Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp. 31 38

Regression Quantile Estimations on Censored Survival Data

Joo Yong Shim[1)]

Abstract

In the case of multiple survival times which might be censored at each covariate vector, we study the regression quantile estimations in this paper. The estimations are based on the empirical distribution functins of the censored times and the sample quantiles of the observed survival times at each covariate vector and the weighted least square method is applied for the estimation of the regression quantile. The estimators are shown to be asymptotically normally distributed under some regularity conditions.

Keywords and *Phrases* : Censoring, Covariate, Emipirical distribution function, Quantile regression model, Sample quantile, Survival data

1. Introductions

The accelerated failure time model in the survival data analysis regresses the logarithms of the survival times on the corresponding covariate vector, which is appealing to the practitioners due to the direct physical interpretations. The mean of the logarithms of survival times are related to the covariate in the accelerated failure time model, which causes difficulty to estimate the mean of survival times or the intercept parameter. Ying(1995) proposed a median regression model as an alternative to the mean regression model for examining the covariate effect on the survival pattern. Yang(1999) proposed a median regression estimator which are based on the weighted empirical survival and hazard functions without estimating the distributions of the censored times, and showed that the estimators are consistent and asymptoticcally

1. Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea

distributed.

The median is a simple and meaningful measure of the center of the thick tailed distribution of the survival times, and can be well estimated even for not too heavy cen soring. But usually large or small quantile depends on the covariate differently from the median, which leads to consider the quantile regression approach. Koenker and Bassett(1978) introduced the quantile regression model, that the quantiles of responses are linearly related to the covariate vector as follows

$$Q(\theta \mid \boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta}(\theta) \quad \text{for} \quad \theta \in (0, 1),$$

where $Q(\theta \mid \boldsymbol{x}_i)$ is the θ-th quantile of the responses given the covariate vector $\boldsymbol{x}_i$. The estimator of the θ-th regression quantile $\boldsymbol{\beta}(\theta)$ is defined as the value of $\boldsymbol{\beta}$ minimizing the objective function,

$$R_\theta(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\theta(y_i - \boldsymbol{x}_i \boldsymbol{\beta}) \quad \text{for} \quad \theta \in (0, 1),$$

over all $\boldsymbol{\beta}$ in some parameter space $B(\theta)$ where $\rho_\theta(\cdot)$ is the check function defined as

$$\rho_\theta(r) = \theta r I(r \geq 0) + (\theta - 1) r I(r < 0),$$

where $I(\cdot)$ is the indicator function. The median estimator is easily seen to be a special case of $\theta = 1/2$. Powell(1986) studied the censored quantile regression, where observations could not be observed below the fixed level 0 in the regression model. The censored regression quantile estimator is defined as the value of $\boldsymbol{\beta}$ minimizing the objective function,

$$Q_n(\boldsymbol{\beta}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \rho_\theta(y_i - \max(0, \boldsymbol{x}_i \boldsymbol{\beta})) .$$

Lindgren(1997) suggested estimating the conditional quantiles nonparametrically with a local Kaplan-Meier estimators of the survival functions for more general censored case.

In this paper we propose the regression quantile estimators for the survival data with multiple observations at each covariate vector, $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k$, under the presence of censorings. Based on the sample quantiles from multiple observations, the regression quantile estimator is obtained by the weighted least square method. We show that the proposed estimators are asymptotically normally distributed and they perform well even

in the heteroscedastic error model via the simulation study.

2. Regression Quantile Estimations

Let T_{ij} be the survival time of the j-th individual corresponding to the covariate vector, $\boldsymbol{x}_i$ or transformation on it, where $j = 1,2,\cdots,n_i$, $i = 1,2,\cdots,k$. Let $\boldsymbol{x}_i$ be the the associated covariate vector with p+1 components, where the first component is set to 1. Let $q_i(\theta)$ be the theta θ-th quantile of T_{ij} given $\boldsymbol{x}_i$ for θ in (0,1) then

$$q_i(\theta) = inf\{t : P(T_{ij} \le t \mid \boldsymbol{x}_i) \ge \theta\}.$$

Assume that the $q_i(\theta)$ is linearly related to the covariate vector $\boldsymbol{x}_i$ as

$$q_i(\theta) = \boldsymbol{x}_i \boldsymbol{\beta}(\theta) \text{ for } i = 1,2,\cdots,k, \tag{2.1}$$

where $\boldsymbol{\beta}(\theta)$ is a (p+1)-dimensional regression quantile modelling the covariate effect on the θ-th quantiles of T_{ij}.

The observed survival time is $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \le C_{ij})$, where C_{ij} is the censored time corresponding to $\boldsymbol{x}_i$ for $j = 1,2,\cdots,n_i$. C_{ij}'s are assumed to be independently distributed with unknown distribution functions G_i's. Since T_{ij} and C_{ij} are assumed to be independent given $\boldsymbol{x}_i$,

$$\begin{aligned} P(Y_{ij} \le q_i(\theta) \mid \boldsymbol{x}_i) &= 1 - P(T_{ij} > q_i(\theta) \mid \boldsymbol{x}_i) P(C_{ij} > q_i(\theta) \mid \boldsymbol{x}_i) \\ &\ge 1 - (1-\theta)(1 - G_i(q_i(\theta))). \end{aligned} \tag{2.2}$$

The right-hand side of (2.2) depends on θ and $G_i(q_i(\theta))$, not on the distribution of the survival times. Denote it by p_i then $q_i(\theta)$ can be set to satisfy

$$q_i(\theta) = inf\{y : P(Y_{ij} \le y \mid \boldsymbol{x}_i) \ge p_i\},$$

that is, $q_i(\theta)$ can be the p_i-th quantile of Y_{ij} given $\boldsymbol{x}_i$. For a fixed value of p_i $q_i(\theta)$ can be estimated by the p_i-th sample quantile of Y_{ij}, which is the value of q_i minimizing the objective function,

$$\sum_{j=1}^{n_i} \rho_{p_i}(y_{ij} - q_i) \text{ for } i = 1, 2, \cdots, k .$$

Given x_i, the distribution function G_i of C_{ij} can be estimated by the empirical distribution function,

$$\hat{G}_i(y) = \frac{1}{m_i} \sum_{j=1}^{n_i} I(Y_{ij} \le y \mid \delta_{ij} = 0) \text{ for } i = 1, 2, \cdots, k \tag{2.3}$$

where $m_i = n_i - \sum_{j=1}^{n_i} \delta_{ij}$.

Then for each $i = 1, 2, \cdots, k$, $q_i(\theta)$ can be estimated by iteration method as follows:

i) Set $q_i(\theta)$ to the initial value $q_i(\theta)^{(0)}$.

ii) $p_i^{(l)} = 1 - (1 - \theta)(1 - \hat{G}_i(q_i(\theta)^{(l)}))$

iii) $q_i(\theta)^{(l+1)}$ is the value of q_i minimizing the objective function,

$$\sum_{j=1}^{n_i} \rho_p(y_{ij} - q_i) \text{ with } p = p_i^{(l)}$$

Under the assumption that given x_i the distribution function F_i of Y_{ij} is not flat in the neighborhood of $F_i^{-1}(\tilde{p}_i)$, $\hat{q}_i(\theta)$ is the consistent estimator of $F_i^{-1}(\tilde{p}_i)$, where $F_i^{-1}(\tilde{p}_i)$ is the $\tilde{p}_i$ -th quantile of Y_{ij} given x_i and $\tilde{p}_i = 1 - (1 - \theta)(1 - \hat{G}_i(q_i(\theta)))$. Under the assumption that for each i, given x_i, Y_{ij}'s are independently distributed with distribution function, F_i, whose density, f_i, is continuous and positive,

$$\hat{q}_i(\theta) \sim AN\left(F_i^{-1}(\tilde{p}_i), \frac{\tilde{p}_i(1 - \tilde{p}_i)}{n_i f_i^2(F_i^{-1}(\tilde{p}_i))}\right).$$

For each fixed y, $-\infty < y < \infty$, the empirical distribution function $\hat{G}_i(y)$ is the consistent estimator of $G_i(y)$, which implies that $\tilde{p}_i$ is the consistent estimator of $p_i = 1 - (1 - \theta)(1 - G_i(q_i(\theta)))$. The limiting distribution of $\hat{q}_i(\theta)$ can be obtained as

$$\begin{aligned}\hat{q}_i(\theta) &\sim AN\left(F_i^{-1}(p_i),\ \frac{p_i(1-p_i)}{n_i f_i^2(F_i^{-1}(p_i))}\right)\\ &= AN\left(q_i(\theta),\ \frac{p_i(1-p_i)}{n_i f_i^2(q_i(\theta))}\right)\end{aligned} \tag{2.4}$$

Using $\hat{q}_i(\theta)$ the θ-th regression quantile $\boldsymbol{\beta}(\theta)$ of T_{ij} can be estimated by the value $\boldsymbol{b}$ minimizing the objective function,

$$\sum_{i=1}^{k} n_i(\hat{q}_i(\theta) - \boldsymbol{x}_i \boldsymbol{b})^2,$$

which is,

$$\hat{\boldsymbol{\beta}}(\theta) = \left(\sum_{i=1}^{k} n_i \boldsymbol{x}_i' \boldsymbol{x}_i\right)^{-1} \sum_{i=1}^{k} n_i \boldsymbol{x}_i' \hat{q}_i(\theta). \tag{2.5}$$

Under the following assumptions:

a1) For each i, given $\boldsymbol{x}_i$, Y_{ij}'s are independently distributed with distribution function F_i which is not flat in the neighborhood of $q_i(\theta)$ and density f_i is continuous and positive.

a2) $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k$ span R^{p+1}.

a3) $\frac{n_i}{N} \rightarrow \lambda_i$ as $N \rightarrow \infty$ where $N = \sum_{i=1}^{k} n_i$.

Using (2.4) and (2.5) the limiting distribution of the estimator of the θ-th regression quantiles can be obtained as

$$\hat{\boldsymbol{\beta}}(\theta) \sim AN\left(\boldsymbol{\beta}(\theta),\ \frac{1}{N} D^{-1} V D^{-1}\right) \tag{2.6}$$

where

$$D = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i' \boldsymbol{x}_i \quad \text{and} \quad V = \sum_{i=1}^{k} \frac{\lambda_i p_i(1-p_i)}{f_i^2(q_i(\theta))} \boldsymbol{x}_i' \boldsymbol{x}_i .$$

3. Numerical Studies

To investigate the performance of the proposed estimators the simulation study is conducted. For the comparison, the median estimators of Ying *et al.*(1995) are obtained. Let T_{ij} be the j-th survival time corresponding to the covariate vector $\boldsymbol{x}_i$ for $i = 1,2,3$. Assume that the survival times T_{ij}'s are from the linear regression model:

$$T_{ij} = \boldsymbol{x}_i \boldsymbol{\beta} + e_{ij} \quad \text{for } j = 1, 2, \cdots, n_i, \quad i = 1, 2, 3, \tag{3.1}$$

where error terms e_{ij}'s follow a logistic distribution with a location parameter of 0 and a scale parameter of $\boldsymbol{x}_i \boldsymbol{\phi}$, which leads

$$P(T_{ij} \leq t \mid \boldsymbol{x}_i) = P\left(T_0 \leq \frac{t - \boldsymbol{x}_i \boldsymbol{\beta}}{\boldsymbol{x}_i \boldsymbol{\phi}}\right),$$

where T_0 follows a standard logistic distribution, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1)'$ The censored times C_{ij}'s are assumed to follow a eponential distribution with a location parameter c_{oi} and a scale parameter $\boldsymbol{x}_i \boldsymbol{\phi}$ as

$$P(C_{ij} \leq t \mid \boldsymbol{x}_i) = P\left(C_0 \leq \frac{t - c_{oi}}{\boldsymbol{x}_i \boldsymbol{\phi}}\right) I(t \geq c_{oi}),$$

where C_0 follows a standard exponential distribution and c_{oi}'s are chosen to be 10% of proportion of censoring.

Set $\boldsymbol{x}_i = (1, i-1)$ and $n_1 = 150$, $n_2 = 200$, $n_3 = 250$ then the regression quantile of the survival times in the model (2.1) is,

$$\boldsymbol{\beta}(\theta) = \boldsymbol{\beta} + F_0^{-1}(\theta)\, \boldsymbol{\phi}, \quad \theta \in (0, 1), \tag{3.2}$$

where $F_0^{-1}(\theta)$ is the θ-th quantile of T_0, which is the θ-th quantile of the standard logistic distribution. For each $\boldsymbol{x}_i$, $i = 1, 2, 3$, with $\boldsymbol{\beta} = (1, 1)'$ and $\boldsymbol{\phi} = (0.5, 0.5)'$, 500 random samples of $\{Y_{ij}, \delta_{ij}, j = 1, 2, \cdots, n_i\}$ are generated. Then for $\theta = 0.25, 0.5, 0.75$, true values of regression quantiles of the survival times are obtained by (3.2) as, respectively,

$$\boldsymbol{\beta}(0.25) = \begin{pmatrix} 0.4507 \\ 0.4507 \end{pmatrix}, \quad \boldsymbol{\beta}(0.5) = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}, \quad \boldsymbol{\beta}(0.75) = \begin{pmatrix} 1.5493 \\ 1.5493 \end{pmatrix}.$$

From each sample of size 600, estimators of 2 components of the θ-th regression quantile, $\hat{\beta}(\theta)$and $\hat{\beta}(\theta)$, for $\theta = 0.25, 0.5, 0.75$, are obtained by (2.5) based on the sample quantiles of T_{ij} estimated by the iteration method. The estimators proposed by Ying *et al.*(1995), $\hat{\beta}(0.5)$ and $\hat{\beta}(0.5)$, are obtained through minimization of the objective function which may have multiple minima, the simulated annealing algorithm

applied to the accelerated failure model by Lin and Geyer(1992) is used for the estimation. Based on 500 samples the empirical means and mean square errors of estimators are obtained.

From Table 1 we can see that the proposed estimators $\hat{\beta}(\theta)$ provides relatively accurate results even for the heteroscedastic error model. For θ = 0.5 the proposed estimators, $\hat{\beta}_0(0.5)$ and $\hat{\beta}_1(0.5)$, have the values of empirical means closer to the true values than the estimators of Ying *et al.*(1995), $\hat{\beta}_{Y0}(0.5)$ and $\hat{\beta}_{Y1}(0.5)$. And the proposed estimators have smaller values of mean square errors than the estimators of Ying *et al.*(1995).

Table 1. Mean and MSE of estimators of $\beta(\theta)$ with 10% Censoring

θ		0.25	0.5	0.75
$\hat{\beta}_0(\theta)$	mean	0.4481	0.9931	1.7074
	mse	0.1066	0.0934	0.5175
$\hat{\beta}_{Y0}(0.5)$	mean		1.0187	
	mse		0.1304	
$\hat{\beta}_1(\theta)$	mean	0.4498	0.9975	1.6268
	mse	0.1283	0.1091	0.5802
$\hat{\beta}_{Y1}(0.5)$	mean		0.9777	
	mse		0.1291	

4. Remarks and Conclusions

Based on the empirical distribution function of censored times and the sample quantiles of the observed survival times, the estimators of regression quantiles are studied in a censored quantile regression model with multiple survival times at each covariate vector. The quantile regression approach detects the reversal of treatment effects which a proportional hazard model does not permit.

The simulation study shows that the proposed estimators perform well, even in the heteroscedastic error model, where the variance of errors depends on the covariate vector. And a rather easy numerical method is required than Ying *et al.*(1995).

References

1. Koenker, R. and Bassett, G. (1978). Regression Quantiles, *Econometrica* 46, 33-50.
2. Lind, D. Y. and Geyer, C. J. (1992). Computational Methods for Semiparametric Linear Regression With Censored Data, *Journal of Computational and Graphical Statistics*, 1, 77-90.
3. Lindgren, A. (1997). Quantile Regression with Censored Data using Generalized L1 Minimization, *Computational Statistics & Data Analysis*, 23, 509-524.
4. Powell, J. (1986). Censored Regression Quantiles. *Journal of Econometrics*, 32, 143-155.
5. Yang, S. (1999). Censored Median Regression Using Weighted Empirical Survival and Hazard Functions, *Journal of the American Statistical Association,* 94, 137-145.
6. Ying, Z., Jung, S. and Wei, L. (1995). Survival Analysis with Censored Median Regression Models, *Journal of the American Statistical Association,* 90, 178-184.

[2002년 1월 접수, 2002년 5월 채택]