}])^{-1}$ is uniformly bounded in $\mathbf{b} \in \mathfrak{B}(d_0)$; $\mathfrak{B}(d_0)$.
- R.7: For any $v \in (0, \tau_U)$, $\inf_{\tau \in [v, \tau_U]} \text{eigmin} E[\mathbf{X}^{\otimes 2} \tilde{f}(e^{\mathbf{X}^\top \boldsymbol{\beta}_0(\tau)}|\mathbf{X}) e^{\mathbf{X}^\top \boldsymbol{\beta}_0(\tau)}] > 0$, where $\text{eigmin}(\cdot)$ denotes the minimum eigenvalue of a matrix.

Theorem 1. Assuming that the regularity conditions R.1-R.7 hold, if $\lim_{n \rightarrow \infty} \|\mathbb{S}_L\| = 0$, 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 that the regularity conditions R.1-R.7 hold, if $\lim_{n \rightarrow \infty} n^{1/2} \|\mathbb{S}_L\| = 0$, then $n^{1/2} \{\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\}$ weakly converges to a zero-mean Gaussian process for $\tau \in [v, \tau_U]$, where $0 < v < \tau_U$.

To prove Theorems 1 and 2, we need to show that $\max_{1 \leq i \leq n} |\lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)| = o_p(1)$. We consider two different types of $g(\mathbf{Z}_i; \boldsymbol{\theta})$. First, $g(\mathbf{Z}_i; \boldsymbol{\theta})$ does not contain the censored observations, as given in (4.1). The second, $g(\mathbf{Z}_i; \boldsymbol{\theta})$, contains the censored observations, as given in (4.5).

In the case of uncensored observations, by Owen (1991) and Lemma 11.2 of Owen (2001), we have $\max_{1 \leq i \leq n} \|g(\mathbf{Z}_i; \boldsymbol{\theta}_0)\| = o_p(\sqrt{n})$. By Lemma 1 of Tang and Leng (2012), we have under the regularity conditions R.2, R.3; the λ_{θ_0} in (2.2) satisfies $\|\lambda_{\theta_0}\| = O_p\left(\frac{1}{\sqrt{n}}\right)$. So,

$$\max_{1 \leq i \leq n} |\lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)| = O_p\left(\frac{1}{\sqrt{n}}\right) o_p(\sqrt{n}) = o_p(1). \quad (3.4)$$

Under the condition R.4; Qin and Jing (2001) proved $\max_{1 \leq i \leq n} |\lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)| = o_p(1)$ for the $g(\cdot)$ with censored observations.

Now following Owen (2001), using Taylor's series expansion of the weights, P_i 's defined in (2.2) can be rewritten as,

$$\begin{aligned} P_i(\boldsymbol{\theta}_0) &= \frac{1}{n \{1 + \lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)\}} \\ &= \frac{1}{n} [1 - \lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0) \{1 + o_p(1)\}] \\ &= \frac{1}{n} [1 - \lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)] + o_p\left(\frac{1}{n}\right); \quad i = 1, 2, \dots, n. \end{aligned}$$

Now we rewrite the $S_n(\boldsymbol{\beta}, \tau)$ as

$$\begin{aligned} S_n(\boldsymbol{\beta}, \tau) &= \frac{1}{n} \sum_{i=1}^n [1 - \lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0)] \mathbf{X}_i \left\{ \mathbb{N}_i(e^{\mathbf{X}_i^\top \boldsymbol{\beta}(\tau)}) - \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^\top \boldsymbol{\beta}(u)}] dH(u) \right\} + o_p\left(\frac{1}{n}\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \left\{ \mathbb{N}_i(e^{\mathbf{X}_i^\top \boldsymbol{\beta}(\tau)}) - \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^\top \boldsymbol{\beta}(u)}] dH(u) \right\} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \lambda_{\theta_0}^\top g(\mathbf{Z}_i; \boldsymbol{\theta}_0) \mathbf{X}_i \left\{ \mathbb{N}_i(e^{\mathbf{X}_i^\top \boldsymbol{\beta}(\tau)}) - \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^\top \boldsymbol{\beta}(u)}] dH(u) \right\} + o_p\left(\frac{1}{n}\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \left\{ \mathbb{N}_i(e^{\mathbf{X}_i^\top \boldsymbol{\beta}(\tau)}) - \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^\top \boldsymbol{\beta}(u)}] dH(u) \right\} \\ &\quad - \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i \left\{ \mathbb{N}_i(e^{\mathbf{X}_i^\top \boldsymbol{\beta}(\tau)}) - \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^\top \boldsymbol{\beta}(u)}] dH(u) \right\} + o_p\left(\frac{1}{n}\right). \end{aligned}$$

Asymptotically, by (3.4) we have $\|\mathbf{W}_i\| = o_p(1)$; $i = 1, 2, \dots, n$. So,

$$S_n(\boldsymbol{\beta}, \tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \left\{ \mathbb{N}_i(e^{\mathbf{X}_i^\top \boldsymbol{\beta}(\tau)}) - \int_0^\tau \mathbb{I}[Y_i \geq e^{\mathbf{X}_i^\top \boldsymbol{\beta}(u)}] dH(u) \right\} + o_p\left(\frac{1}{n}\right).$$

Asymptotically this estimating function, $S_n(\boldsymbol{\beta}, \tau)$ is equivalent to that in Peng and Huang (2008). Following the similar arguments of Peng and Huang (2008), we complete the proofs of Theorems 1 and 2.

As indicated in Peng and Huang (2008), the estimation of asymptotic variance of the quantile regression estimates is not easy since the covariance matrix of the limiting process of $\sqrt{n} \{\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\}$ involves unknown density function $f(y|\mathbf{X})$ and $\tilde{f}(y|\mathbf{X})$. Instead of using a smoothing or other numerical approximations, we suggest a simple bootstrap approach to estimate the standard errors of the regression estimates. This approach is used in our performance analysis discussed in next section.

4 Performance analysis

We conduct extensive simulation studies to compare the performance between our proposed EL based weighted censored quantile regression estimator and the standard censored quantile regression estimator. For our simulation, we use the models discussed in Tang and Leng (2012).

The simulation models used to generate the logarithmic event time (T_r) and logarithmic censoring time (C_r) for the r^{th} ($r = 1, 2, \dots, N$) subject are given in Table 4.1 under four Cases (i)-(iv).

Table 4.1
Four simulation models to generate event and censoring times

Cases	Models	Error Distribution
(i)	$T_r = \theta_0 + \theta_1 x_{1r} + \theta_2 x_{2r} + u_r,$ $C_r = \gamma_0 + \gamma_1 x_{1r} + \gamma_2 x_{2r} + v_r.$	$u_r, v_r \sim N(0, 1)$
(ii)	$T_r = \theta_0 + \theta_1 x_{1r} + \theta_2 x_{2r} + u_r,$ $C_r = \gamma_0 + \gamma_1 x_{1r} + \gamma_2 x_{2r} + v_r.$	$u_r, v_r \sim t(3)$
(iii)	$T_r = \theta_0 + \theta_1 x_{1r} + \theta_2 x_{2r} + (\pi_0 + \pi_0 x_{1r} + \pi_2 x_{2r}) u_r,$ $C_r = \gamma_0 + \gamma_1 x_{1r} + \gamma_2 x_{2r} + (\pi_0 + \pi_0 x_{1r} + \pi_2 x_{2r}) v_r.$	$u_r, v_r \sim N(0, 1)$
(iv)	$T_r = \theta_0 + \theta_1 x_{1r} + \theta_2 x_{2r} + (\pi_0 + \pi_0 x_{1r} + \pi_2 x_{2r}) u_r,$ $C_r = \gamma_0 + \gamma_1 x_{1r} + \gamma_2 x_{2r} + (\pi_0 + \pi_0 x_{1r} + \pi_2 x_{2r}) v_r.$	$u_r, v_r \sim t(3)$

In Cases (i) and (ii), event times and censoring times are generated from the homoscedastic models and in Cases (iii) and (iv), we considered heteroscedastic models to examine the efficiency gain of our proposed method over the standard censored quantile regression. We set the parameter values as $\theta^\top = (0, -1, 0.2)$, $\pi^\top = (0.3, -0.1, 0.1)$ and selected γ^\top to maintain approximately 30% of the censoring proportion in each case. We generated explanatory variables from zero mean bivariate normal distribution with covariance,

$$\Sigma = \begin{bmatrix} 1 & \sigma_{x_1, x_2} \\ \sigma_{x_1, x_2} & 1 \end{bmatrix}.$$

We considered two different ways to compute the EL based probability weights. In numerical study -I, we compute P_i 's based on the auxiliary information related to the failure time, T_i , whereas in numerical study -II, P_i 's are computed using the observed survival time, $Y_i = \min(T_i, C_i)$. In numerical study -II, we employ the synthetic variable approach (Koul et al., 1981; Qin and Jing, 2001; Li and Wang, 2003) to compute the EL based data driven probability weights.

4.1 Numerical study -I

To compute P_i 's, first we need to have a known population parameter, θ , or its estimate. We considered a linear relation between T and $\mathbf{X} = (X_1, X_2)$ with slopes (θ_1 and θ_2) and intercept (θ_0) as the auxiliary information. We estimated θ using the standard linear regression (least square) based on a large, finite

population with size, $N = 10,000$. We need to generate censoring times as well to compute the event indicator, $\delta_i = \mathbb{I}(T_i \leq C_i)$ and survival time, $Y_i = \min(T_i, C_i)$ to estimate the censored quantile regression parameters. To fit the weighted censored quantile regression model given in (1.2), we generated another n observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$ with $n \ll N$, using the same models given in Table 4.1. We considered the sample sizes, $n = 100$ and 200 and three quantiles, $\tau = 0.25, 0.50, 0.75$. For our proposed method, we estimated P_i 's using the estimating function, $g(t_i, \mathbf{x}_i; \boldsymbol{\theta})$ defined based on the normal equations of the linear least squares method as,

$$g_i(\mathbf{z}_i; \boldsymbol{\theta}) = g(t_i, \mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{x}_i(t_i - \mathbf{x}_i^\top \hat{\boldsymbol{\theta}}), \quad i = 1, 2, \dots, n. \quad (4.1)$$

For a given quantile, τ , the true value of the censored quantile regression parameters $\boldsymbol{\beta}_0(\tau)$ are estimated from the population of size, $N = 10,000$. In general, under a linear model assumption, the true value of the censored quantile regression slope parameters are the same as the $\boldsymbol{\theta}$ (i.e., $\beta_1(\tau) = \theta_1$, $\beta_2(\tau) = \theta_2$). But for the intercept, it is $\beta_0(\tau) = \theta_0 + F^{-1}(\tau)$, where F is the error distribution. We conducted 1,000 simulations and computed mean bias, standard error (SE) and 95% coverage probability (CP) of the model parameter estimates for different sample sizes using 250 bootstrap samples. We compared the performance of our proposed method (CQR-EL1) with the standard censored quantile regression (CQR) model. We present the simulation results in Tables 4.2 to 4.5 respectively for Cases (i)-(iv) with $\sigma_{x_1, x_2} = 0$.

Table 4.2

Bias, SE and CP of regression parameters for Case (i) model with independent covariates ($\sigma_{x_1, x_2} = 0$)

	n	$\tau \rightarrow$	CQR			CQR-EL1		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0042	0.0170	0.0647	0.0027	0.0180	0.0771
		β_1	0.0029	0.0035	0.0094	-0.0014	-0.0048	0.0030
		β_2	-0.0049	-0.0141	-0.0100	-0.0047	-0.0124	-0.0171
	200	β_0	0.0218	0.0298	0.0501	0.0199	0.0322	0.0635
		β_1	0.0016	0.0026	0.0057	0.0008	0.0028	0.0048
		β_2	-0.0020	-0.0032	-0.0078	-0.0010	0.0001	-0.0071
SE	100	β_0	0.1449	0.1404	0.2268	0.1103	0.1086	0.2110
		β_1	0.1533	0.1515	0.2141	0.1159	0.1109	0.2000
		β_2	0.1519	0.1525	0.2198	0.1149	0.1109	0.2082
	200	β_0	0.0973	0.0929	0.1292	0.0720	0.0703	0.1221
		β_1	0.1040	0.1029	0.1341	0.0746	0.0718	0.1173
		β_2	0.1041	0.1027	0.1354	0.0752	0.0717	0.1177
CP	100	β_0	93.3	93.4	95.7	95.8	96.6	97.0
		β_1	94.7	95.8	96.5	95.1	96.2	97.9
		β_2	96.0	96.3	96.4	96.4	96.4	96.9
	200	β_0	92.3	91.9	92.7	92.7	92.5	94.8
		β_1	94.5	96.2	95.0	95.0	95.5	96.9
		β_2	93.6	95.0	95.2	94.2	94.9	95.8

Table 4.3**Bias, SE and CP of regression parameters for Case (ii) model with independent covariates ($\sigma_{x_1, x_2} = 0$)**

	n	$\tau \rightarrow$	CQR			CQR-EL1		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0105	0.0288	0.1088	0.0119	0.0270	0.1062
		β_1	0.0063	0.0214	0.0169	0.0005	0.0102	0.0066
		β_2	0.0164	0.0096	-0.0170	0.0152	0.0079	-0.0184
	200	β_0	0.0267	0.0355	0.0821	0.0276	0.0340	0.0760
		β_1	0.0006	-0.0032	0.0050	0.0042	0.0032	0.0024
		β_2	0.0112	0.0025	0.0051	0.0029	-0.0038	-0.0057
SE	100	β_0	0.1871	0.1538	0.2980	0.1522	0.1304	0.2914
		β_1	0.1946	0.1664	0.2698	0.1555	0.1318	0.2480
		β_2	0.1955	0.1676	0.2733	0.1556	0.1327	0.2543
	200	β_0	0.1235	0.1029	0.1621	0.0998	0.0871	0.1556
		β_1	0.1301	0.1146	0.1663	0.1010	0.0893	0.1473
		β_2	0.1315	0.1149	0.1671	0.1023	0.0897	0.1465
CP	100	β_0	95.5	93.1	94.7	96.2	94.8	97.2
		β_1	95.6	93.5	96.4	95.7	95.6	97.8
		β_2	95.9	95.4	96.4	96.0	95.0	97.2
	200	β_0	93.1	91.2	94.0	93.0	93.8	95.7
		β_1	95.0	95.5	95.4	94.8	95.5	96.2
		β_2	95.5	95.7	95.5	95.0	95.2	96.3

Table 4.4**Bias, SE and CP of regression parameters for Case (iii) model with independent covariates ($\sigma_{x_1, x_2} = 0$)**

	n	$\tau \rightarrow$	CQR			CQR-EL1		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0062	0.0088	0.0224	0.0055	0.0085	0.0254
		β_1	0.0042	0.0051	0.0076	0.0034	0.0016	0.0057
		β_2	-0.0038	-0.0039	-0.0069	-0.0013	0.0003	-0.0010
	200	β_0	0.0064	0.0072	0.0167	0.0064	0.0089	0.0195
		β_1	0.0012	0.0038	0.0033	-0.0006	-0.0003	-0.0014
		β_2	-0.0015	-0.0031	-0.0017	-0.0004	0.0002	0.0023
SE	100	β_0	0.0472	0.0466	0.0767	0.0349	0.0338	0.0737
		β_1	0.0566	0.0570	0.0796	0.0424	0.0411	0.0708
		β_2	0.0567	0.0575	0.0807	0.0425	0.0418	0.0720
	200	β_0	0.0313	0.0301	0.0402	0.0225	0.0213	0.0345
		β_1	0.0371	0.0377	0.0489	0.0276	0.0267	0.0402
		β_2	0.0367	0.0376	0.0488	0.0270	0.0267	0.0401
CP	100	β_0	94.4	95.0	96.1	94.3	96.0	97.1
		β_1	95.0	95.2	95.5	95.2	95.3	97.4
		β_2	96.6	96.7	97.3	95.4	96.6	98.0
	200	β_0	94.1	93.4	94.9	93.2	94.0	94.1
		β_1	94.0	94.9	96.0	93.0	95.1	95.9
		β_2	94.6	95.0	95.3	94.4	95.3	94.8

Table 4.5**Bias, SE and CP of regression parameters for Case (iv) model with independent covariates ($\sigma_{x_1, x_2} = 0$)**

	n	$\tau \rightarrow$	CQR			CQR-EL1		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0066	0.0097	0.0364	0.0048	0.0076	0.0273
		β_1	0.0031	0.0039	0.0041	0.0026	0.0043	0.0036
		β_2	0.0008	-0.0009	-0.0018	0.0008	-0.0035	-0.0028
	200	β_0	0.0083	0.0089	0.0243	0.0100	0.0103	0.0258
		β_1	-0.0020	0.0016	0.0017	-0.0022	-0.0008	-0.0018
		β_2	0.0008	-0.0012	-0.0031	0.0026	0.0012	0.0004
SE	100	β_0	0.0600	0.0507	0.1103	0.0466	0.0407	0.1038
		β_1	0.0667	0.0592	0.0993	0.0514	0.0468	0.0885
		β_2	0.0677	0.0600	0.1014	0.0525	0.0470	0.0921
	200	β_0	0.0395	0.0327	0.0521	0.0305	0.0260	0.0464
		β_1	0.0429	0.0386	0.0568	0.0331	0.0298	0.0491
		β_2	0.0429	0.0389	0.0580	0.0331	0.0301	0.0501
CP	100	β_0	93.5	95.0	97.7	94.7	95.5	97.8
		β_1	95.6	96.6	97.0	96.0	96.3	97.3
		β_2	96.0	96.2	97.3	95.8	96.7	97.0
	200	β_0	93.0	93.9	94.9	93.5	93.4	94.1
		β_1	95.6	95.8	94.7	94.5	95.2	95.4
		β_2	94.5	95.9	95.5	94.5	96.0	95.2

From Tables 4.2-4.5, we see that our proposed estimator has approximately zero bias. A comparison of SE of CQR-EL1 with CQR indicates that the SE of CQR-EL1 reduces remarkably for all the parameters irrespective of any quantile. For example, we consider the scenario of $n = 100$ and $\tau = 0.25$ for comparison purposes throughout this paper. From Table 4.2, for CQR, SE of $\hat{\beta}_1$ is 0.1533 and for CQR-EL1, SE of $\hat{\beta}_1$ is reduced to 0.1159. When the sample size is increased to 200, SE of $\hat{\beta}_1$ of our proposed method further is reduced to 0.0746. If we compare the CP of our proposed method with the nominal level of 95%, CQR-EL1 provides approximately 95% coverage and becomes more stable when the sample size increases. Similar conclusions can be reached for Case (ii) (results are in Table 4.3) even though we considered heavy tailed distribution for the failure time compared to Case (i). For example, SE of $\hat{\beta}_1$ using CQR is 0.1946, whereas it is only 0.1555 for the CQR-EL1 based estimate. We also observed that SE is comparatively high in Case (ii) compared to Case (i).

In Cases (iii) and (iv), the error depends on the covariates. Simulation results for these Cases (Tables 4.4 and 4.5) are almost similar to the cases where error is independent of covariates. For example, in Case (iii) (Table 4.4), SE of $\hat{\beta}_1$ is 0.0566 and 0.0424 for CQR and CQR-EL1 respectively. Similarly, in Case (iv) (Table 4.5), SE of $\hat{\beta}_1$ is 0.0667 and 0.0514 for CQR and CQR-EL1 respectively. Here, we could also see a slight increase in the SE of estimates for Case (iv) because of the heavy tailed distribution assumption for the failure time compared to Case (iii).

4.2 Numerical study -II

In most of the survival data with random right censoring, the observed data are the triplet $\{Y = \min(T, C), \mathbf{X}, \delta\}$. We consider a linear relationship between the survival time (Y) and the covariates as the auxiliary information. Here we cannot use the EL estimating function, $g(\cdot)$ defined in (4.1) because of the censoring. There are other methods available in the literature which take care of the right censoring in the linear regression.

Koul et al. (1981) introduced a synthetic data approach by transforming the survival time, Y_r to a synthetic variable, \tilde{Y}_r as

$$\tilde{Y}_r = \frac{\delta_r Y_r}{1 - G(Y_r)}; \quad r = 1, 2, \dots, N, \quad (4.2)$$

where δ_r is the censoring indicator and $G(\cdot)$ is the distribution of the censoring time. $E(\tilde{Y} | \mathbf{X}) = E(Y | \mathbf{X})$ if C is independent of both \mathbf{X} and Y . When $G(\cdot)$ is unknown, we can replace it with its Kaplan-Meier estimator. The estimator of $G(\cdot)$ using the Kaplan-Meier (Kaplan and Meier, 1958) estimator is

$$1 - \hat{G}_N(t) = \prod_{r=1}^N \left(\frac{N - r}{N - r + 1} \right)^{\mathbb{I}(Y_{(r)} \leq t, \delta_{(r)} = 0)}, \quad (4.3)$$

where $Y_{(r)}$'s are ordered and the corresponding censoring indicator is $\delta_{(r)}$. We can estimate $\boldsymbol{\theta}$ as

$$\tilde{\boldsymbol{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{Y}}. \quad (4.4)$$

Qin and Jing (2001) and Li and Wang (2003) independently provided the estimating function to compute the EL based data driven probabilities as

$$g_i(\mathbf{z}_i; \tilde{\boldsymbol{\theta}}) = g(y_i, \mathbf{x}_i, \delta_i; \tilde{\boldsymbol{\theta}}) = \mathbf{x}_i (\tilde{y}_i - \mathbf{x}_i^\top \tilde{\boldsymbol{\theta}}), \quad i = 1, 2, \dots, n. \quad (4.5)$$

We can compute the \tilde{y}_i and $\hat{G}_n(t)$ using the sample analogues of (4.2) and (4.3) respectively.

To compute P_i 's, we consider a linear relation between Y and $\mathbf{X} = (X_1, X_2)$ with slopes (θ_1 and θ_2) and intercept (θ_0). We estimate $\boldsymbol{\theta}$ using (4.4) based on a large, finite population with size, $N = 10,000$. To fit the weighted censored quantile regression model given in (1.2), we generate another n observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$ with $n \ll N$ using the same models given in Table 4.1. For our proposed method, we estimate P_i 's using the estimating function, $g(y_i, \mathbf{x}_i, \delta_i; \tilde{\boldsymbol{\theta}})$ given in (4.5).

Similar to numerical study -I, we present the results based on 1,000 simulations and report the bias, standard error (SE) and empirical coverage probability (CP) for the nominal level of 95% based on 250 bootstrap samples. We provide the summary of the simulation results for this study in Tables 4.6-4.9.

Similar to the population information related to T (numerical study -I), conclusions are almost similar for uncorrelated covariates. From Tables 4.6-4.9 we see that our proposed method (CQR-EL2) provides unbiased estimates irrespective of any sample size and quantile. If we consider the coverage probability,

both CQR and CQR-EL2 provide approximately 95% coverage. For any quantile, there is a reduction in the standard error of CQR-EL2 parameter estimates compared to CQR parameter estimates. If we consider Case (i) as a basic model, CQR-EL2 with Case (ii) has reasonably higher SE along with CQR because of the heavy tailed distribution of the observed survival time. When the error depended on the covariates (Cases (iii) & (iv)), the SE of CQR-EL2 reduced considerably.

We also conducted large number of simulations with correlated covariates with $\sigma_{x_1, x_2} = 0.5$ as well as constructed weights based on simple relationship with one covariate only for both numerical studies. The results of these simulations are not provided here to save the space. The conclusions arrived are almost similar to the uncorrelated covariate cases.

In numerical study -I, we noticed that there is a slight reduction in SE of $\hat{\beta}_2$ using heteroscedastic models for CQR-EL1. But use of the estimating function, $g(y_i, x_{1i}, \delta_i; \tilde{\theta})$ (CQR-EL2), does not reduce the SE of $\hat{\beta}_2$ under heteroscedastic models. Since we utilized only partial population information in relation to X_1 , the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$ reduced for CQR-EL2 compared to CQR. The standard error of $\hat{\beta}_2$ was not changed.

Our simulation studies reveal that auxiliary information greatly enhances the efficiency of estimation, if the population information related to both X_1 and X_2 is available. If the population information is only related to X_1 , the efficiency gain is limited to β_0 and β_1 only. However, under heteroscedastic models, the efficiency of estimating β_2 slightly improved in numerical study -I, but not in numerical study -II.

Table 4.6

Bias, SE and CP of regression parameters for Case (i) model with independent covariates ($\sigma_{x_1, x_2} = 0$)

	n	$\tau \rightarrow$	CQR			CQR-EL2		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0042	0.0170	0.0647	0.0217	0.0275	0.0720
		β_1	0.0029	0.0035	0.0094	-0.0491	-0.0411	-0.0090
		β_2	-0.0049	-0.0141	-0.0100	0.0116	-0.0029	-0.0194
	200	β_0	0.0218	0.0298	0.0501	0.0220	0.0323	0.0562
		β_1	0.0016	0.0026	0.0057	-0.0295	-0.0273	-0.0119
		β_2	-0.0020	-0.0032	-0.0078	0.0034	0.0053	-0.0011
SE	100	β_0	0.1449	0.1404	0.2268	0.1273	0.1233	0.2160
		β_1	0.1533	0.1515	0.2141	0.1475	0.1416	0.2075
		β_2	0.1519	0.1525	0.2198	0.1416	0.1414	0.2162
	200	β_0	0.0973	0.0929	0.1292	0.0840	0.0798	0.1239
		β_1	0.1040	0.1029	0.1341	0.0970	0.0921	0.1278
		β_2	0.1041	0.1027	0.1354	0.0957	0.0936	0.1304
CP	100	β_0	93.3	93.4	95.7	94.3	96.1	96.8
		β_1	94.7	95.8	96.5	94.6	96.1	96.9
		β_2	96.0	96.3	96.4	95.4	95.4	97.4
	200	β_0	92.3	91.9	92.7	92.9	92.3	94.3
		β_1	94.5	96.2	95.0	95.3	95.3	94.8
		β_2	93.6	95.0	95.2	93.5	94.9	95.9

Table 4.7**Bias, SE and CP of regression parameters for Case (ii) model with independent covariates ($\sigma_{x_1, x_2} = 0$)**

	n	$\tau \rightarrow$	CQR			CQR-EL2		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0105	0.0288	0.1088	0.0306	0.0461	0.1139
		β_1	0.0063	0.0214	0.0169	-0.0841	-0.0503	-0.0216
		β_2	0.0164	0.0096	-0.0170	0.0329	0.0260	-0.0094
	200	β_0	0.0267	0.0355	0.0821	0.0419	0.0508	0.0921
		β_1	0.0006	-0.0032	0.0050	-0.0022	-0.0010	-0.0188
		β_2	0.0112	0.0025	0.0051	0.0251	0.0137	0.0133
SE	100	β_0	0.1871	0.1538	0.2980	0.1619	0.1379	0.2768
		β_1	0.1946	0.1664	0.2698	0.1863	0.1595	0.2548
		β_2	0.1955	0.1676	0.2733	0.1787	0.1549	0.2632
	200	β_0	0.1235	0.1029	0.1621	0.1048	0.0900	0.1551
		β_1	0.1301	0.1146	0.1663	0.1214	0.1052	0.1575
		β_2	0.1315	0.1149	0.1671	0.1185	0.1044	0.1606
CP	100	β_0	95.5	93.1	94.7	95.9	94.2	97.5
		β_1	95.6	93.5	96.4	94.8	93.3	96.7
		β_2	95.9	95.4	96.4	94.2	94.2	96.3
	200	β_0	93.1	91.2	94.0	93.5	93.0	94.7
		β_1	95.0	95.5	95.4	94.5	94.0	94.9
		β_2	95.5	95.7	95.5	94.8	94.5	95.4

Table 4.8**Bias, SE and CP of regression parameters for Case (iii) model with independent covariates ($\sigma_{x_1, x_2} = 0$)**

	n	$\tau \rightarrow$	CQR			CQR-EL2		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0062	0.0088	0.0224	0.0127	0.0146	0.0302
		β_1	0.0042	0.0051	0.0076	-0.0071	-0.0043	0.0021
		β_2	-0.0038	-0.0039	-0.0069	0.0018	0.0017	-0.0040
	200	β_0	0.0064	0.0072	0.0167	0.0094	0.0105	0.0197
		β_1	0.0012	0.0038	0.0033	-0.0042	-0.0026	-0.0007
		β_2	-0.0015	-0.0031	-0.0017	0.0009	-0.0003	0.0015
SE	100	β_0	0.0472	0.0466	0.0767	0.0448	0.0445	0.0801
		β_1	0.0566	0.0570	0.0796	0.0541	0.0549	0.0830
		β_2	0.0567	0.0575	0.0807	0.0538	0.0558	0.0833
	200	β_0	0.0313	0.0301	0.0402	0.0292	0.0283	0.0396
		β_1	0.0371	0.0377	0.0489	0.0348	0.0356	0.0484
		β_2	0.0367	0.0376	0.0488	0.0344	0.0359	0.0488
CP	100	β_0	94.4	95.0	96.1	93.9	94.7	96.9
		β_1	95.0	95.2	95.5	94.6	94.7	96.3
		β_2	96.6	96.7	97.3	95.8	96.4	97.3
	200	β_0	94.1	93.4	94.9	93.9	93.8	94.9
		β_1	94.0	94.9	96.0	94.1	94.3	95.0
		β_2	94.6	95.0	95.3	94.0	95.4	94.3

Table 4.9**Bias, SE and CP of regression parameters for Case (iv) model with independent covariates ($\sigma_{x_1, x_2} = 0$)**

	n	$\tau \rightarrow$	CQR			CQR-EL2		
			0.25	0.50	0.75	0.25	0.50	0.75
Bias	100	β_0	0.0066	0.0097	0.0364	0.0189	0.0169	0.0419
		β_1	0.0031	0.0039	0.0041	-0.0138	-0.0073	-0.0000
		β_2	0.0008	-0.0009	-0.0018	0.0074	0.0060	0.0024
	200	β_0	0.0083	0.0089	0.0243	0.0124	0.0119	0.0273
		β_1	-0.0020	0.0016	0.0017	-0.0097	-0.0051	-0.0032
		β_2	0.0008	-0.0012	-0.0031	0.0019	0.0004	-0.0020
SE	100	β_0	0.0600	0.0507	0.1103	0.0548	0.0486	0.1159
		β_1	0.0667	0.0592	0.0993	0.0618	0.0581	0.1018
		β_2	0.0677	0.0600	0.1014	0.0616	0.0578	0.1066
	200	β_0	0.0395	0.0327	0.0521	0.0359	0.0304	0.0516
		β_1	0.0429	0.0386	0.0568	0.0397	0.0364	0.0558
		β_2	0.0429	0.0389	0.0580	0.0397	0.0368	0.0579
CP	100	β_0	93.5	95.0	97.7	92.9	95.2	97.6
		β_1	95.6	96.6	97.0	94.2	95.5	97.4
		β_2	96.0	96.2	97.3	96.3	97.0	97.6
	200	β_0	93.0	93.9	94.9	93.3	94.2	95.8
		β_1	95.6	95.8	94.7	94.0	95.5	95.2
		β_2	94.5	95.9	95.5	94.9	96.0	94.7

Note that the value of the auxiliary parameter value plays a big role in the efficiency of the weighted censored quantile regression parameter estimates. If the estimate of θ based present study data and previous study (or known θ value) are very close, then all weights will be close to $1/n$ and solutions to (1.1) and (1.2) remain the same. If data on previous studies are not available, we can make of the data available in the present study to estimate the value of θ . In this case, if dimensions of θ and estimating equation $g(z, \theta)$ are same, then all weights will be equal to $1/n$ and solutions to (1.1) and (1.2) remain same. However, if the dimensions of $g(z, \theta)$ is greater than that of θ , the weights $p(\hat{\theta})$ is no longer equal to $1/n$ and this scheme provides an efficiency gain over the conventional QR estimates (Tang and Leng, 2012).

4.3 Case example

The North Central Cancer Treatment Group (NCCTG) was initiated by a group of physicians from the north central region of the United States of America and the Mayo Clinic in Rochester, Minnesota. This study was conducted by NCCTG to determine whether the conclusions from the patient-completed questionnaire and those already obtained by the patient's physician were independent or not (Loprinzi, Laurie, Wieand, Krook, Novotny, Kugler, Bartel, Law, Bateman and Klatt, 1994). They used the performance scores (ECOG and Karnofsky) to assess the patient's daily activities. The dataset is available in the "survival" package of R software with readings of 228 patients. Because of the incompleteness of some of the variables, we had to limit the dataset to 167 observations. For the illustration of our proposed method, we changed our focus to identify the effect of following covariates over the observed survival time at different quantiles. We considered "age", patient's age in years; "sex", (Male = 1 Female = 2); "ph.ecog",

ECOG performance score measured by physician (0 = good 5 = dead); “meal.cal”, calories consumed at meals and “wt.loss”, weight loss in the last six months as the covariates. After removing the incomplete patient readings, the available ECOG scores were 0,1 and 2 only. We defined two dummy categorical variables for “ph.ecog” as follows.

$$\text{ecog1} = \begin{cases} 1, & \text{if ph.ecog} = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{ecog2} = \begin{cases} 1, & \text{if ph.ecog} = 2 \\ 0, & \text{otherwise.} \end{cases}$$

To demonstrate the usefulness of our proposed method, we randomly selected a part (100 observations) of the complete data (167 observations) by considering it to be the data available from the previous study. We assumed that there exists a linear relation between the logarithm of the observed survival time and all the continuous explanatory variables (age, meal.cal and wt.loss) as the available auxiliary information. We estimated the $\theta = (\theta_0, \theta_{\text{age}}, \theta_{\text{meal}}, \theta_{\text{wt}})$ by the least square method based on 100 observations where the response is the synthetic variable defined by (4.2). Then we computed the EL based data driven probability weights for the present study data points (67 observations). After computing the weights, we estimated the weighted censored quantile regression parameters using Peng and Huang (2008) method with all the covariates. For the present study data, the censoring proportion is 0.283. Interestingly, we estimated the regression parameters using CQR up to the 86th quantile, where as we could estimate to the 90th quantile using CQR-EL2. Along with the estimates for the quantiles, $\tau = 0.25, 0.50, 0.75$, we report standard error (SE) and 95% confidence limits using 250 bootstrap samples as well in Table 4.10.

Table 4.10
Estimates, SE and 95% CI for regression parameters of NCCTG lung cancer data

	$\tau \rightarrow$	CQR			CQR-EL2		
		0.25	0.50	0.75	0.25	0.50	0.75
$\hat{\beta}$	Intercept	5.4777	4.2651	5.5380	4.7531	4.1729	6.4258
	Age	-0.0168	0.0179	0.0040	-0.0047	0.0202	-0.0032
	Sex	0.7201	0.6180	0.4181	0.7606	0.6638	0.3651
	ECOG1	-0.7059	-0.5449	-0.2029	-0.5701	-0.5355	-0.2884
	ECOG2	-0.8677	-0.9402	-0.8336	-1.1584	-1.0612	-1.0192
	MealCal	0.0004	0.0001	0.0001	0.0004	0.0001	-0.0000
	WtLoss	-0.0007	-0.0084	-0.0023	-0.0023	-0.0100	-0.0135
SE	Intercept	1.9235	1.4314	1.7494	1.6628	1.4149	1.4666
	Age	0.0277	0.0188	0.0225	0.0256	0.0184	0.0176
	Sex	0.5610	0.3389	0.3716	0.5374	0.3317	0.2809
	ECOG1	0.6521	0.3436	0.3375	0.6498	0.3493	0.2434
	ECOG2	1.0317	0.5410	0.6061	0.9336	0.5413	0.3879
	MealCal	0.0009	0.0006	0.0008	0.0009	0.0006	0.0005
	WtLoss	0.0181	0.0128	0.0231	0.0157	0.0124	0.0100
CI	Intercept	(1.6, 9.14)	(2.38, 8)	(2.08, 8.94)	(1.79, 8.31)	(2.32, 7.87)	(3.14, 8.89)
	Age	(-0.07, 0.04)	(-0.04, 0.04)	(-0.04, 0.05)	(-0.06, 0.04)	(-0.03, 0.04)	(-0.03, 0.04)
	Sex	(-0.45, 1.74)	(0, 1.33)	(-0.13, 1.33)	(-0.39, 1.71)	(-0.04, 1.27)	(-0.07, 1.03)
	ECOG1	(-1.75, 0.81)	(-1.15, 0.2)	(-0.97, 0.35)	(-1.86, 0.69)	(-1.18, 0.19)	(-0.78, 0.18)
	ECOG2	(-2.88, 1.16)	(-2, 0.12)	(-2.11, 0.26)	(-2.83, 0.83)	(-2.13, -0.01)	(-1.73, -0.21)
	MealCal	(-0.04, 0.03)	(-0.03, 0.02)	(-0.05, 0.04)	(-0.04, 0.02)	(-0.03, 0.01)	(-0.04, 0)
	WtLoss	(-0.04, 0.03)	(-0.03, 0.02)	(-0.05, 0.04)	(-0.04, 0.02)	(-0.03, 0.01)	(-0.04, 0)

From Table 4.10, we see that the standard error of the estimates of all the continuous variable parameters and the intercept reduced considerably because we considered the auxiliary information related to them. For the remaining variables, a reduction of standard error can also be seen, even though we did not consider any auxiliary information related to them. In the censored quantile regression with the EL based data driven probability weights, we see narrower 95% confidence limits for all the variables compared to those using the standard censored quantile regression.

5 Conclusions

We proposed a method which effectively use the auxiliary information to improve the efficiency of the censored quantile regression estimator. We developed a methodology to transform the population information available from previous clinical trials or from some existing facts into non-parametric empirical likelihood based data driven probabilities. We developed the EL based data driven probability computation for both known and unknown cases of prior information regarding population parameters. We applied these probabilities as the weights into Peng and Huang (2008) censored quantile regression model. Our proposed method is efficient compared to standard censored quantile regression and provides consistent estimators of regression coefficients with asymptotic normality. Our simulations studies showed that the standard error of the parameter estimates based on our proposed methods (CQR-EL1 and CQR-EL2) is lower than the standard method (CQR) when we use all the covariates for computing the EL based data driven probability weights. Our proposed weighted censored quantile regression method provides almost the same coverage probability compared to the nominal level. In the case of heteroscedastic models, even the use of the auxiliary information regarding a subset of population parameters improved the efficiency of the estimates of all the parameters by using CQR-EL1. But in CQR-EL2, the efficiency improvement was limited to the corresponding subset of variables and intercept. In homoscedastic models, the use of auxiliary information regarding a subset of population parameters improved the efficiency only for that particular subset of parameters and the intercept in both CQR-EL1 and CQR-EL2. In the real data analysis, we observed that our proposed method provides more efficient quantile estimates and narrower confidence limits compared to the standard censored quantile regression.

Acknowledgements

The authors thank the Editor, Associate Editor, and referees whose suggestions greatly contributed to improving this paper. The research of Variyath and Fan are partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

- Bahadur, R.R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37, 577-580.

- Buckley, J., and James, I. (1979). Linear regression with censored data. *Biometrika*, 66, 429-436.
- Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.
- Fang, K.-T., Li, G., Lu, X. and Qin, H. (2013). An empirical likelihood method for semiparametric linear regression with right censored data. *Computational and Mathematical Methods in Medicine*, 1-9.
- Fleming, T.R., and Harrington, D.P. (2011). *Counting Processes and Survival Analysis*, New York: John Wiley & Sons, Inc.
- Fygenson, M., and Ritov, Y. (1994). Monotone estimating equations for censored data. *The Annals of Statistics*, 22, 732-746.
- Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481.
- Koenker, R. (2005). *Quantile Regression*, Cambridge.
- Koul, H., Susarla, V. and Ryzin, J.V. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, 9, 1276-1288.
- Kuk, A.Y.C., and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of Royal Statistical Society, Series B*, 51, 261-269.
- Li, G., and Wang, Q.H. (2003). Empirical likelihood regression analysis for right censored data. *Statistica Sinica*, 13, 51-68.
- Loprinzi, C.L., Laurie, J.A., Wieand, H.S., Krook, J.E., Novotny, P.J., Kugler, J.W., Bartel, J., Law, M., Bateman, M. and Klatt, N.E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, 12, 1, 601-607.
- Owen, A. (1988). Empirical likelihood ratio confidence interval for a single functional. *Biometrika*, 75, 237-249.
- Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19, 1725-1747.
- Owen, A. (2001). *Empirical Likelihood*, Chapman & Hall/CRC.
- Peng, L., and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103, 637-649.
- Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98, 1001-1012.
- Powell, J. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25, 303-325.
- Powell, J. (1986). Censored regression quantiles. *Journal of Econometrics*, 32, 143-155.
- Qin, G., and Jing, B.-Y. (2001). Empirical likelihood for censored linear regression. *Scandinavian Journal of Statistics*, 28, 661-673.
- Qin, J., and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300-325.

- Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- Tang, C., and Leng, C. (2012). An empirical likelihood approach to quantile regression with auxiliary information. *Statistics & Probability Letters*, 82, 29-36.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24-36.