应用概率统计　第二十六卷
第二期　2010年4月

Chinese Journal of Applied Probability
and Statistics Vol.26 No.2 Apr. 2010

Median Regression for Left and Right Censored Data *

Zheng Ming　　Yu Wen*
(Department of Statistics, Fudan University, Shanghai, 200433)

Abstract

In many applications involving follow-up studies, individuals' lifetimes may be subject to both left and right censoring. In this paper we consider regressing the median of the lifetime variable or a transformation thereof on the corresponding covariates when the lifetimes are subject to both left and right censoring and the values of both censoring variables are always observable. A semi-parametric inferential procedure is proposed and the asymptotic properties of the proposed method are discussed. We also propose an alternative empirical likelihood based inferential procedure for the regression coefficients vector. Moreover, we make some discussion on classical doubly censored data whose censoring variables cannot always be observed. The proposed methods are illustrated by some simulation studies.

Keywords: Left and right censoring, empirical likelihood, least absolute deviations, median regression.

AMS Subject Classification: Primary 62N01, secondary 62G10.

§1. Introduction

In survival analysis, individuals' lifetimes are often subject to different types of censoring. Right censoring is one of the most common situations. Similar to right censoring, left censoring occurs if the individual is observed to fail prior to some time while the actual failure time is unknown. When lifetimes are said to be subject to double censoring, it means both left and right censoring exist. That is to say, lifetime T is observed when it lies in a random interval $[X, Y]$ ($X \le Y$ with probability 1) and is censored when it is smaller than X or larger than Y. Classical double censoring is defined and widely discussed in literature (Gehan [1]; Turnbull [2]; DeGruttola and Lagakos [3]; Zhang and Li [4]; Cai and Cheng [5]), and it is different from interval censoring. For interval censoring, one can only observe that the lifetime T lies in some random intervals without knowing its exact value all the time.

In some situations of double censoring, both censoring variables, X and Y, might be observed regardless of T bing censored or not. For example, a Stanford psychiatrist wanted

*The project supported by NSFC Grant (10971033) and Shanghai Leading Academic Discipline Project (B210).
*Corresponding author, E-mail: wenyu@fudan.edu.cn.
Received January 23, 2007. Revised March 24, 2009.

to establish norms for infant development for a community in Kenya by investigating how long would infants take to learn a particular task in order to make comparisons with the known standards in the United States and United Kingdom. In his sample there were children born between July 1 and December 31, 1969. When he started the study in January 1970, some children had already learned to perform the task; whereas losses occurred when some infants were still unsuccessful by the end of the study (Leiderman et al. [6]). In this case, T would represent the time from birth to first learn to perform the particular task and was subject to double censoring. The censoring variables, X and Y, representing the age of these infants at the beginning and the end of the study, could be observed for each individual in the sample.

When the lifetimes appear with some corresponding covariates, regression models are proposed. For survival data, the accelerated failure time (AFT) model is widely used and discussed (Miller [7]; Buckley and James [8]; Leurgans [9]). The AFT model regresses the mean of the logarithm of the lifetime on its covariates. Its ease of interpretation is rather attracting to practitioners. However, the presence of censoring may preclude the estimation of intercept parameter in the model (Meier [10]), and the AFT model usually assumes the error terms follow identical distribution, which is not reasonable in some applications.

The median is another important parameter for measuring the center of a long-tailed or skew survival population. Unlike the mean, it can be well estimated when the censoring is not too heavy while the mean is often not estimatable under such circumstances. Therefore the regression of the median of the lifetime or a monotone transformation thereof on the covariates is a natural alternative to the AFT model. For uncensored data, robust estimation of median regression models obtained by least absolute deviations (LAD) method and its asymptotic properties are studied by Bassett and Koenker [11], Koenker and Bassett [12] and Chen et al. [13], among others. For right censored data, Powell [14, 15] analyzed median regression models with the right censoring variables being always observable. For general random right censoring situation, recent works include Ying et al. [16], Jung [17], Yang [18], Band and Tsiatis [19], Portnoy [20], Liu and Ren [21], etc.

Though the median regression for right censored data has been widely discussed, the existing methods can not be applied to doubly censored data directly. In this paper we propose a semiparametric inferential procedure for median regression models with left and right censored data when both censoring variables are always observable. The main tool we use is the estimation equation for such observations, which is a nature extension of the one used by Ying et al. [16] for right censored data. Moreover, according to the estimating equation, we propose an alternative empirical likelihood based inferential procedure for the regression coefficients vector. The empirical likelihood method was introduced by Owen [22]. The method has been extended to many areas and its advantages have been well-recognized. A nice summary can be found in Owen [23]. Qin and Tsao [24], Liu and Ren

[21] used the empirical likelihood method to make inference about right censored median regression models. Our procedure is a parallel extension to left and right censored data.

The paper is structured as follows. In the Section 2, we discuss the median model mainly by proposing an estimating equation and using extended LAD method to solve it. In the Section 3, we define the empirical likelihood ratio for the regression coefficients vector and show the corresponding limiting distribution. In the Section 4, some simulation studies are carried out to evaluate the finite sample performances of our methods. We also make some discussion on the usage of the proposed estimating equations for classical doubly censored data in Section 5. All technical derivations are summarized in the Appendix.

§2. Median Regression Analysis

Let T be the continuous lifetime or transformation of it. Let $\mathbf{U}$ be a $p\times 1$ vector of covariates for T and denote $\mathbf{Z}=(1,\mathbf{U}')'$. Conditional on $\mathbf{Z}$, let the median of the conditional distribution of T be denoted by m. Suppose there exists an unknown $(p+1)\times 1$ vector β such that

$$m=\beta'\mathbf{Z}\,. \tag{2.1}$$

Let β_0 be the true value of β. Let X denote the random left censoring time with continuous distribution function $(1-H)$, and Y be the random right censoring time with distribution function $(1-G)$. Neither H nor G depends on $\mathbf{Z}$. It is also assumed that T is independent of (X,Y), while X and Y may be dependent and $\mathsf{P}(X\leq Y)=1$. We observe $N=Y\wedge(T\vee X)$, where $\wedge$ means minimum and $\vee$ means maximum, the censoring variables (X,Y) and the covariate vector $\mathbf{Z}$. The observations are $(N_i,X_i,Y_i,\mathbf{Z}_i)$ $i=1,2,\cdots,n$, which are i.i.d. copies of $(N,X,Y,\mathbf{Z})$. Let the conditional distribution of $T_i-\beta_0'\mathbf{Z}_i$ be denoted by F_i, which is continuous, completely unspecified and may depend on the covariate vector $\mathbf{Z}_i$.

For the uncensored case, the LAD estimator for β_0 is obtained by minimizing

$$\sum_{i=1}^n|T_i-\beta'\mathbf{Z}_i|. \tag{2.2}$$

Note that the minimizer for (2.2) is a root of the following estimating equation:

$$\mathbf{U}_n(\beta)=\sum_{i=1}^n\Big[I(T_i\geq\beta'\mathbf{Z}_i)-\frac{1}{2}\Big]\mathbf{Z}_i\approx 0, \tag{2.3}$$

where $I(\cdot)$ is the indicator function. The approximation sign is used here because $\mathbf{U}_n(\beta)$ is a discontinuous function of β. Conditional on $\mathbf{Z}_i$, the expected value of $\mathbf{U}_n(\beta_0)$ is zero under model (2.1), therefore $\mathbf{U}_n(\beta)$ is a reasonable estimating function for β.

Since given $\mathbf{Z}_i$, the expected value of $I(N_i\geq\beta_0'\mathbf{Z}_i)$ is $(1/2)[G(\beta_0'\mathbf{Z}_i)+H(\beta_0'\mathbf{Z}_i)]$, which is shown in Appendix A.1, a natural estimating equation, which resembles (2.3), for β_0 is

$$\mathbf{S}_n^0(\beta)=\sum_{i=1}^n\Big[\frac{I(N_i\geq\beta'\mathbf{Z}_i)}{G(\beta'\mathbf{Z}_i)+H(\beta'\mathbf{Z}_i)}-\frac{1}{2}\Big]\mathbf{Z}_i\approx 0.$$

When G and H are not known, we propose the following estimating equation

$$\mathbf{S}_n(\beta)=\sum_{i=1}^{n}\Big[\frac{I(N_i\ge\beta'\mathbf{Z}_i)}{\widehat{G}(\beta'\mathbf{Z}_i)+\widehat{H}(\beta'\mathbf{Z}_i)}-\frac{1}{2}\Big]\mathbf{Z}_i\approx 0, \tag{2.4}$$

where $\widehat{H}$ and $\widehat{G}$ are the empirical distributions based on X_i, $i=1,2,\cdots,n$ and Y_i, $i=1,2,\cdots,n$, respectively.

A "root" $\widehat{\beta}$ of (2.4) may be defined as the minimizer of the function $\|\mathbf{S}_n(\beta)\|$, where $\|a\|^2=a'a$ for any column vector a. Since (2.4) is a discrete function for β, grid search method is used to locate its root. Next we discuss the large sample properties of $\widehat{\beta}$. Assume that the following regularity conditions hold:

1. g, h and f_i, which are the derivatives of $-G$, $-H$ and F_i, are uniformly bounded;

2. The covariates vector $\mathbf{Z}$ is bounded;

3. the true value β_0 of β is in the interior of a bounded convex region D;

4. For $\beta\in D$, suppose that there exists constants $\widetilde{t}$ such that $\mathsf{P}(Y\ge\widetilde{t}|\mathbf{Z})+\mathsf{P}(X\ge\widetilde{t}|\mathbf{Z})>0$ and $\beta'\mathbf{Z}\le\widetilde{t}$ with probability 1;

5. The matrix $\mathsf{E}[\mathbf{Z}\mathbf{Z}'f(0|\mathbf{Z})]$ is positive definite, where $f(\cdot|\mathbf{z})$ denotes the conditional density of $T-\beta_0'\mathbf{Z}$ given $\mathbf{Z}=\mathbf{z}$.

Theorem 2.1 Under the conditions 1-5, $\widehat{\beta}\to\beta_0$ a.s., as $n\to\infty$.

The theorem is proved in Appendix A.2. In addition to the point estimate, one may also consider the problem of testing the null hypothesis $H_0:\beta_0=\beta$, or constructing confidence regions for β_0. Usually, a Wald-type statistic based on $\widehat{\beta}$ and its large sample distribution can be used. In Appendix A.4 we demonstrate that $\sqrt{n}(\widehat{\beta}-\beta_0)$ is asymptotically normally distributed with mean zero and certain variance-covariance matrix. However, the variance-covariance matrix of the limiting distribution depends on the unknown density functions f_i. Because of the existence of left and right censored data, they may not be well estimated.

One alternative way is to use the estimating function $\mathbf{S}_n(\beta)$ directly. We show in Appendix A.3 that the distribution of $n^{-1/2}\mathbf{S}_n(\beta_0)$ is approximately normal with mean zero and some variance-covariance matrix Γ. Similarly, we need to estimate Γ. However, since the explicit expression of Γ is not easy to get, we turn to the well-known jackknife variance estimator, which could be calculated quite simply. Denote $\widehat{\Gamma}_{jack}$ as the jackknife estimator for Γ, then

$$\widehat{\Gamma}_{\text{jack}}=\frac{n-1}{n^2}\sum_{i=1}^{n}\Big[\mathbf{S}_{n-1,i}(\widehat{\beta})-\frac{1}{n}\sum_{j=1}^{n}\mathbf{S}_{n-1,j}(\widehat{\beta})\Big]^{\otimes 2},$$

where

$$\mathbf{S}_{n-1,i}(\widehat{\beta}) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \Big[\frac{I(N_j \geq \widehat{\beta}'\mathbf{Z}_j)}{\widehat{G}(\widehat{\beta}'\mathbf{Z}_j) + \widehat{H}(\widehat{\beta}'\mathbf{Z}_j)} - \frac{1}{2}\Big]\mathbf{Z}_j, \qquad i = 1, 2, \cdots, n,$$

and $a^{\otimes 2} = aa'$ for any column vector a. Let $\widehat{\Gamma}$ be any consistent estimator of Γ. From the general theory of Jackknife, one may show that $\widehat{\Gamma}_{\text{jack}} = \widehat{\Gamma} + o_p(1)$ under some regular conditions (See, for example, Shao and Tu [25]). Thus, a natural test statistic based on $\mathbf{S}_n$ for testing H_0 would be $n^{-1}\mathbf{S}_n'(\beta_0)\widehat{\Gamma}_{\text{jack}}^{-1}\mathbf{S}_n(\beta_0)$, which has the same approximately chi-square distribution with $(p+1)$ degrees of freedom as $n^{-1}\mathbf{S}_n'(\beta_0)\widehat{\Gamma}^{-1}\mathbf{S}_n(\beta_0)$.

One may also hope to check the model assumption (2.1). We propose one to do this as Ying et al. [16] did in Section 4 of their paper, by just replacing $\widehat{G}(\widehat{\beta}'\mathbf{Z}_i)^{-1}I(Y_i \geq \widehat{\beta}'\mathbf{Z}_i) - 1/2$ there with $[\widehat{G}(\widehat{\beta}'\mathbf{Z}_i) + \widehat{H}(\widehat{\beta}'\mathbf{Z}_i)]^{-1}I(N_i \geq \widehat{\beta}'\mathbf{Z}_i) - 1/2$.

§3. Empirical Likelihood Inference for β_0

Note that

$$\mathsf{E}\Big[\Big(\frac{I(N_i \geq \beta_0'\mathbf{Z}_i)}{G(\beta_0'\mathbf{Z}_i) + H(\beta_0'\mathbf{Z}_i)} - \frac{1}{2}\Big)\mathbf{Z}_i\Big] = 0, \qquad i = 1, 2, \cdots, n,$$

it is possible to construct an empirical likelihood based confidence region for β_0. As discussed in Section 1, the empirical likelihood confidence region may complement the normal approximation method effectively. Let $p = (p_1, p_2, \cdots, p_n)$ be a probability vector, i.e., $\sum p_i = 1$ and $p_i \geq 0$ for all i. For $i = 1, 2, \cdots, n$, define

$$S_{ni}(\beta) = \Big[\frac{I(N_i \geq \beta'\mathbf{Z}_i)}{\widehat{G}(\beta'\mathbf{Z}_i) + \widehat{H}(\beta'\mathbf{Z}_i)} - \frac{1}{2}\Big]\mathbf{Z}_i.$$

Then, an estimated empirical likelihood function is defined by

$$L_n(\beta) = \sup\Big\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i S_{ni}(\beta) = 0, \sum_{i=1}^{n} p_i = 1, p_i \geq 0\Big\}.$$

By the Lagrange multiplier and the similar arguments in Qin and Tsao [24], the corresponding empirical log-likelihood ratio, evaluated at β_0, is defined as

$$l_n(\beta_0) = -2\log L_n(\beta_0)/n^{-n} = 2\sum_{i=1}^{n} \log\{1 + \lambda' S_{ni}(\beta_0)\},$$

where $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_{p+1})'$ is the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \frac{S_{ni}(\beta_0)}{1 + \lambda' S_{ni}(\beta_0)} = 0.$$

To derive the limiting distribution, we need to assume that the conditions 1-5 hold. Furthermore, let

$$\Gamma_1 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{I(N_i \geq \beta_0'\mathbf{Z}_i)}{G(\beta_0'\mathbf{Z}_i) + H(\beta_0'\mathbf{Z}_i)} - \frac{1}{2}\Big]^2 \mathbf{Z}_i\mathbf{Z}_i'.$$

The following theorem gives the limiting distribution of $l_n(\beta_0)$.

Theorem 3.1 Under the conditions 1-5, we have

$$l_n(\beta_0) \to l_1\chi^2_{1,1} + l_2\chi^2_{2,1} + \cdots + l_{p+1}\chi^2_{p+1,1}$$

in distribution as $n \to \infty$, where the weights l_i's are the eigenvalues of $\Gamma_1^{-1}\Gamma$ and $\chi^2_{i,1}$ $(i = 1, 2, \cdots, p+1)$ are independent chi-square random variables with one degree of freedom.

The theorem is proved in Appendix A.5. In order to apply Theorem 3.1, the weights l_i's have to be estimated. Under the assumptions of Theorem 3.1, Γ_1 can be consistently estimated by

$$\widehat{\Gamma}_1 = \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{I(N_i \geq \widehat{\beta}'\mathbf{Z}_i)}{\widehat{G}(\widehat{\beta}'\mathbf{Z}_i) + \widehat{H}(\widehat{\beta}'\mathbf{Z}_i)} - \frac{1}{2}\Big]^2 \mathbf{Z}_i\mathbf{Z}_i'.$$

Meanwhile, Γ could be estimated by $\widehat{\Gamma}_{\text{jack}}$ as discussed in Section 2. It follows that the l_i's can be estimated by the $\widehat{l}_i$'s, which are the eigenvalues of $\widehat{\Gamma}_1^{-1}\widehat{\Gamma}_{\text{jack}}$.

Now Let $R_\alpha(\beta) = \{\beta : l_n(\beta) \leq c_\alpha\}$, where c_α is the $(1-\alpha)$th quantile of the weighted chi-square distribution $\widehat{l}_1\chi^2_{1,1} + \widehat{l}_2\chi^2_{2,1} + \cdots + \widehat{l}_{p+1}\chi^2_{p+1,1}$. By Theorem 3.1, $R_\alpha(\beta)$ is an approximate confidence region for β_0 with asymptotically correct coverage probability $1-\alpha$.

Note that the limiting distribution of our empirical likelihood ratio test statistic is different from that of the standard empirical likelihood method. This is because we replace the unknown nuisance parameters H and G by their consistent estimators in $S_{ni}(\beta_0)$. By appropriately modifying the construction of $S_{ni}(\beta_0)$, it is possible to get an empirical likelihood ratio test statistic with standard chi-squared limiting distribution.

§4. Simulation

In this section, simulation studies were conducted to investigate finite sample properties of the proposed inference procedures. We considered the following median regression models:

Model A: The covariates u_i's were generated from uniform distribution $U[0.5, 1.5]$. For a given u_i value, T_i was generated from a normal distribution with mean u_i and constant variance 1. The left censoring variable X_i's were generated from $U[-1, 1]$ distribution and the right censoring variable Y_i's were from the distribution of $U[1, 3]$ distribution.

Model B: Same as Model A except that given u_i, T_i was generated from a normal distribution with mean u_i and variance u_i, i.e., the error items in model (2.1) depended on u_i's.

Model C: The covariates u_i's took two values of 1, 2 with probability 0.5, respectively. For a given u_i value, T_i was generated from an exponential distribution with median u_i and mean $u_i/\log(2)$. The left censoring variable X_i was generated from $U[0,1]$ distribution and the right censoring variable Y_i was from the distribution of $U[1,8]$ distribution.

According the generating schemes, in all the models the true values of the intercept and slope parameters are 0 and 1, respectively, and both left and right censoring proportion are about 20%. We got $\widehat{\beta}$ by minimizing $\|\mathbf{S}_n(\beta)\|$. For each model and each sample size, we generated $M = 100$ sets of data and calculated the averages and standard errors of the estimates. The results are listed in Table 1.

Table 1

	n	intercept	std	slope	std
	10	0.089	0.388	0.883	0.370
	20	-0.051	0.356	0.951	0.351
Model A	60	-0.036	0.357	1.037	0.329
	100	-0.043	0.329	1.081	0.319
	200	-0.037	0.300	1.034	0.282
	10	0.087	0.420	0.845	0.363
	20	-0.007	0.343	0.931	0.352
Model B	60	-0.103	0.329	1.068	0.349
	100	-0.052	0.270	1.072	0.311
	200	-0.050	0.252	1.063	0.287
	10	-0.287	0.374	0.803	0.382
	20	-0.173	0.361	0.988	0.368
Model C	60	-0.064	0.376	1.030	0.339
	100	-0.008	0.363	1.045	0.316
	200	-0.037	0.328	1.023	0.262

For all the models (especially for the Model C), the proposed estimator seems to be a little bit biased when the sample size is small (say, < 60). With the increase of the sample size, the averages go closer to the true value of the regression coefficients, and the standard errors decrease.

Then we compared the performance of confidence regions based on the normal approximation based method (Norm) and the empirical likelihood based method (EL). For each model and each sample size, we generated $M = 1000$ sets of data and simulated the empirical coverage probabilities of the two methods. The nominal levels were chose to be 0.90 and 0.95. The results are summarized in Table 2.

Table 2

nominal level	n	0.90		0.95	
		Norm	EL	Norm	EL
Model A	60	0.838	0.876	0.896	0.936
	100	0.859	0.884	0.913	0.946
	200	0.878	0.910	0.919	0.956
Model B	60	0.845	0.898	0.889	0.946
	100	0.866	0.891	0.925	0.940
	200	0.880	0.899	0.948	0.945
Model C	60	0.816	0.885	0.868	0.929
	100	0.849	0.896	0.897	0.940
	200	0.879	0.895	0.938	0.945

Results in Table 2 led to the following observations: For all the models, the empirical likelihood based method performs better than the normal approximation based method when the sample size is not large. With the increase of the sample size, both methods have quite good empirical coverage probabilities. Moreover, the empirical likelihood based method seems to be less affected by the model assumptions. According to these observations, one is recommended to use the empirical likelihood based confidence region in general. When the sample size is large, the normal approximation based one can also be used.

§5. Further Discussion

Here we give some discussion on classical doubly censored data. Assume X_i, Y_i, T_i, N_i and $\mathbf{Z}_i$ have the same meaning as those in Section 2. For classical doubly censored data, we observe N_i, $\mathbf{Z}_i$ and a double censoring indicator $\delta_i = 1^{I(T_i<X_i)}2^{I(X_i\le T_i\le Y_i)}3^{I(T_i>Y_i)}$, but the censoring variables are not available for those observed T_i's.

Under this data structure, we still intend to use (2.4) to estimate β. However, since the observations of X_i and Y_i are not complete here, empirical function $\widehat{G}$ and $\widehat{H}$ are no longer proper estimators for G and H. Note that $N_i = Y_i \wedge (T_i \vee X_i) = X_i \vee (T_i \wedge Y_i)$, which means that Y is right censored by $T \vee X$ and X is left censored by $T \wedge Y$ as well. Thus, in the light of Kaplan-Meier PL estimator (Kaplan and Meier [26]), we estimate G and H by

$$\widetilde{G}(t) = \prod_{i:N_{(i)}\le t}\Big(1-\frac{1}{n-i+1}\Big)^{\eta_{(i)}},$$

where $\eta_i = 1$ for $\delta_i = 3$ and 0 for others, $N_{(i)}$'s are the ordered observations of N_i's and $\eta_{(i)}$ is the value of η associated with $N_{(i)}$, and

$$\widetilde{H}(t) = 1 - \prod_{i:N_{(i)}>t} \Big(1 - \frac{1}{i}\Big)^{\zeta_{(i)}},$$

where $\zeta_i = 1$ for $\delta_i = 1$ and 0 for others and $\zeta_{(i)}$ is the value of ζ associated with $N_{(i)}$. We rewrite (2.4) into

$$\widetilde{\mathbf{S}}_n(\beta) = \sum_{i=1}^{n} \Big[\frac{I(N_i \geq \beta'\mathbf{Z}_i)}{\widetilde{G}(\beta'\mathbf{Z}_i) + \widetilde{H}(\beta'\mathbf{Z}_i)} - \frac{1}{2}\Big]\mathbf{Z}_i.$$

By minimizing $\|\widetilde{\mathbf{S}}_n(\beta)\|$ here we also get the estimator $\widetilde{\beta}$. Naturally we are concerned with the asymptotic properties of $\widetilde{\beta}$ and $\widetilde{\mathbf{S}}_n(\beta_0)$ as considered above, but the situation becomes complicated here. Although $\widetilde{G}$ has the same form as the Kaplan-Meier PL estimator and $\widetilde{H}$ is a corresponding estimator for left censoring data, Y is not independent of the "right censoring variable" $T \vee X$ and the same problem exists between X and $T \wedge Y$, since Y and X may be dependent. Thus $\widetilde{G}$ here is not the common Kaplan-Meier PL estimator for lifetimes with independent censoring variables. Thus, we are not sure about the corresponding large sample properties of such estimator so far, and we are not clear if $\widetilde{\beta}$ and $\widetilde{\mathbf{S}}_n(\beta_0)$ here have the same large sample behaviors as $\widehat{\beta}$ and $\mathbf{S}_n(\beta_0)$ discussed in Section 2 and 3. Though theoretical justifications are not available, we still did some simulation with Model A, B and C in Section 4 by using such estimators to explore their large sample behaviors. The results are listed in Table 3 and Table 4.

Table 3

	n	intercept	std	slope	std
Model A	10	-0.258	0.363	0.931	0.409
	20	0.017	0.372	0.905	0.354
	60	0.003	0.355	1.010	0.330
	100	0.016	0.318	0.987	0.309
	200	0.011	0.317	0.990	0.299
Model B	10	-0.217	0.368	0.904	0.392
	20	0.008	0.351	0.931	0.373
	60	0.016	0.350	0.985	0.341
	100	0.017	0.332	0.991	0.323
	200	-0.001	0.249	1.019	0.298
Model C	10	-0.362	0.319	0.820	0.425
	20	-0.325	0.314	0.934	0.415
	60	-0.125	0.350	0.992	0.344
	100	-0.059	0.322	1.049	0.300
	200	0.026	0.342	0.987	0.284

Table 4

nominal level	n	0.90 Norm	0.90 EL	0.95 Norm	0.95 EL
Model A	60	0.843	0.851	0.904	0.930
	100	0.869	0.856	0.928	0.936
	200	0.862	0.869	0.931	0.926
Model B	60	0.839	0.845	0.913	0.930
	100	0.869	0.864	0.921	0.930
	200	0.874	0.869	0.931	0.934
Model C	60	0.844	0.870	0.914	0.927
	100	0.875	0.872	0.936	0.937
	200	0.900	0.883	0.946	0.953

The simulation results show some useful hints. When the sample size is small (say, $<$ 60), the estimator does not perform well. The averages of the estimates indicate that $\widetilde{\beta}$ may be biased, especially for the Model C, in which T has a long tail distribution. However, with the increase of the sample size, the bias disappears and the empirical coverage probabilities go closer to the nominal ones. It seems that the large sample behaviors of $\widehat{\beta}$ and $\mathbf{S}_n(\beta_0)$ are still valid for $\widetilde{\beta}$ and $\widetilde{\mathbf{S}}_n(\beta_0)$ under our simulation designs. However, strict theoretical proof of the properties of such estimators is still an open question.

Appendix

A.1 The Expected Value of $I(N_i \geq \beta_0'\mathbf{Z}_i)$ Given $\mathbf{Z}_i$

For any β,

$$\begin{aligned} & \mathsf{P}(N_i \geq \beta'\mathbf{Z}_i|\mathbf{Z}_i) \\ = \; & \mathsf{P}(Y_i \geq \beta'\mathbf{Z}_i, X_i \vee T_i \geq \beta'\mathbf{Z}_i|\mathbf{Z}_i) \\ = \; & \mathsf{P}(Y_i \geq \beta'\mathbf{Z}_i, X_i \geq \beta'\mathbf{Z}_i, T_i \geq \beta'\mathbf{Z}_i|\mathbf{Z}_i) \\ & + \mathsf{P}(Y_i \geq \beta'\mathbf{Z}_i, X_i \geq \beta'\mathbf{Z}_i > T_i|\mathbf{Z}_i) + \mathsf{P}(Y_i \geq \beta'\mathbf{Z}_i, T_i \geq \beta'\mathbf{Z}_i > X_i|\mathbf{Z}_i). \end{aligned}$$

By the fact that $Y_i \geq X_i$ with probability 1 and the independence between T_i and (X_i, Y_i),

$$\mathsf{P}(N_i \geq \beta'\mathbf{Z}_i|\mathbf{Z}_i) = \mathsf{P}(X_i \geq \beta'\mathbf{Z}_i|\mathbf{Z}_i)\mathsf{P}(T_i < \beta'\mathbf{Z}_i|\mathbf{Z}_i) + \mathsf{P}(Y_i \geq \beta'\mathbf{Z}_i|\mathbf{Z}_i)\mathsf{P}(T_i \geq \beta'\mathbf{Z}_i|\mathbf{Z}_i).$$

Since T_i is continuous and its conditional median $m_i = \beta_0'\mathbf{Z}_i$, given $\mathbf{Z}_i$, we get that

$$\mathsf{P}(N_i \geq \beta_0'\mathbf{Z}_i|\mathbf{Z}_i) = \frac{1}{2}[G(\beta_0'\mathbf{Z}_i) + H(\beta_0'\mathbf{Z}_i)].$$

A.2 Proof of Theorem 2.1

Let

$$\overline{\mathbf{S}}_n(\beta)=\sum_{i=1}^{n}\Big\{\frac{1}{2}-\frac{F_i((\beta-\beta_0)'\mathbf{Z}_i)G(\beta'\mathbf{Z}_i)+[1-F_i((\beta-\beta_0)'\mathbf{Z}_i)]H(\beta'\mathbf{Z}_i)}{G(\beta'\mathbf{Z}_i)+H(\beta'\mathbf{Z}_i)}\Big\}\mathbf{Z}_i.$$

Similar to the proof of Appendix A in Ying et al. [16], we get that

$$\sup_{\beta\in D}\|n^{-1}\mathbf{S}_n(\beta)-n^{-1}\overline{\mathbf{S}}_n(\beta)\|=o(n^{-1/2+\varepsilon})\qquad \text{a.s..}\tag{A.1}$$

Denote

$$\mathbf{A}_n(\beta)=\frac{1}{n}\frac{\partial\overline{\mathbf{S}}_n(\beta)}{\partial\beta}=-\frac{1}{n}\sum_{i=1}^{n}\frac{f_iG^2-f_iH^2+2F_igH-gH-2F_ihG+hG}{(G+H)^2}\mathbf{Z}_i\mathbf{Z}_i',$$

where $F_i=F_i((\beta-\beta_0)'\mathbf{Z}_i)$, $f_i=f_i((\beta-\beta_0)'\mathbf{Z}_i)$, $G=G(\beta'\mathbf{Z}_i)$, $H=H(\beta'\mathbf{Z}_i)$, $g=g(\beta'\mathbf{Z}_i)$, $h=h(\beta'\mathbf{Z}_i)$. Notice that $F_i(0)=1/2$, so with probability 1,

$$\begin{aligned}\mathbf{A}_n(\beta_0)&=-\frac{1}{n}\sum_{i=1}^{n}\frac{f_i(0)(G^2(\beta_0'\mathbf{Z}_i)-H^2(\beta_0'\mathbf{Z}_i))}{(G(\beta_0'\mathbf{Z}_i)+H(\beta_0'\mathbf{Z}_i))^2}\mathbf{Z}_i\mathbf{Z}_i'\\&\to-\mathsf{E}\Big[\frac{f(0|\mathbf{Z})(G^2(\beta_0'\mathbf{Z})-H^2(\beta_0'\mathbf{Z}))}{(G(\beta_0'\mathbf{Z})+H(\beta_0'\mathbf{Z}))^2}\mathbf{Z}\mathbf{Z}'\Big].\end{aligned}$$

Because $\mathsf{P}(X_i\le Y_i)=1$, we can get $G(\beta_0'\mathbf{Z})>H(\beta_0'\mathbf{Z})$. Since the matrix $\mathsf{E}[\mathbf{Z}\mathbf{Z}'f(0|\mathbf{Z})]$ is positive definite, the matrix

$$\mathsf{E}\Big[\frac{f(0|\mathbf{Z})(G^2(\beta_0'\mathbf{Z})-H^2(\beta_0'\mathbf{Z}))}{(G(\beta_0'\mathbf{Z})+H(\beta_0'\mathbf{Z}))^2}\mathbf{Z}\mathbf{Z}'\Big]$$

is positive definite, and then $A_n(\beta_0)$ is negative definite. Because $\overline{\mathbf{S}}_n(\beta_0)=0$, it follows that $n^{-1}\overline{\mathbf{S}}_n(\beta)$ is bounded away from zero for any $\beta\ne\beta_0$. Coupled with (A.1) and the definition of $\widehat{\beta}$, implies that $\widehat{\beta}\to\beta_0$ a.s..

A.3 Asymptotic Normality of $n^{-1/2}\mathbf{S}_n(\beta_0)$

Since $n^{-1/2}\mathbf{S}_n(\beta_0)$ is a sum of dependent random variable, we will look for a sum of independent random variable to approximate it. By definition,

$$\begin{aligned}\mathbf{S}_n(\beta_0)&=\sum_{i=1}^{n}\Big[\frac{I(N_i\ge\beta_0'\mathbf{Z}_i)}{\widehat{G}(\beta_0'\mathbf{Z}_i)+\widehat{H}(\beta_0'\mathbf{Z}_i)}-\frac{1}{2}\Big]\mathbf{Z}_i\\&=\sum_{i=1}^{n}\Big[\frac{I(N_i\ge\beta_0'\mathbf{Z}_i)}{G(\beta_0'\mathbf{Z}_i)+H(\beta_0'\mathbf{Z}_i)}-\frac{1}{2}\Big]\mathbf{Z}_i-n\int_{-\infty}^{\infty}\frac{\widehat{G}(t)-G(t)}{(\widehat{G}(t)+\widehat{H}(t))(G(t)+H(t))}\mathrm{d}\mathbf{Q}(\mathbf{t})\\&\quad-n\int_{-\infty}^{\infty}\frac{\widehat{H}(t)-H(t)}{(\widehat{G}(t)+\widehat{H}(t))(G(t)+H(t))}\mathrm{d}\mathbf{Q}(\mathbf{t}),\end{aligned}$$

where $\mathbf{Q}(t)=n^{-1}\sum_{i=1}^{n} I(\beta_0'\mathbf{Z}_i\le t\vee N_i)\mathbf{Z}_i$.

Note that empirical distribution is a special case of Kaplan-Meier estimator, which means $\widehat{G}$ and $\widehat{H}$ are also the Kaplan-Meier estimators for G and H, respectively. Then, applying a well-known martingale representation for both $(\widehat{G}-G)/G$ and $(\widehat{H}-H)/H$ (see, for example, Gill [27]) and using the similar arguments in Appendix B in Ying et al. [16], we can show that the statistic $\mathbf{S}_n(\beta_0)$ is asymptotically equivalent to $\sum_{i=1}^{n}\eta_i$, where

$$\begin{aligned}\eta_i \;=\;& \Big[\frac{I(N_i\ge\beta_0'\mathbf{Z}_i)}{G(\beta_0'\mathbf{Z}_i)+H(\beta_0'\mathbf{Z}_i)}-\frac{1}{2}\Big]\mathbf{Z}_i-\frac{1}{2}\int_{-\infty}^{\infty}\frac{\mathbf{q}_1(\mathbf{t})}{h_1(t)}\Big[\mathrm{d}I(Y_i\le t)-I(Y_i\ge t)\mathrm{d}\Lambda_1(t)\Big]\\&-\frac{1}{2}\int_{-\infty}^{\infty}\frac{\mathbf{q}_2(\mathbf{t})}{h_2(t)}\Big[\mathrm{d}I(X_i\le t)-I(X_i\ge t)\mathrm{d}\Lambda_2(t)\Big],\end{aligned}$$

and

$$\mathbf{q}_1(\mathbf{t})=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{G(\beta_0'\mathbf{Z}_i)}{G(\beta_0'\mathbf{Z}_i)+H(\beta_0'\mathbf{Z}_i)}I(\beta_0'\mathbf{Z}_i\ge\mathbf{t})\mathbf{Z}_i,\qquad h_1(t)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}I(Y_i\ge t);$$

$$\mathbf{q}_2(\mathbf{t})=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{H(\beta_0'\mathbf{Z}_i)}{G(\beta_0'\mathbf{Z}_i)+H(\beta_0'\mathbf{Z}_i)}I(\beta_0'\mathbf{Z}_i\ge\mathbf{t})\mathbf{Z}_i,\qquad h_2(t)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}I(X_i\ge t),$$

$\Lambda_1(\cdot)$ and $\Lambda_2(\cdot)$ are the cumulative hazard functions for the right and left censoring variables, respectively. Thus, by the multivariate central limit theorem we can show that the distribution of $n^{-1/2}\mathbf{S}_n(\beta_0)$ is asymptotically normal with mean zero and some variance-covariance matrix Γ.

A.4 Local Linearity for $\mathbf{S}_n(\beta)$ and Asymptotic Normality for $\widehat{\beta}$

Before we show the local linearity of $\mathbf{S}_n(\beta)$, two lemmas are needed in the proof.

Lemma A.1 Let μ be a continuously differentiable function. Then for any fixed constant c,

$$\sup_{|s-t|\le cn^{-1/3},s,t\le\widetilde{t}}|\mu(\widehat{G}(t)+\widehat{H}(t))-\mu(G(t)+H(t))-\mu(\widehat{G}(s)+\widehat{H}(s))+\mu(G(s)+H(s))|=o_p(n^{-1/2}).$$

Lemma A.2 Let ν_i be a sequence of constants. Then

$$\begin{aligned}\sup_{\|\beta-\beta_0\|\le cn^{-1/3}}\Big|&\sum_{i=1}^{n}\nu_iI(N_i\ge\beta'\mathbf{Z}_i)-\sum_{i=1}^{n}\nu_iI(N_i\ge\beta_0'\mathbf{Z}_i)+\sum_{i=1}^{n}\nu_i\frac{1}{2}(G(\beta_0'\mathbf{Z}_i)+H(\beta_0'\mathbf{Z}_i))\\&-\sum_{i=1}^{n}\nu_i[(1-F_i((\beta-\beta_0)'\mathbf{Z}_i))G(\beta'\mathbf{Z}_i)+F_i((\beta-\beta_0)'\mathbf{Z}_i)H(\beta'\mathbf{Z}_i)]\Big|=o_p(n^{1/2}).\end{aligned}$$

In particular we have

$$\sup_{\|\beta-\beta_0\|\le cn^{-1/3}}\sum_{i=1}^{n}|I(N_i\ge\beta'\mathbf{Z}_i)-I(N_i\ge\beta_0'\mathbf{Z}_i)|=O_p(n^{2/3}).$$

Recall that β is in the $n^{-1/3}$-neighborhood of β_0. Therefore, by Lemma A.1 and A.2, we can show that

$$\mathbf{S}_n(\beta) = \mathbf{S}_n(\beta_0) + \overline{\mathbf{S}}_n(\beta) + o_p(n^{1/2}). \tag{A.2}$$

The details of proof are similar to Appendix C in Ying et al. [16], we omit them here.

Supposing that $\mathbf{A} = \lim\limits_{n\to\infty} \mathbf{A}_n(\beta_0)$ to be nonsingular, for any fixed constant c and all β in $\|\beta - \beta_0\| < cn^{-1/3}$, by taking Taylor's expansion of $\overline{\mathbf{S}}_n(\beta)$ in (A.2) at β_0, we get that

$$\mathbf{S}_n(\beta) = \mathbf{S}_n(\beta_0) + n\mathbf{A}(\beta - \beta_0) + o_p(\max(\sqrt{n}, n\|\beta - \beta_0\|)).$$

Since $n^{-1/2}\mathbf{S}_n(\beta_0) \to N(0, \Gamma)$ in distribution as $n \to \infty$, $n^{1/2}(\widehat{\beta} - \beta_0) \to N(0, \mathbf{A}^{-1}\Gamma\mathbf{A}^{-1})$ in distribution. This completes the proof of asymptotic normality for $\widehat{\beta}$.

A.5 Proof of Theorem 3.1

We now have the asymptotic Normality of both $n^{-1/2}\mathbf{S}_n(\beta_0)$ and $\widehat{\beta} - \beta_0$. The later implies that $\widehat{\beta} - \beta_0 = O_p(n^{-1/2})$. In order to prove Theorem 3.1, two other lemmas are needed.

Lemma A.3 Under the conditions as in Theorem 3.1, we have

$$\text{(i)}\ \frac{1}{n}\sum_{i=1}^n S_{ni}(\beta_0)S_{ni}(\beta_0)' \to \Gamma_1, \qquad \text{(ii)}\ \widehat{\Gamma}_1 \to \Gamma_1$$

in probability as $n \to \infty$.

Proof Similar to the proof of Lemma A.1 in Qin and Tsao [24]. □

Lemma A.4 Let $X \to N(0, I_p)$ in distribution as $n \to \infty$, where I_p is the $p \times p$ identity matrix. Let Σ be a $p\times p$ nonnegative definite matrix with eigenvalues $l_1, l_2, \cdots, l_p$, Then,

$$X'\Sigma X \to l_1\chi^2_{1,1} + l_2\chi^2_{2,1} + \cdots + l_p\chi^2_{p,1}$$

in distribution as $n \to \infty$, where $\chi^2_{i,1}$ $(i = 1, 2, \cdots, p)$ are independent chi-square random variables each with one degree of freedom.

Proof See Qin and Jing [28]. □

Now, by using the above two lemmas and the similar argument in appendix in Qin and Tsao [24], we have

$$\begin{aligned} l_n(\beta_0) &= \sum_{i=1}^n \lambda' S_{ni}(\beta_0)S_{ni}(\beta_0)'\lambda + o_p(1) \\ &= \Big(\Gamma^{-1/2}n^{-1/2}\sum_{i=1}^n S_{ni}(\beta_0)\Big)'(\Gamma^{1/2}\Gamma_1^{-1}\Gamma^{1/2})\Big(\Gamma^{-1/2}n^{-1/2}\sum_{i=1}^n S_{ni}(\beta_0)\Big) + o_p(1). \end{aligned}$$

Because of the asymptotic Normality of $n^{-1/2}\mathbf{S}_n(\beta_0)$, we have $\Gamma^{-1/2}(n^{-1/2}\sum_{i=1}^{n} S_{ni}(\beta_0)) \to N(0, I_{p+1})$ in distribution as $n \to \infty$. Also note that $\Gamma^{1/2}\Gamma_1^{-1}\Gamma^{-1/2}$ and $\Gamma_1^{-1}\Gamma$ have the same eigenvalues. Thus, from Lemma A.4, Theorem 3.1 is proved.

References

[1] Gehan, E.A., A generalized Wilcoxon test for comparing doubly censored data, *Biometrika*, **52**(1965), 650–653.

[2] Turnbull, B.W., Nonparametric estimation of a survivorship function with doubly censored data, *J. Amer. Statist. Assoc.*, **69**(1974), 169–173.

[3] DeGruttola, V. and Lagakos, S.W., Analysis of doubly-censored survival data, with application to AIDS, *Biometrics*, **45**(1989), 1–11.

[4] Zhang, C.H. and Li, X., Linear regression with doubly censored data, *Ann. Statist.*, **24**(1996), 2720–2734.

[5] Cai, T. and Cheng, S., Semiparametric regression analysis for doubly censored data, *Biometrika*, **91**(2004), 277–290.

[6] Leiderman, P.H., Babu, D., Kagia, J., Kraemer, H.C. and Leiderman, G.F., African infant precocity and some social influences during the first year, *Nature*, **242**(1973), 247–249.

[7] Miller, R.G., Least squares regression with censored data, *Biometrika*, **63**(1976), 449–464.

[8] Buckley, J. and James, I., Linear regression with censored data, *Biometrika*, **66**(1979), 429–436.

[9] Leurgans, S., Linear models, random censoring and synthetic data, *Biometrika*, **74**(1987), 301–309.

[10] Meier, P., Estimation of distribution function from incomplete observations, in *Perspectives in Probability and Statistics* (1975), ed. J. Gani, London: Academic Press, 67–87.

[11] Bassett, G. and Koenker, R., Asymptotic theory of least absolute error regression, *J. Amer. Statist. Assoc.*, **73**(1978), 618–622.

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

[13] Chen, X., Bai, Z., Zhao, L. and Wu, Y., Asymptotic normality of minimum L_1-norm estimates in linear models, *Sci. China*, **33**(1990), 1311–1328.

[14] Powell, J.L., Least absolute deviations estimation for censored regression model, *J. Econometrics*, **25**(1984), 303–325.

[15] Powell, J.L., Censored regression quantiles, *J. Econometrics*, **32**(1986), 143–155.

[16] Ying, Z., Jung, S.H. and Wei, L.J., Survival analysis with median regression models, *J. Amer. Statist. Assoc.*, **90**(1995), 178–184.

[17] Jung, S., Quasi-likelihood for median regression models, *J. Amer. Statist. Assoc.*, **91**(1996), 251–257.

[18] Yang, S., Censored median regression using weighted empirical survival and hazard functions, *J. Amer. Statist. Assoc*, **94**(1999), 137–145.

[19] Bang, H. and Tsiatis, A.A., Median regression with censored cost data, *Biometrics*, **58**(2002), 643–649.

[20] Portnoy, S., Censored regression quantiles, *J. Amer. Statist. Assoc*, **98**(2003), 1001–1012.

[21] Liu, W. and Ren, Y., Empirical likelihood infernece for censored median regression model, *Mathematica Applicata*, **18**(2005), 489–496.

[22] Owen, A.B., Empirical likelihood ratio confidence intervals for single functional, *Biometrika*, **75**(1988), 237–249.

[23] Owen, A.B., *Empirical Likelihood*, Chapman and Hall, 2001.

[24] Qin, G. and Tsao, M., Empirical likelihood inference for median regression models for censored survival data, *J. Multivariate Anal.*, **85**(2003), 416–430.

[25] Shao, J. and Tu, D., *The Jackknife and Bootstrap*, New York: Springer-Verlag, 1995.

[26] Kaplan, E.L. and Meier, P., Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**(1958), 457–481.

[27] Gill, R.D., Large sample behaviour of the product-limit estimator on the whole line, *Ann. Statist.*, **11**(1993), 49–58.

[28] Qin, G. and Jing, B., Empirical likelihood for censored linear regression, *Scand. J. of Statist.*, **28**(2001), 661–673.

双侧截断数据的中位数回归

郑 明 郁 文

(复旦大学管理学院统计学系, 上海, 200433)

在很多实际应用中, 个体寿命时间可能被同时左截断与右截断. 本文在左右截断变量都能被观察到的假设下, 提出了一种半参数推断方法, 来分析协变量对于相应寿命时间或其某种变换的中位数的影响, 并讨论了所得估计量的渐近性质. 此外, 本文还提供了一种基于经验似然的回归参数推断方法, 并讨论了将这些方法推广到经典双侧截断数据的可能性. 一些模拟计算被用于展示这些方法的有效性.

关键词: 左截断与右截断, 经验似然, 最小绝对离差, 中位数回归.

学科分类号: O212.1.