Censored Quantile Regression with Covariate Measurement Errors

Yuanshan Wu[1], Yanyuan Ma[2], and Guosheng Yin[3]
[1]Wuhan University, Wuhan, Hubei 430072, China
[2]Texas A&M University, College Station, Texas 77843, U.S.A.
[3]The University of Hong Kong Pokfulam Road, Hong Kong
Corresponding author: Guosheng Yin, email: gyin@hku.hk

Abstract

Censored quantile regression has become an important alternative to the Cox proportional hazards model in survival analysis. In contrast to the central covariate effect from the mean-based hazard regression, quantile regression can effectively characterize the covariate effects at different quantiles of the survival time. When covariates are measured with errors, it is known that naively treating mismeasured covariates as error-free would result in estimation bias. Under censored quantile regression, we propose corrected estimating equations to obtain consistent estimators. We establish consistency and asymptotic normality for the proposed estimators of quantile regression coefficients. Compared with the naive estimator, the proposed method can eliminate the estimation bias under various measurement error distributions and model error distributions. We conduct simulation studies to examine the finite-sample properties of the new method and apply our model to a lung cancer study.

Keywords: Check function; Corrected estimating equations; Measurement errors; Kernel smoothing.

1 Introduction

Mean-based regression models have been extensively studied for randomly censored survival data. For example, the Cox (1972) proportional hazards model characterizes the hazard as a function of different covariates; and the accelerated failure time (AFT) model directly formulates linear regression between the logarithm of the failure time and covariates. However, neither the Cox nor the AFT model can differentiate the covariate effect at higher or lower quantiles of survival times, as they only provide the mean effect. In survival analysis with random censoring, censored quantile regression (CQR) has been proposed and is gaining much popularity (Ying, Jung, and Wei, 1995; Lindgren, 1997; Yang, 1999; Koenker and Geling, 2001; Bang and Tsiatis, 2002; Chernozhukov and Hong, 2002; Portnoy, 2003; Peng and Huang, 2008; and Wang and Wang, 2009). In practice, covariates are often subject to measurement errors. The most common measurement error structure is $\mathbf{W} = \mathbf{Z} + \mathbf{U}$, where

$\mathbf{W}$ is the observed surrogate, $\mathbf{Z}$ is the true but unobserved covariate, and $\mathbf{U}$ is the random measurement error. For a comprehensive coverage of various measurement error models and inference procedures with mean-based regression, see Carroll et al. (2006). In the context of quantile regression with measurement errors, Brown (1982) examined median regression and described the difficulty involved in parameter estimation. He and Liang (2000) proposed root-n consistent estimators for linear and partially linear quantile regression models. Their method assumes that the random error in the response and the measurement errors in the covariates follow a spherical symmetric distribution. Wei and Carroll (2009) proposed a novel approach to quantile regression with measurement errors by utilizing the derivative property of the quantile function when the same quantile regression structure is assumed for all the quantile levels. Recently, Wang, Stefanski, and Zhu (2012) developed a corrected-loss function for the smoothed check function, a substantial advance in this area. However, there is limited research on quantile regression with covariate measurement errors under censoring.

2 CQR Model with Measurement Errors

Let T denote the transformed failure time under a known monotone transformation, e.g., the logarithm function. Let C denote the censoring time under the same transformation. Let $\mathbf{Z}$ be a p-vector of covariates, $X = T \wedge C$ be the observed time, and $\Delta = I(T \leq C)$ be the censoring indicator. For $\tau \in (0, 1)$, the conditional τth quantile function of survival time T given covariate $\mathbf{Z}$ is defined as $Q_T(\tau|\mathbf{Z}) = \inf\{t\colon P(T \leq t|\mathbf{Z}) \geq \tau\}$. The quantile regression model associated with covariate $\mathbf{Z}$ has the form

$$Q_T(\tau|\mathbf{Z}) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}(\tau), \tag{2.1}$$

where $\boldsymbol{\beta}(\tau)$ is an unknown p-vector of regression coefficients, representing the effect of $\mathbf{Z}$ on the τth quantile of the transformed survival time.

In reality covariate $\mathbf{Z}$ may be measured with errors, so that we do not directly observe $\mathbf{Z}$ but its surrogate $\mathbf{W}$. We assume the classical error structure

$$\mathbf{W} = \mathbf{Z} + \mathbf{U},$$

where $\mathbf{U}$ is a p-variate random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

We first introduce notation: $F_T(t|\mathbf{Z}) = P(T \leq t|\mathbf{Z})$, $\Lambda_T(t|\mathbf{Z}) = -\log\{1 - P(T \leq t|\mathbf{Z})\}$, $N(t) = \Delta I(X \leq t)$ and $M(t) = N(t) - \Lambda_T(t \wedge X|\mathbf{Z})$. Following the argument in Fleming and Harrington (1991), it is easy to show that evaluated at $\boldsymbol{\beta}_0(\tau)$, the true value of $\boldsymbol{\beta}(\tau)$, $M(t)$ is a martingale process associated with the counting process $N(t)$. Furthermore, because $E\{M(t)|\mathbf{Z}\} = 0$ at $\boldsymbol{\beta}_0(\tau)$ for $t \geq 0$, we have

$$E\left(\mathbf{Z}\left[N\{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_0(\tau)\} - \Lambda_T(\{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_0(\tau)\} \wedge X|\mathbf{Z})\right]\right) = \mathbf{0} \tag{2.2}$$

for $\tau \in (0,1)$. Under model (2.1), after some algebraic manipulations, we obtain that

$$\Lambda_T(\{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_0(\tau)\} \wedge X|\mathbf{Z}) \;=\; \int_0^\tau I\{X \geq \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_0(u)\}\mathrm{d}H(u), \tag{2.3}$$

where $H(u) = -\log(1-u)$ for $0 \leq u < 1$.

Based on (2.2) and (2.3), when all $\mathbf{Z}_i$'s are observed, Peng and Huang (2008) proposed an estimating equation for $\{\boldsymbol{\beta}(\tau)\colon \tau \in (0,1)\}$,

$$\mathbf{0} \;=\; n^{-1}\sum_{i=1}^n \left[\Delta_i\mathbf{Z}_i - \Delta_i\mathbf{Z}_i I\{X_i > \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)\} - \int_0^\tau \mathbf{Z}_i I\{X_i > \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}(u)\}\mathrm{d}H(u)\right]. \tag{2.4}$$

We denote the observed data $\mathcal{O} \equiv (X, \Delta, \mathbf{W})$ and the unobserved data $\mathcal{U} \equiv (X, \Delta, \mathbf{Z})$. In view of the estimating equation (2.4), if we can find some function $g^*\{\mathcal{O}, \boldsymbol{\beta}(\tau)\}$ such that for $\tau \in (0,1)$,

$$E[g^*\{\mathcal{O}, \boldsymbol{\beta}(\tau)\}|\mathcal{U}] = \mathbf{Z}I\{X > \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}(\tau)\},$$

following the corrected score argument (Stefanski, 1989; Nakamura, 1990), we can construct the (corrected) unbiased estimating equation as

$$\mathbf{0} = n^{-1}\sum_{i=1}^n \left[\Delta_i\mathbf{W}_i - \Delta_i g^*\{\mathcal{O}_i, \boldsymbol{\beta}(\tau)\} - \int_0^\tau g^*\{\mathcal{O}_i, \boldsymbol{\beta}(u)\}\mathrm{d}H(u)\right].$$

However, the cusp in the indicator function makes it difficult to find such a function. Assume that a smooth function $K(\cdot)$ satisfies $\lim_{x\to-\infty} K(x) = 0$ and $\lim_{x\to\infty} K(x) = 1$. If we consider a positive scale parameter h_n that converges to zero as sample size $n \to \infty$, $K(x/h_n)$ may provide an adequate approximation to $I(x > 0)$ as $n \to \infty$, where h_n behaves like the bandwidth in the kernel smoothing.

As a result, we aim to find a function $G\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\}$ such that

$$E[G\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\}|\mathcal{U}] \;\approx\; \{X - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}(\tau)\}I\{X > \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}(\tau)\}.$$

If there exists such a function $G\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\}$, we may set

$$g\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\} = -\frac{\partial G\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\}}{\partial \boldsymbol{\beta}(\tau)},$$

and conclude that $E[g\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\}|\mathcal{U}]$ is close to $\mathbf{Z}I\{X > \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}(\tau)\}$. Thus, we can construct an approximately corrected estimating equation

$$\mathbb{P}_n\Delta\bar{g}\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\} - \int_0^\tau \mathbb{P}_n g\{\mathcal{O}, \boldsymbol{\beta}(u); h_n\}\mathrm{d}H(u) = \mathbf{0}, \tag{2.5}$$

where $\mathbb{P}_n$ is the empirical measure with respect to $\mathcal{O}$, and $\bar{g}\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\} = \mathbf{W} - g\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\}$. Noting that it is very challenging to obtain the functional solution to the integral equation

(2.5), we propose a grid-based estimation procedure for $\boldsymbol{\beta}_0(\cdot)$. Assume that τ_U is a deterministic constant in $(0,1)$ subject to certain identifiability constraints (see the **??**). Due to the inherent nonidentifiability of the regression quantiles beyond the τ_U-th level, we confine our estimation for $\boldsymbol{\beta}_0(\tau)$ with $\tau \in (0, \tau_U]$. We denote a partition over the interval $[0, \tau_U]$ by $\mathcal{S}_{q_n} = \{0 \equiv \tau_0 < \tau_1 < \cdots < \tau_{q_n} \equiv \tau_U\}$, where the number of grid points q_n depends on n. We consider an estimator of $\boldsymbol{\beta}_0(\tau)$ that is a right-continuous piecewise constant function and jumps only at grid points in $\mathcal{S}_{q_n}$. Noting that $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_0(\tau_0) = -\infty$, we intuitively set $g\{\mathcal{O}, \widehat{\boldsymbol{\beta}}(\tau_0); h_n\} = \mathbf{W}$. For a given h_n, employing the Newton–Raphson algorithm, the estimates $\widehat{\boldsymbol{\beta}}(\tau_j),\ j = 1, \ldots, q_n$, can be obtained sequentially by solving

$$\mathbb{P}_n \Delta \bar{g}\{\mathcal{O}, \boldsymbol{\beta}(\tau); h_n\} - \sum_{k=0}^{j-1} \mathbb{P}_n g\{\mathcal{O}, \widehat{\boldsymbol{\beta}}_n(\tau_k); h_n\}\{H(\tau_{k+1}) - H(\tau_k)\} = \mathbf{0}. \tag{2.6}$$

Denote $a_n = \max_{1 \le j \le q_n} |\tau_j - \tau_{j-1}|$, the maximum distance between two adjacent points belonging to $\mathcal{S}_{q_n}$. The asymptotic properties of the resultant estimator $\widehat{\boldsymbol{\beta}}(\tau)$ are summarized in the following two theorems.

Theorem 1 *If $a_n = o(1)$, then $\sup_{\tau \in [\nu, \tau_U]} \|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\| \to 0$ in probability for any $\nu \in (0, \tau_U]$ as $n \to \infty$.*

Theorem 2 *If $a_n = o(n^{-1/2})$, then $n^{1/2}\{\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\}$ converges weakly to a mean zero Gaussian random field over $\tau \in [\nu, \tau_U]$ for any $\nu \in (0, \tau_U]$ as $n \to \infty$.*

3 Remarks

We have proposed a corrected estimating equation approach to CQR models for survival data when covariates are measured with errors. Using a smooth function to approximate the indicator function, the resultant estimating function becomes smooth, and thus the root-finding procedure can be carried out by the conventional iterative methods such as the Newton–Raphson algorithm.

References

Brown, M. L. (1982). Robust line estimation with error in both variables. *Journal of the American Statistical Association* **77**, 71–79. Correction in **78**, 1008.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. London: CRC Press.

He, X. and Liang, H. (2000). Quantile regression estimates for a class of linear and partially linear errors-in-variables models. *Statistica Sinica* **10**, 129–140.

Hong, H. and Tamer, E. (2003). A simple estimator for nonlinear error in variable models. *Journal of Econometrics* **117**, 1–19.

Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60**, 505–531.

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

Koenker, R. and Bassett, G. J. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association* **103**, 637–649.

Wang, H., Stefanski, L. A., and Zhu, Z. (2012). Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika* **99**, 405–421.