Journal of Multivariate Analysis 101 (2010) 240–251

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Empirical likelihood for median regression model with designed censoring variables

Pingshou Zhong [a,b], Hengjian Cui [a,*]

[a] Department of Statistics and Financial Mathematics, Beijing Normal University, Beijing 100875, China

[b] Department of Statistics, Iowa State University, Ames, Iowa 50011, USA

ARTICLE INFO

Article history:
Received 1 November 2008
Available online 3 August 2009

AMS 2000 subject classifications:
62G20
62N02

Keywords:
Empirical likelihood
Designed censoring
Fixed censoring
Median regression model
Confidence region

ABSTRACT

We propose a new and simple estimating equation for the parameters in median regression models with designed censoring variables, and then apply the empirical log likelihood ratio statistic to construct confidence region for the parameters. The empirical log likelihood ratio statistic is shown to have a standard chi-square distribution, which makes this method easy to implement. At the same time, another empirical log likelihood ratio statistic is proposed based on an existing estimating equation and the limiting distribution of the empirical likelihood ratio statistic is shown to be a sum of weighted chi-square distributions. We compare the performance of the empirical likelihood confidence region based on the new estimating equation, with that based on the existing estimating equation and a normal approximation method by simulation studies.

1. Introduction

Censored data often appear in econometrics, medical follow-up research, industrial life-testing and other studies. There are several mechanisms that may lead to censoring. The most simple case is the Type I fixed censoring, in which a sample of subjects are followed for a fixed time C. For example, we follow up with some unemployed people to observe the duration time without jobs in one study, but the study ends while still a certain percentage remains unemployed. So the only thing that we know about them is that the duration times of those people are greater than the design censoring time C (for more examples, see Lee [1]).

The designed censoring we mentioned in the title is a generalization of the fixed censoring scheme, which is also called the generalized Type I fixed censoring in some literature. Each subject in the studies has a potential censoring time, which may vary from one subject to another but is nevertheless known by design. For example, in many clinical trials and epidemiological studies, it is difficult to enroll all subjects simultaneously and recruitment usually takes place over an interval, but the study ends at a fixed time point, which results in different censoring times for different subjects. Another generalization of the fixed censoring is random censoring, where it is assumed that the censoring time is not known for the subject for which we have complete data. However, the fact that the follow-up time is designed in advance may be an important practical advantage in a follow-up study. In the present paper, we focus on the designed censoring and regard the censoring times as design points generated from some distribution.

Suppose, in the example mentioned in the first paragraph, we are interested in analyzing the relationship between the duration time and some exogenous variables, which can be years of education, age, *etc*. We may consider a linear regression

* Corresponding author.
E-mail addresses: pszhong@iastate.edu (P. Zhong), hjcui@bnu.edu.cn (H. Cui).

0047-259X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.07.008

between the duration time and the exogenous variables. There are two common ways to estimate the parameters in a linear regression with fixed censoring data. One way is the maximum likelihood method, assuming that the error terms are normally distributed (see Heckman [2,3]). It is well known that this method is not robust. Another way is to apply more robust methods. The most often used method is median regression, which assumes that the median of the response, such as the duration time in the example, is a parametric function of the covariates. Many authors including Powell [4], Rao [5], McKeague [6] and Subramanian [7] applied median regression in their research.

Our objective in this paper is to conduct inference on the regression parameters in a linear median regression model with designed censoring variables. Powell [4] considered the least absolute deviation (LAD) method in censored regression models with fixed censoring variables and established the asymptotic normality of the LAD estimator. They assumed that the responses are censored at 0. Zhou and Wang [8] discussed the LAD estimators for the parameters in nonlinear regression models with designed censoring variables. They proved the asymptotic normality of the estimator, but the asymptotic covariance matrices depend on the error density and are therefore difficult to estimate reliably. Hence, it is not easy to use the asymptotic normality for statistical inference in practice. It is also well known that confidence regions based on the asymptotic normality could encounter large coverage errors in small and medium sample sizes.

To overcome the difficulty of variance estimation in the normal approximation inference method, we consider an empirical likelihood based inference as an alternative for the parameters in median regression with designed censoring variables. Owen [9] introduced empirical likelihood as a general inference procedure for the parameters defined in estimating equations. Since then, empirical likelihood has proven to be useful in diverse statistical applications, for example, Chen and Hall [10], Cui and Chen [11], Hall and La Scala [12], Qin and Lawless [13], Qin and Tsao [14], Shi and Lau [15], among others. Furthermore, empirical likelihood has some attractive properties. For example, the empirical log likelihood ratio satisfies the nonparametric Wilks' theorem (Owen [9]) and the confidence regions based on the empirical likelihood are Bartlett correctable (DiCicco, Hall and Romano [16], Chen and Cui [17]).

Qin and Tsao [14] applied the empirical likelihood method for parameters in median regression models with censored data based on the estimating equation proposed by Ying et al. [18]. However, the estimating equation was proposed for the median regression models with random censoring. We adapt the estimating equation to our data structure, which involved with a secondary estimation of the censoring distribution. Due to the different data structure, we use the empirical cumulative distribution to replace the Kaplan–Meier estimate for the censoring distribution in our proposed empirical log likelihood ratio statistic. We show that the limiting distribution of the empirical log likelihood ratio is a weighted sum of chi-square distributions rather than a standard chi-square distribution.

The main advantage of the empirical likelihood based on the new estimating equation for parameters in median regression models with designed censoring variables is that the resulting empirical log likelihood ratio statistic satisfies the standard nonparametric Wilks' theorem. No secondary estimation is needed when applying the new empirical likelihood method and thus it is more convenient to use. The new proposed empirical likelihood approach is more accurate than the normal approximation in many situations when the sample size is not large and the underlying distribution is non-normal. In addition, it can be readily used when the dependency exists between censoring variables and covariates without modification.

This paper is organized as follows. We introduce the models and give a new and simple estimating equation in Section 2. In Section 3, three ways are used to construct confidence regions. The asymptotic distributions of two empirical likelihood based statistics for inference are derived and the main results are given. We present some simulation results in Section 4. Summary of the paper is given in Section 5. All the conditions and technical proofs are put in the Appendix.

2. Models and a new estimating equation

Let T_i be the response of interest, for example, T_i may be the duration time of unemployment for individual i ($i = 1, \ldots, n$). Let $\mathbf{Z}_i$ be exogenous covariates thought to influence the response. We want to study the relationship between T_i and $\mathbf{Z}_i$. We assume that median of T_i is a linear function of $\mathbf{Z}_i$, specifically,

$$T_i = \beta_0' \mathbf{Z}_i + e_i, \quad \text{for } i = 1, \ldots, n \tag{1}$$

where $e_1, \ldots, e_n$ are i.i.d. random variables with zero median and have continuous density $h(t)$ satisfying condition (A1) in the Appendix, β_0 is a $p \times 1$ vector.

Due to designed censoring, we do not observe T_i directly if T_i is greater than C_i. Instead, we observe the vector $(Y_i, \mathbf{Z}_i', \delta_i, C_i)$, where $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$ is an indicator variable. We assume that $C_1, \ldots, C_n$ are design fixed constants generated from a distribution $G(t)$.

We define the estimator $\hat{\beta}_n$ of β_0 to be a solution of the equation

$$\bar{g}(\beta) \equiv \frac{1}{n} \sum_{i=1}^{n} g_i(\beta) = 0, \tag{2}$$

where $g_i(\beta) = \mathrm{sgn}(Y_i - \min(C_i, \beta' \mathbf{Z}_i))(1 + \mathrm{sgn}(C_i - \beta' \mathbf{Z}_i))\mathbf{Z}_i$ and $\mathrm{sgn}(x) = I(x \ge 0) - I(x \le 0)$, the sign function. Because of the discontinuity of $\bar{g}(\beta)$, an exact solution to (2) may not exist, but we can define the estimator $\hat{\beta}_n$ satisfying $\bar{g}(\hat{\beta}_n) \approx 0$.

There are two primary justifications for the estimating equation. Firstly, the estimation equation approximate the derivation of the objective function of the LAD estimator of β_0

$$\hat{\beta}_{LAD} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} |Y_i - \min(C_i, \beta'\mathbf{Z}_i)|. \tag{3}$$

The definition of the LAD estimator for this model is based on the fact that, for any scalar random variable Y satisfying $E|Y| < +\infty$, the function $E|Y-b|$ is minimized by choosing b to be the median of the distribution of Y. Hence, if the median of Y given C, $\mathbf{Z}$ and β, is some known function $m(C, \mathbf{Z}, \beta)$ of the censoring variables, the regressors and unknown parameters, a sample analogue to the conditional median can be defined by choosing β to minimize the function $\sum_{i=1}^{n} |Y_i - m(C_i, \mathbf{Z}_i, \beta)|$. Using a method similar to that of Powell [4], under condition (A1) in the Appendix, we can verify that the conditional median (given C_i and $\mathbf{Z}_i$) of Y_i is $m(C_i, \mathbf{Z}_i, \beta_0) = \min(C_i, \beta_0'\mathbf{Z}_i)$. Secondly, let Ω be a bounded parameter space which contains β_0 as an interior point. Then under some mild conditions (A1)–(A4) in the Appendix, we may show that the $E[\bar{g}(\beta)] = 0$ if and only if $\beta = \beta_0$ for $\beta \in \Omega$ (see Lemma 1 in the Appendix). So $\bar{g}(\beta) = 0$ is also a reasonable estimating equation for β_0 from this standpoint.

Ying et al. [18] have established an estimating equation for parameters in the median regression model under random censoring. Under the assumption that C_i are independent of $\mathbf{Z}_i$ and T_i are independent of C_i given $\mathbf{Z}_i$, they used the fact that $E\{I(Y_i \geq \beta_0'\mathbf{Z}_i)|\mathbf{Z}_i\} = (1 - G(\beta_0'\mathbf{Z}_i))/2$ where $G(\cdot)$ is the distribution function of C_i. However, the relationship is no longer hold if C_i and $\mathbf{Z}_i$ are dependent. In fact, we need to estimate the conditional distribution of C_i condition on $\mathbf{Z}_i$ when C_i are dependent on $\mathbf{Z}_i$, which could be difficult in some cases.

3. Confidence regions and main results

3.1. Empirical likelihood based on the new estimating equation

Recall that from the discussion in Section 2, under some conditions, $E[\bar{g}(\beta)] = 0$ if and only if $\beta = \beta_0$. This motivates us to construct empirical likelihood confidence region for β_0 in model (1). According to Owen [9], the empirical likelihood ratio for β can be defined as

$$R(F) = \max_{p_i} \left\{ \prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g_i(\beta) = 0, p_i > 0 \right\},$$

which corresponds to the empirical log likelihood ratio evaluated at β, that is,

$$\ell(\beta) = -2 \min_{\substack{\sum_{i=1}^{n} p_i = 1, \\ \sum_{i=1}^{n} p_i g_i(\beta) = 0}} \sum_{i=1}^{n} \log(np_i).$$

By introducing a Lagrange multiplier $\lambda \in \mathbb{R}^p$, standard derivations in empirical likelihood lead to

$$\ell(\beta) = 2\sum_{i=1}^{n} \log\{1 + \lambda' g_i(\beta)\},$$

where λ satisfies

$$\sum_{i=1}^{n} \frac{g_i(\beta)}{1 + \lambda' g_i(\beta)} = 0.$$

We investigate the nonparametric version of Wilks' theorem for the empirical log likelihood ratio statistic, the convexity of the confidence region based on the empirical log likelihood statistic and the power of the empirical log likelihood ratio test in the following theorems.

Theorem 1. *Under conditions* (A1)–(A4) *in the Appendix, if β_0 is the true value of β, then the limiting distribution of $\ell(\beta_0)$ is chi-square distributed with p degrees of freedom, that is as $n \to \infty$, $\ell(\beta_0) \xrightarrow{d} \chi_p^2$, where $\xrightarrow{d}$ stands for converging in distribution.*

We note that the limiting distribution of $\ell(\beta_0)$ is a standard chi-squared distribution, which is free of any tuning parameter. This is due to the fact that $g_i(\beta_0)$ are independent random variables without any unknown parameter. More specific reason is that $\frac{1}{n}\sum g_i(\beta_0)g_i'(\beta_0) \to V_{\beta_0}$ a.s. and $\frac{1}{\sqrt{n}}\sum g_i(\beta_0) \xrightarrow{d} N(0, V_{\beta_0})$. According to the proof given in the Appendix,

$$\ell(\beta_0) = \left\{\frac{1}{\sqrt{n}}\sum g_i'(\beta_0)\right\} \left\{\frac{1}{n}\sum g_i(\beta_0)g_i'(\beta_0)\right\}^{-1} \left\{\frac{1}{\sqrt{n}}\sum g_i(\beta_0)\right\} + o_p(1).$$

Hence, $\ell(\beta_0) \xrightarrow{d} \chi^2_p$. By applying Theorem 1, we can construct a $1-\alpha$ confidence region for β_0 as

$$CR_\alpha = \{\beta : \ell(\beta) \le c_\alpha\}, \tag{4}$$

where c_α is the $1-\alpha$ quantile of a χ^2_p distribution satisfying $P\{\chi^2_p < c_\alpha\} = 1-\alpha$.

Usually convex confidence regions are appealing for purpose of interpretation. In particular, one would hope that as the sample size increases and we acquire more information near β_0, this increased information would be reflected in a high probability of obtaining a convex and more accurate confidence regions. Theorem 2 says this fact.

Theorem 2. *CR_α is asymptotically convex, that is the gap between CR_α and a convex region attract zero probability, as $n \to \infty$.*

Let us consider the problem of testing the null hypothesis $H_0 : \beta = \beta_0$. We can use the empirical log likelihood ratio $\ell(\beta)$ as a test statistic. Therefore, if we want to know the power of the test, the following theorem should be considered.

Theorem 3. *Under conditions* (A1)–(A4) *in the Appendix,* $\lim_{n\to\infty} P\{\tilde\beta \in CR_\alpha\} = 0$ *for any fixed* $\tilde\beta \ne \beta_0$ *and* $\lim_{n\to\infty} P\{\tilde\beta_n \in CR_\alpha\} = P\{\chi^2_p(\|\gamma\|^2) < c_\alpha\}$ *for any fixed* $\tilde\beta_n = \beta_0 + \frac{1}{\sqrt{n}} h(0)^{-1} V_{\beta_0}^{-1/2}\gamma$, *where* $\chi^2_p(\|\gamma\|^2)$ *stands for the noncentral* χ^2 *random variable with p degrees of freedom and noncentrality parameter* $\|\gamma\|^2$, *for a fixed* $\gamma \in \mathbb{R}^p$ *and* $V_{\beta_0} = \lim_{n\to\infty} \frac{4}{n}\sum_{i=1}^n E[\mathbf{Z}_i\mathbf{Z}_i'(1 - G(\beta_0'\mathbf{Z}_i))]$.

The first result of Theorem 3 reveals that the power of test $H_0 : \beta = \beta_0$ is asymptotically 1 as $n \to \infty$. The second result of this Theorem gives the asymptotic distribution of $\ell(\beta)$ under the local alternative $H_1 : \beta = \beta_0 + \frac{1}{\sqrt{n}} h(0)^{-1} V_{\beta_0}^{-1/2}\gamma$, where V_{β_0} was given in Theorem 3. The result tells us that the empirical log likelihood ratio test has a non-trivial power to test the departure from the null hypothesis of order $O(1/\sqrt{n})$.

3.2. Empirical likelihood based on the existing estimating equation

In this section, we present an alternative way to construct empirical likelihood confidence region based on the estimating equation proposed by Ying et al. [18]. The estimating equation was proposed for a different data structure with random censoring, i.e., C_i are only observed when $\delta_i = 0$. Based on this estimating equation, Qin and Tsao [14] have constructed an empirical likelihood confidence region for the regression coefficients in median regression with random censoring. However, in designed censoring situation, the complete observations of C_i can be obtained. It seems reasonable and useful to use all the observations of C_i rather than part of them. By replacing the Kaplan–Meier estimate in Ying et al. with an empirical distribution function of C_i, i.e., $\hat G_n(t) = \frac{1}{n}\sum_{i=1}^n I(C_i \le t)$, we propose the following estimating equation

$$W_n(\beta) = \frac{1}{n}\sum_{i=1}^n W_{ni}(\beta) \equiv \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\left[I(Y_i \ge \beta'\mathbf{Z}_i) - \frac{1}{2}(1 - \hat G_n(\beta'\mathbf{Z}_i))\right] = 0. \tag{5}$$

As the same as Section 3.1, we can define the following empirical log likelihood ratio based on (5),

$$\ell_E(\beta) = 2\sum_{i=1}^n \log\{1 + \lambda_E' W_{ni}(\beta)\},$$

where λ_E satisfies $\sum_{i=1}^n \frac{W_{ni}(\beta)}{1+\lambda_E' W_{ni}(\beta)} = 0$.

Let $B_{ij}(\beta_0) = G(\beta_0'\mathbf{Z}_i)(1 - G(\beta_0'\mathbf{Z}_j))$ and denote

$$\Gamma_1 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n E\left[\left\{I(Y_i \ge \beta_0'\mathbf{Z}_i) - \frac{1}{2}(1 - G(\beta_0'\mathbf{Z}_i))\right\}^2 \mathbf{Z}_i\mathbf{Z}_i'\right]$$

$$\Gamma_2 = \lim_{n\to\infty} \frac{1}{4n^2}\sum_{i\ne j} E\left[B_{ij}(\beta_0)(\mathbf{Z}_i\mathbf{Z}_j' + \mathbf{Z}_j\mathbf{Z}_i') - \min\{B_{ij}(\beta_0), B_{ji}(\beta_0)\}\mathbf{Z}_i\mathbf{Z}_j'\right].$$

We give the following limiting distribution of $\ell_E(\beta_0)$:

Theorem 4. *Suppose that the conditions* (A1)–(A4) *in the Appendix hold. If β_0 is the true value of β, then the limiting distribution of $\ell_E(\beta_0)$ is a weighted sum of chi-square distributions with 1 degree of freedom, that is,*

$$\ell_E(\beta_0) \xrightarrow{d} l_1\chi^2_{1,1} + \cdots + l_p\chi^2_{p,1},$$

where the weights l_i's are the eigenvalues of $\Gamma_1^{-1}(\Gamma_1 - \Gamma_2)$ and $\chi^2_{i,1}(i = 1, 2, \ldots, p)$ are independent chi-square random variables each with one degree of freedom.

To apply Theorem 4 for constructing confidence region for β_0, we have to estimate the weights l_i. We first estimate Γ_1 and Γ_2 by replacing $G(t)$ with $\hat G_n(t)$, the empirical cumulative distribution function and β_0 with $\hat\beta_{LAD}$ or $\hat\beta_n$, a solution to (5).

Then estimating the l_i's by the eigenvalues of $\hat{\Gamma}_1^{-1}(\hat{\Gamma}_1 - \hat{\Gamma}_2)$. Now a $1-\alpha$ confidence region for β_0 can be formed as

$$CR_\alpha^E = \{\beta : \ell_E(\beta) < c_\alpha^E\},$$

where c_α^E is the $1-\alpha$ quantile of the weighted sum of chi-square distributions $\hat{l}_1\chi_{1,1}^2 + \cdots + \hat{l}_p\chi_{p,1}^2$.

As compared with Theorem 1, we notice that the limiting distribution is a weighted sum of chi-square distributions rather than a standard chi-square distribution. This is essentially due to the estimation of $G(\cdot)$, which causes the dependency between $W_{ni}(\beta_0)$. Because of the estimation of Γ_1, Γ_2, β_0 and c_α^E, the procedure for constructing confidence region CR_α^E become much complicated than using CR_α as a confidence region.

3.3. Normal approximation based confidence region

Under the same data structure with us and certain conditions, Zhou and Wang [8] have shown that

$$\sqrt{n}(\hat{\beta}_{LAD} - \beta_0) \xrightarrow{d} N(0, (4h^2(0)S)^{-1}), \tag{6}$$

where $S = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i' \int_{(\beta_0'\mathbf{Z}_i)_-}^{+\infty} dG(u)$ and $(a)_- = -\min(a, 0)$.

Now let

$$\widehat{S} = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i' \int_{(\hat{\beta}_{LAD}'\mathbf{Z}_i)_-}^{+\infty} d\hat{G}(u), \qquad \hat{h}(0) = \int K_b(u)d\hat{F}^{KM}(u),$$

where $\hat{G}(u) = \frac{1}{n}\sum_{i=1}^n \int_{-\infty}^u K_a(t - C_i)dt$, $\hat{F}^{KM}(u)$ is the Kaplan–Meier estimate of the distribution function of $T_i - \hat{\beta}_{LAD}'\mathbf{Z}_i$ and $K_a(\cdot) = K(\cdot/a)/a$, $K(\cdot)$ is a symmetric probability kernel function and a, b are two bandwidths.

Therefore, we can formulate the following normal approximation based confidence region with significant level $1-\alpha$:

$$NCR_\alpha = \{\beta : 4n\hat{h}^2(0)(\hat{\beta}_{LAD} - \beta)'\hat{S}(\hat{\beta}_{LAD} - \beta) \le c_\alpha\},$$

where c_α is the $1-\alpha$ quantile of the standard chi-square distribution with p-degrees of freedom.

The empirical likelihood for constructing confidence regions in the previous sections is known as a computer intensive method. However, to apply the normal approximation method to construct confidence region for β_0, one also has to estimate β_{LAD} first. Because the objective function $F_n(\beta) = \sum_{i=1}^n |Y_i - \min(C_i, \beta'\mathbf{Z}_i)|$ is not convex, Newton-type algorithm is not guaranteed to converge to a global minimum, we applied the Genetic algorithm to find the global minimum (see, Zhou and Wang [19]). This increases computation time for using the normal approximation method.

4. Simulation study

In this section, we use Monte Carlo simulation to evaluate the performance of the empirical likelihood methods proposed in Sections 3.1 and 3.2, and compare them with the normal approximation inference method of Zhou and Wang [8]. Throughout this section, we use NCR_α to denote the $1-\alpha$ confidence region constructed by the normal approximation method, CR_α for the $1-\alpha$ empirical likelihood based confidence region proposed in Section 3.1 and CR_α^E for the $1-\alpha$ empirical likelihood confidence region proposed in Section 3.2. Let R_α be the $1-\alpha$ empirical likelihood confidence region for median regression with random censoring proposed by Qin and Tsao.

Firstly, to show the asymptotic convexity of the confidence region CR_α and compare the coverage probabilities and interval lengths between CR_α and NCR_α, we simulated the following model:

Model I: Let $T_i = \beta_0 Z_i + e_i$, where $\beta_0 = 1$, $Z_i \sim U[0, 1]$ and $e_i \sim N(0, 1)$. The censoring variable $C_i \sim \log U_i$, where $U_i \sim U[0, c]$ and $Y_i = \min(T_i, C_i)$. The value of the constant c in the model determines the censoring proportion.

In Fig. 1, we present two curves to demonstrate the asymptotic convexity of the confidence region CR_α for β_0 in model I with censoring rate 15%. We see that as the sample size increased from 60 to 300, the curve near the true value is almost convex. We also note that the confidence interval became narrower as the sample size increased and both of them included the true value $\beta_0 = 1$, as we would expect.

Next, we compare CR_α with NCR_α using model I. The coverage probabilities of CR_α and NCR_α are summarized in Table 1, which were estimated by the frequency of the true values falling into the confidence intervals in 1000 simulations. On our PC with 2 G memory and Vista system, the average CPU time for constructing one confidence interval using NCR_α in model I was 0.42 s, while the CPU time using CR_α was 1.36 s. The confidence intervals based on CR_α were consistently superior than that based on NCR_α in term of coverage probabilities. The coverage probabilities of NCR_α were usually lower than the nominal levels, especially when the censoring proportion was large. In contrast, the intervals CR_α maintained the nominal level well even as the censoring proportion increases.

Furthermore, the improved coverage probabilities of CR_α does not come at the expense of the increased interval width. Table 1 also compares the length of CR_α (denote as L_{el}) with the length of NCR_α (denote as L_{norm}) for various sample sizes, censoring proportions and nominal levels based on 100 simulations. It revealed that the length of CR_α usually approximately equals to the length of NCR_α, even with 40% censoring proportion. In small censoring proportion (15%), the length of NCR_α was usually longer than that of CR_α. As the sample size increased, the difference between L_{norm} and L_{el} decreased.

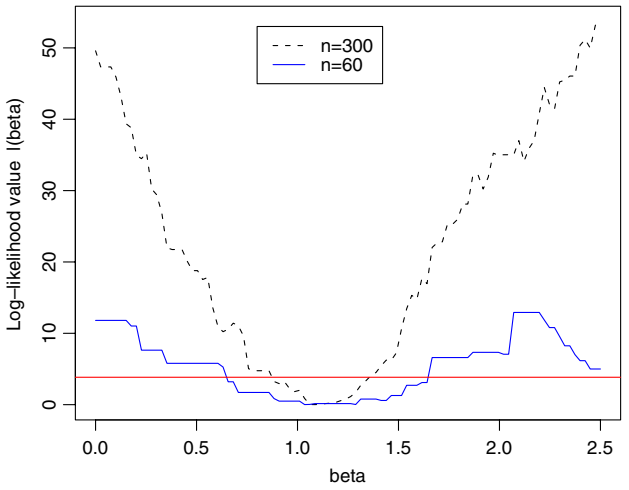

Fig. 1. The asymptotic convexity of the confidence interval CR_α is shown in the plot. The solid curve is the value of $\ell(\beta)$ against β with sample size $n = 60$ and the dashed line is for sample size $n = 300$. The horizontal line is the cut-off line at 3.84.

Table 1
Comparison of coverage probability (CP) and interval length between the empirical likelihood CR_α and the normal approximation confidence interval NCR_α. Defining $LD = L_{norm} - L_{el}$, where L_{norm} is the length of NCR_α and L_{el} is the length of CR_α.

α	Model	Sample size	Censoring proportion	NCR_α, CP	CR_α, CP	Mean of LD	Std. of LD	Proportion of LD $>$ 0
5	I	60	15	0.930	0.947	−0.004	0.378	0.57
			25	0.920	0.946	−0.096	0.490	0.48
			40	0.864	0.948	−0.060	0.240	0.35
	I	100	15	0.926	0.938	0.001	0.227	0.52
			25	0.932	0.956	−0.017	0.225	0.50
			40	0.889	0.963	−0.072	0.212	0.36
10	I	60	15	0.869	0.894	0.080	0.270	0.68
			25	0.851	0.897	−0.047	0.334	0.54
			40	0.825	0.908	−0.053	0.379	0.46
	I	100	15	0.843	0.861	0.050	0.225	0.70
			25	0.852	0.890	−0.015	0.236	0.55
			40	0.845	0.927	−0.027	0.189	0.43

Table 2
Empirical power and size for testing $H_0 : (\beta_{01}, \beta_{02}) = (1, 1)$. The tests were based on the empirical likelihood ratio statistic proposed in Section 3.1 and the normal approximation method (given in parenthesis).

α (%)	Censoring proportion (%)	Models for (β_{01}, β_{02})			
		(1, 1)	(0, 1)	(1, 0.5)	(0, 0)
5	15	5.6 (3.6)	76.1 (61.4)	73.7 (68.1)	100.0 (100.0)
	25	3.9 (5.8)	80.1 (71.8)	82.9 (85.0)	100.0 (100.0)
	40	6.7 (14.3)	88.4 (84.8)	40.3 (42.3)	100.0 (100.0)
10	15	14.3 (8.0)	84.2 (76.6)	83.3 (78.7)	100.0 (100.0)
	25	10.3 (8.4)	88.9 (83.2)	92.5 (90.9)	100.0 (100.0)
	40	13.2 (18.3)	95.1 (92.6)	96.7 (97.9)	100.0 (100.0)

Secondly, to illustrate Theorem 3 and compare CR_α with the normal approximation method, we simulated data from the following model:

Model II: $T_i = Z_{i1}\beta_{01} + Z_{i2}\beta_{02} + e_i$, where Z_{i1} and Z_{i2} are two i.i.d. random variables from $U(0, 1)$ and $Exponential(1)$ respectively, Z_{i1} and Z_{i2} are independent, and $e_i \sim N(0, 1)$. The censoring variable $C_i \sim \log U_i$, where $U_i \sim U[0, c]$ and $Y_i = \min(T_i, C_i)$, where c is also used to adjust censoring proportion as in Model I.

In Table 2, the regression parameter $\beta_0 = (\beta_{01}, \beta_{02})$ in the model II was chosen from one of the following four cases: (1, 1), (0, 1), (1, 0.5) and (0, 0). We applied the empirical likelihood method based on the new estimating equation and the normal approximation method to test the null hypothesis $H_0 : \beta_0 = (1, 1)$.

Each entry in Table 2 is based on 1000 simulations. The numbers in Table 2 are the empirical powers or sizes (in %) of tests applying the empirical likelihood method proposed in Section 3.1 and the normal approximation inference method of Zhou and Wang [8] (given in parentheses). From the results we see that both tests give the appropriate type I errors when the null hypotheses are true. The power of empirical likelihood test was usually bigger than that of the normal approximation

Table 3
Comparison of coverage probability of three types of empirical likelihood confidence regions. R_α stands for the empirical likelihood confidence region for median regression with random censoring suggested by Qin and Tsao (we quote the column for comparison), CR_α^E is the empirical likelihood proposed in Section 3.2 and CR_α is the empirical likelihood proposed in Section 3.1.

Nominal levels	Models	Censoring proportion (%)	R_α	CR_α^E	CR_α
0.90	A	15	0.87	0.896	0.894
		25	0.90	0.894	0.896
		40	0.84	0.905	0.908
	B	15	0.87	0.891	0.903
		25	0.91	0.893	0.895
		40	0.75	0.901	0.895
	C	15	0.89	0.743	0.884
		25	0.88	0.666	0.890
		40	0.85	0.598	0.885
	D	15	0.90	0.774	0.894
		25	0.91	0.704	0.896
		40	0.85	0.570	0.905
0.95	A	15	0.93	0.950	0.950
		25	0.95	0.949	0.948
		40	0.90	0.954	0.962
	B	15	0.94	0.945	0.950
		25	0.95	0.939	0.941
		40	0.82	0.943	0.940
	C	15	0.95	0.836	0.936
		25	0.93	0.784	0.947
		40	0.91	0.725	0.957
	D	15	0.95	0.847	0.933
		25	0.94	0.774	0.941
		40	0.90	0.686	0.938

method. When β was far away from β_0, for example in the last column of Table 2, we observe that both methods had an empirical power 1.

Finally, for comparisons among R_α, CR_α^E and CR_α, we assume the following four simple linear models where the true values of intercept and slope parameters are 0 and 1, respectively and let $\mathbf{Z}_i = (1, Z_{i2})^\tau$. The confidence region for $\beta_0 = (0, 1)'$ is of interest here.

Model A: The covariates Z_{i2} are i.i.d random variables from uniform distribution $U[0, 1]$, and e_i are random variables with standard normal distribution. The censoring time is given by $C_i = \log U_i$, where $U_i \sim U[0, c]$.
Model B: The same as Model A except that e_i are generated from a normal distribution with mean 0 and variance Z_{i2}.
Model C: For this model, we set $Z_{i2} = U_i$, $T_i = Z_{i2} + 0.5e_i$, $C_i = c + Z_{i2} + 0.5\eta_i$. where U_i are i.i.d. $U[0, 1]$, and e_i's and η_i's are i.i.d. standard normal, which are independent.
Model D: The same model as Model C except that the covariates are now set to $Z_{i2} = i/n$, $i = 1, \ldots, n$.

As in models I and II, we set c to achieve 15%, 25% and 40% censoring proportions, C_p. In models A and B we use $E(1 - \Phi(c)) = C_p$ to determine c, and in models C and D, $\Phi(1 - C_p)$ was used as c, where Φ is the distribution function of the standard normal distribution. The sample size was set to be $n = 60$. We note that models A and B satisfy our model assumption, but models C and D do not, because C_i are no longer independent of $\mathbf{Z}_i$. Under the model structure C, C_i has marginal distribution $G(u) = \int_{-\infty}^{u}(\Phi(2y-2c) - \Phi(2y-2c-2))dy$ and $P(C_i > \beta_0'\mathbf{Z}_i|\mathbf{Z}_i)$ is $1-\Phi(-2c)$. Then the conditional probability $P(C_i > \beta_0'\mathbf{Z}_i|\mathbf{Z}_i) \neq 1 - G(\beta_0'\mathbf{Z}_i)$. Thus, we would expect that the empirical likelihood confidence region CR_α^E is not valid in models C and D.

Table 3 summarizes the empirical coverage accuracy of R_α, CR_α^E and CR_α based on 1000 simulations. The empirical coverage of R_α, where C_i were assumed to be only observed when $\delta_i = 0$, was quoted from Qin and Tsao [14] for comparison. From Table 3, we can get the following observations: firstly, CR_α^E and CR_α both had good coverage probabilities in models A and B. As we would expect, CR_α^E and CR_α obtain more information about censoring variables C_i such that we should have better coverage accuracy than R_α. Indeed, when the censoring proportion attained 40%, we find that CR_α^E and CR_α had better coverage probabilities in models A and B. Secondly, two proposed empirical likelihood confidence regions CR_α^E and CR_α had almost the same coverage accuracy in models A and B. Thirdly, we notice that the coverage probabilities of CR_α^E in models C and D were consistently lower than the nominal levels $1 - \alpha$ as we would expect from the previous analysis. However, we did not see much effect of the dependence of C_i and $\mathbf{Z}_i$ on the coverage probabilities of CR_α. These phenomena revealed that the CR_α^E was sensitive to the dependence between C_i and $\mathbf{Z}_i$ and CR_α was more robust in this case.

5. Summary

In this paper, we propose two empirical likelihood methods to make inference for the parameters in a median regression model with designed censoring variables. The empirical likelihood proposed based on the new estimating equation is much simple and easy to implement than the empirical likelihood based on the existing estimating equation and the normal

approximation method. The limiting distribution of the two empirical log likelihood ratio statistics evaluated at the true parameter were derived for constructing confidence regions. It was shown that one of the limiting distribution is a standard chi-square distribution and the other one is a weighted sum of chi-square distributions.

We note that, the difference between the designed censoring data we assumed in this paper and the random censoring data is that we may obtain more information about the censoring variables. As we can see from the simulation studies, the proposed empirical likelihood methods make use of the additional information about censoring variable to make more accurate the confidence region of the parameters in median regression than that from random censoring data. The proposed empirical likelihood based on the existing estimating equation is not robust against the departure from the independence of the censoring variable C_i's and covariate $\mathbf{Z}_i$'s. However, the empirical likelihood based on the new estimating equation performed well.

Acknowledgments

The research of Hengjian Cui was partially supported by the Natural Science Foundation of China (No: 10771017) and by the Key Project of MOE, PRC (No: 309007). We would like to thank the referees for their valuable comments and suggestions which led to great improvements of the presentation of the paper.

Appendix. Proofs of the main results

In this section, we give the assumptions and the proofs of the results given in Section 3.

A1. The e_i $(i = 1, \ldots, n)$ are i.i.d. random variables with $H(0) = 1/2$, where $H(t)$ is the distribution function of e_i with a symmetric continuous density $h(t)$ satisfying $h(0) > 0$.

A2. The C_i are design censoring variables, which are generated from a distribution $G(t)$ and $G(t) < 1$ for any bounded t. C_i are assumed to be independent of $\mathbf{Z}_i$ and conditionally independent of T_i given covariates $\mathbf{Z}_i$.

A3. The matrix $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E\{\mathbf{Z}_i\mathbf{Z}_i'\}$ exists and is positive definite. In addition, $\lim_{M\to\infty} \sup_i E\{\|\mathbf{Z}_i\|^2 I(\|\mathbf{Z}_i\| \geq M)\} \to 0$.

A4. There exists a distribution function $F(z)$ such that $\sup_z |F_n(z) - F(z)| \to 0$ and $\int zz'\mathrm{d}F_n(z) \to \int zz'\mathrm{d}F(z) < +\infty$, where $F_n(z) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{Z}_i \leq z)$.

Remark 1. The condition (A1) was used in Wang and Zhou [8]. The independence between C_i and $\mathbf{Z}_i$ in condition (A2) is not necessary for the empirical likelihood proposed in Section 3.1, but it is necessary for the empirical likelihood proposed in Section 3.2. (A3) is a similar condition with that in Ying et al. [18]. (A4) assures the existence of the limits of $\int \varphi(z)zz'\mathrm{d}F_n(z)$ for any bounded function $\varphi(z)$.

Lemma 1. *Let Ω be a bounded parameter space which contains β_0 as an interior point. Assume that* (A1)–(A4) *hold, then $E(\bar{g}(\beta)) = 0$ holds for $\beta \in \Omega$ if and only if $\beta = \beta_0$.*

Proof. We are going to show that if $\beta = \beta_0$, then $E(g_i(\beta)) = 0$. In fact, we have

$$\begin{aligned} g_i(\beta_0) &= \mathrm{sgn}(\min(C_i, T_i) - \min(C_i, \beta_0'\mathbf{Z}_i))(1 + \mathrm{sgn}(C_i - \beta_0'\mathbf{Z}_i))\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i) \\ &\quad + \mathrm{sgn}(\min(C_i, T_i) - \min(C_i, \beta_0'\mathbf{Z}_i))(1 + \mathrm{sgn}(C_i - \beta_0'\mathbf{Z}_i))\mathbf{Z}_i I(C_i = \beta_0'\mathbf{Z}_i) \\ &\quad + \mathrm{sgn}(\min(C_i, T_i) - \min(C_i, \beta_0'\mathbf{Z}_i))(1 + \mathrm{sgn}(C_i - \beta_0'\mathbf{Z}_i))\mathbf{Z}_i I(C_i < \beta_0'\mathbf{Z}_i). \end{aligned}$$

Therefore, taking expectation on both side yields

$$\begin{aligned} E(g_i(\beta_0)) &= 2E(\mathrm{sgn}(\min(C_i, T_i) - \min(C_i, \beta_0'\mathbf{Z}_i))\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i)) + E(\mathrm{sgn}(\min(C_i, T_i) - \min(C_i, \beta_0'\mathbf{Z}_i))\mathbf{Z}_i I(C_i = \beta_0'\mathbf{Z}_i)) \\ &= -2E(\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i) I(T_i < \beta_0'\mathbf{Z}_i)) \\ &\quad + 2E(\mathrm{sgn}(\min(C_i, T_i) - \beta_0'\mathbf{Z}_i)\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i) I(T_i > \beta_0'\mathbf{Z}_i) I(C_i > T_i)) \\ &\quad + 2E(\mathrm{sgn}(\min(C_i, T_i) - \beta_0'\mathbf{Z}_i)\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i) I(T_i > \beta_0'\mathbf{Z}_i) I(C_i \leq T_i)) - E(\mathbf{Z}_i I(C_i = \beta_0'\mathbf{Z}_i) I(T_i < \beta_0'\mathbf{Z}_i)) \\ &= -2E(\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i) I(T_i < \beta_0'\mathbf{Z}_i)) + 2E(\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i) I(T_i > \beta_0'\mathbf{Z}_i)) - E(\mathbf{Z}_i I(C_i = \beta_0'\mathbf{Z}_i) I(T_i < \beta_0'\mathbf{Z}_i)). \end{aligned}$$

Since C_i's and T_i's are independent given $\mathbf{Z}_i$, it follows that

$$\begin{aligned} E(Z_i I(C_i > \beta_0'\mathbf{Z}_i) I(T_i < \beta_0'\mathbf{Z}_i)) &= E\left\{E(\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i) I(T_i < \beta_0'\mathbf{Z}_i)|\mathbf{Z}_i)\right\} \\ &= E\left\{\mathbf{Z}_i P(C_i > \beta_0'\mathbf{Z}_i|\mathbf{Z}_i) P(T_i < \beta_0'\mathbf{Z}_i|\mathbf{Z}_i)\right\}. \end{aligned}$$

From the assumption that e_i's have continuous density functions, which have median 0, we know that $P(T_i < \beta_0'\mathbf{Z}_i|\mathbf{Z}_i) = P(T_i > \beta_0'\mathbf{Z}_i|\mathbf{Z}_i)$ and $P(C_i = \beta_0'\mathbf{Z}_i) = 0$ which implies that $E(g_i(\beta_0)) = 0$.

On the other hand, we know that $E(g_i(\beta)) = 0$ if and only if

$$E\{\mathbf{Z}_i I(C_i > \beta'\mathbf{Z}_i)(I(T_i > \beta'\mathbf{Z}_i) - I(T_i < \beta'\mathbf{Z}_i))\} = 0. \tag{7}$$

This is equivalent to

$$E\{(\beta - \beta_0)'\mathbf{Z}_i(1 - G(\beta'\mathbf{Z}_i))(1 - 2H((\beta - \beta_0)'\mathbf{Z}_i))\} = 0.$$

Since $H(0) = 1/2$, we have

$$(\beta - \beta_0)' E\{\mathbf{Z}_i(1 - G(\beta'\mathbf{Z}_i))(H(0) - H((\beta - \beta_0)'\mathbf{Z}_i))\} = 0.$$

Now suppose $\beta \neq \beta_0$ and $\beta \in \Omega$. By condition (A1), then $|H(t)-H(0)| \geq b|t|/(1+|t|)$ and $t[H(t)-H(0)] > bt^2/(1+|t|)$ with some $b > 0$. For M_0 large enough, we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\beta - \beta_0)' E\{\mathbf{Z}_i(1 - G(\beta'\mathbf{Z}_i))(H((\beta - \beta_0)'\mathbf{Z}_i) - H(0))\} \\
&\quad \geq b(\beta - \beta_0)' \frac{1}{n}\sum_{i=1}^{n} E\left\{\mathbf{Z}_i\mathbf{Z}_i' \frac{1 - G(\beta'\mathbf{Z}_i)}{1 + |(\beta - \beta_0)'\mathbf{Z}_i|}\right\}(\beta - \beta_0) \\
&\quad \geq b(\beta - \beta_0)' \frac{1}{n}\sum_{i=1}^{n} E\left\{\mathbf{Z}_i\mathbf{Z}_i' I(\|\mathbf{Z}_i\| < M_0) \frac{1 - G(\beta'\mathbf{Z}_i)}{1 + |(\beta - \beta_0)'\mathbf{Z}_i|}\right\}(\beta - \beta_0) \\
&\quad \geq b(\beta - \beta_0)' \frac{1}{n}\sum_{i=1}^{n} E\{\mathbf{Z}_i\mathbf{Z}_i' I(\|\mathbf{Z}_i\| < M_0)\}(\beta - \beta_0) \frac{1 - G(\|\beta\| M_0)}{1 + \|\beta - \beta_0\| M_0}. \qquad (8)
\end{aligned}$$

Because of the positive definite of $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E\{\mathbf{Z}_i\mathbf{Z}_i'\}$ from condition (A3), for large n and M_0, $\frac{1}{n}\sum_{i=1}^{n} E\{\mathbf{Z}_i\mathbf{Z}_i' I(\|\mathbf{Z}_i\| < M_0)\}$ is positive definite. Condition (A2) and $\|\beta\| < \infty$ assure that $\frac{1-G(\|\beta\| M_0)}{1+\|\beta-\beta_0\| M_0}$ is positive. It follows that (8) is positive. This is a contradiction to (7), thus we conclude that $E(\bar{g}(\beta)) = 0$ implies $\beta = \beta_0$. Hence Lemma 1 is proved. □

Proof of Theorem 1. We firstly note that

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{g_i(\beta_0)}{1 + \lambda' g_i(\beta_0)} \qquad (9)$$

and by central limit theorem, if Z_i's are i.i.d random variables,

$$\frac{1}{\sqrt{n}} V_{\beta_0}^{-\frac{1}{2}} \sum_{i=1}^{n} g_i(\beta_0) \xrightarrow{d} N(0, I_p), \qquad (10)$$

where $V_{\beta_0} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(g_i(\beta_0)) = \frac{4}{n}\sum_{i=1}^{n} E\{\mathbf{Z}_i\mathbf{Z}_i'(1-G(\beta_0'\mathbf{Z}_i))\}$ and it is positive from assumption (A3). When $\mathbf{Z}_i$'s are fixed design points, $g_i(\beta_0)$'s are independent but not identically distributed. Form condition (A4), the Lindberg condition holds in this case, then the CLT also holds for $g_i(\beta_0)$. Then, we have

$$\frac{1}{n}\sum_{i=1}^{n} g_i(\beta_0) = O_p(n^{-\frac{1}{2}})$$

and by using the same approach as Owen [9], we get $\|\lambda\| = O_p(n^{-\frac{1}{2}})$.

Employing the Taylor expansion, we have

$$\ell(\beta_0) = 2\sum_{i=1}^{n} \log\{1 + \lambda' g_i(\beta_0)\} = 2\sum_{i=1}^{n}\left[\lambda' g_i(\beta_0) - \frac{1}{2}(\lambda' g_i(\beta_0))^2\right] + \eta_n, \qquad (11)$$

with

$$|\eta_n| \leq C\sum_{i=1}^{n} (\lambda' g_i(\beta_0))^3.$$

From assumption (A4), with probability 1,

$$\max_{1\leq i\leq n} \|g_i(\beta_0)\| \leq 2 \max_{1\leq i\leq n} \|\mathbf{Z}_i\| = o_p(\sqrt{n})$$

and it follows that $|\eta_n| \leq Cn\|\lambda\|^3 \max_{1\leq i\leq n} \|g_i(\beta_0)\| = o_p(1)$.

From (9), we know that

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{g_i(\beta_0)}{1 + \lambda' g_i(\beta_0)} = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta_0) - \frac{1}{n}\left(\sum_{i=1}^{n} g_i(\beta_0) g_i'(\beta_0)\right)\lambda + \frac{1}{n}\sum_{i=1}^{n} \frac{g_i(\beta_0)(\lambda' g_i(\beta_0))^2}{1 + \lambda' g_i(\beta_0)}. \qquad (12)$$

The final term in (12) is bounded by $\frac{1}{n}\sum_{i=1}^{n}\|g_i(\beta_0)\|^3\|\lambda\|^2|1+\lambda' g_i(\beta_0)|^{-1} = o_p(n^{-\frac{1}{2}})$, where we use $\frac{1}{n}\sum_{i=1}^{n}\|g_i(\beta_0)\|^3 = o(n^{\frac{1}{2}})$, which can be derived from $E\|g_i(\beta_0)\|^2 < \infty$. This is right using assumption (A4). Then

$$\lambda = \left(\sum_{i=1}^{n} g_i(\beta_0)g_i'(\beta_0)\right)^{-1}\sum_{i=1}^{n} g_i(\beta_0) + o_p(1).$$

From (12), we have $\sum_{i=1}^{n}\lambda' g_i(\beta_0) = \sum_{i=1}^{n}(\lambda' g_i(\beta_0))^2 + o_p(1)$. Combining it with (11) and the above expression for λ, we get

$$\begin{aligned}\ell(\beta_0) &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i'(\beta_0)\right)\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\beta_0)g_i'(\beta_0)\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\beta_0)\right) + o_p(1)\\ &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i'(\beta_0)\right)V_{\beta_0}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\beta_0)\right) + o_p(1).\end{aligned}$$

Using (10), we have

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i'(\beta_0)\right)V_{\beta_0}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\beta_0)\right) \to \chi_p^2.$$

Therefore, $\ell(\beta_0) \xrightarrow{d} \chi_p^2$. Thus, Theorem 1 is proved. □

Proof of Theorem 2. From the proof of Lemma 1,

$$\begin{aligned}E(g_i(\beta)) &= -2E(\mathbf{Z}_i I(C_i > \beta'\mathbf{Z}_i)I(T_i < \beta'\mathbf{Z}_i)) + 2E(\mathbf{Z}_i I(C_i > \beta'\mathbf{Z}_i)I(T_i > \beta'\mathbf{Z}_i))\\ &= -2E\left[\mathbf{Z}_i P(C_i > \beta'\mathbf{Z}_i|\mathbf{Z}_i)\{P(T_i < \beta'\mathbf{Z}_i|\mathbf{Z}_i) - P(T_i > \beta'\mathbf{Z}_i|\mathbf{Z}_i)\}\right]\\ &= -2E\left[\mathbf{Z}_i\{1-G(\beta'\mathbf{Z}_i)\}\{2H((\beta-\beta_0)'\mathbf{Z}_i) - 1\}\right].\end{aligned}$$

Because e_i's have density functions $h(t)$ and $H(0) = 1/2$. A Taylor expansion may lead to

$$\begin{aligned}E(g_i(\beta)) &= -2E[\mathbf{Z}_i(1-G(\beta_0'\mathbf{Z}_i))(2H(0)-1)] - 4E[\mathbf{Z}_i(1-G(\beta_0'\mathbf{Z}_i))h(0)\mathbf{Z}_i'(\beta-\beta_0)] + O\{(\beta-\beta_0)^2\}\\ &= -4E[\mathbf{Z}_i(1-G(\beta_0'\mathbf{Z}_i))h(0)\mathbf{Z}_i'(\beta-\beta_0)] + O\{(\beta-\beta_0)^2\}.\end{aligned}$$

And we know that

$$\begin{aligned}\mathrm{Var}(g_i(\beta)) &= E[(1+\mathrm{sign}(C_i-\beta'\mathbf{Z}_i))^2\mathbf{Z}_i'\mathbf{Z}_i] - E\{g_i(\beta)\}E\{g_i'(\beta)\}\\ &= 4E[\mathbf{Z}_i\mathbf{Z}_i'P(C_i > \beta'\mathbf{Z}_i|\mathbf{Z}_i)] - 4E\{\mathbf{Z}_i[1-G(\beta'\mathbf{Z}_i)][2H((\beta-\beta_0)'\mathbf{Z}_i) - 1]\}^{\otimes 2}\end{aligned}$$

where $A^{\otimes 2} = AA'$. Plugging in β_0, we get

$$V_{\beta_0} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(g_i(\beta_0)) = \frac{4}{n}\sum_{i=1}^{n}E\{\mathbf{Z}_i\mathbf{Z}_i'(1-G(\beta_0'\mathbf{Z}_i))\}.$$

From the proof of Theorem 1, we know that

$$\ell(\beta) = nE\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\beta)\right)' V_{\beta_0}^{-1}E\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\beta)\right) + o_p(1).$$

For $\beta = \beta_0 + O(n^{-1/2})$, it follows that

$$\ell(\beta) = nh(0)^2(\beta-\beta_0)'V_{\beta_0}(\beta-\beta_0) + o_p(1).$$

Following from the quadratic form of the leading term of $\ell(\beta)$, we conclude that CR_α is asymptotic convex. □

Proof of Theorem 3. From the standard empirical likelihood theory, we have $\ell(\tilde{\beta}) \to \infty$ a.s. as $n \to \infty$. Therefore, $P\{\tilde{\beta} \in CR_\alpha\} = P\{\ell(\tilde{\beta}) < c_\alpha\} \to 0$.

For notation simplification, we denote $c = h(0)^{-1}V_{\beta_0}^{-1/2}$. From the proof of Lemma 1, without loss of generality, assume $c'\mathbf{Z}_i > 0$ a.s., then we know that $\bar{g}(\tilde{\beta}_n) = \bar{g}(\beta_0) + R_n$, where $\bar{g}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} g_i(\cdot)$ and

$$R_n = \frac{1}{n}\sum_{i=1}^{n}\left\{-4\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i)I\left(\left(\beta_0 + \frac{1}{\sqrt{n}}c\right)'\mathbf{Z}_i > T_i > \beta_0'\mathbf{Z}_i\right)\right. \tag{13}$$

$$+2\mathbf{Z}_i I\left(\left(\beta_0+\frac{1}{\sqrt{n}}c\right)'\mathbf{Z}_i > C_i > \beta_0'\mathbf{Z}_i\right) I(T_i < \beta_0'\mathbf{Z}_i) \tag{14}$$

$$-2\mathbf{Z}_i I\left(\left(\beta_0+\frac{1}{\sqrt{n}}c\right)'\mathbf{Z}_i > C_i > \beta_0'\mathbf{Z}_i\right) I(T_i > \beta_0'\mathbf{Z}_i) \tag{15}$$

$$+4\mathbf{Z}_i I\left(\left(\beta_0+\frac{1}{\sqrt{n}}c\right)'\mathbf{Z}_i > C_i > \beta_0'\mathbf{Z}_i\right) I\left(\left(\beta_0+\frac{1}{\sqrt{n}}c\right)'\mathbf{Z}_i > T_i > \beta_0'\mathbf{Z}_i\right)\Bigg\}. \tag{16}$$

Now we show that (13)–(15) are $O_p(1/\sqrt{n})$, and (16) is $O_p(1/n)$. Since the proofs are similar, we only show that (13) is $O_p(1/\sqrt{n})$. In fact, if Z_{ij} is the jth component of vector Z_i, then

$$\begin{aligned}&\mathrm{Var}\left\{\mathbf{Z}_{ij}I(C_i > \beta_0'\mathbf{Z}_i)I\left(\left(\beta_0+\frac{1}{\sqrt{n}}c\right)'\mathbf{Z}_i > T_i > \beta_0'\mathbf{Z}_i\right)\right\}\\ &\quad=\frac{1}{\sqrt{n}}E\left\{\mathbf{Z}_{ij}^2(1-G(\beta_0'\mathbf{Z}_i))h(0)c'\mathbf{Z}_i\right\}+o\left(\frac{1}{\sqrt{n}}\right)=O\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

and

$$\begin{aligned}&E\left\{\mathbf{Z}_i I(C_i > \beta_0'\mathbf{Z}_i)I\left(\left(\beta_0+\frac{1}{\sqrt{n}}c\right)'\mathbf{Z}_i > T_i > \beta_0'\mathbf{Z}_i\right)\right\}\\ &\quad=\frac{1}{\sqrt{n}}E(\mathbf{Z}_i(1-G(\beta_0'\mathbf{Z}_i))h(0)c'\mathbf{Z}_i)+o\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Therefore (13) is $O_p(\frac{1}{\sqrt{n}})$. It follows that

$$\bar{g}(\tilde{\beta}_n)=\bar{g}(\beta_0)-\frac{1}{\sqrt{n}}V_{\beta_0}^{1/2}\gamma+o_p\left(\frac{1}{\sqrt{n}}\right),$$

where V_{β_0} is defined in Theorem 3. From the proof of Theorem 1, we know that

$$\begin{aligned}\ell(\tilde{\beta}_n) &= n\bar{g}'(\tilde{\beta}_n)V_{\beta_0}^{-1}\bar{g}(\tilde{\beta}_n)+o_p(1)\\ &= n\left[\bar{g}(\beta_0)-\frac{1}{\sqrt{n}}V_{\beta_0}^{1/2}\gamma\right]'V_{\beta_0}^{-1}\left[\bar{g}(\beta_0)-\frac{1}{\sqrt{n}}V_{\beta_0}^{1/2}\gamma\right]+o_p(1)\\ &\xrightarrow{d}\chi_p^2(\|\gamma\|^2).\end{aligned}$$

This completes the proof of Theorem 3. □

In order to prove Theorem 4, we need to show the following Lemmas.

Lemma 2. *Under conditions* (A1)–(A4), *we have*

$$\sqrt{n}W_n(\beta_0)\xrightarrow{d}N(0,\,\Gamma_1-\Gamma_2)$$

where Γ_1, Γ_2 *are defined in Section 3.2.*

Proof. From the definition of $W_n(\beta_0)$, we know that

$$\begin{aligned}W_n(\beta_0) &= \frac{1}{n}\sum_{i=1}^n\left[I(Y_i\geq\beta_0'\mathbf{Z}_i)-\frac{1}{2}(1-G(\beta_0'\mathbf{Z}_i))\right]\mathbf{Z}_i-\frac{1}{2n}\sum_{i=1}^n(G(\beta_0'\mathbf{Z}_i)-\hat{G}_n(\beta_0'\mathbf{Z}_i))\mathbf{Z}_i\\ &\equiv P_1-P_2.\end{aligned}$$

Thus, it is easy to see that $\sqrt{n}P_1$ is normal distributed with mean 0 and variance

$$\Gamma_1\equiv\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E\left\{\left[I(Y_i\geq\beta_0'\mathbf{Z}_i)-\frac{1}{2}(1-G(\beta_0'\mathbf{Z}_i))\right]^2\mathbf{Z}_i\mathbf{Z}_i'\right\}.$$

For the second part, we know that P_2 is a two sample U-statistics with kernel

$$\phi(C_j;\mathbf{Z}_i)=(G(\beta_0'\mathbf{Z}_i)-\hat{G}_n(\beta_0'\mathbf{Z}_i))\mathbf{Z}_i.$$

Using the standard U-statistics theory (Randles and Wolfe [20]), one can show that the distribution of $\sqrt{n}P_2$ is asymptotically normal with mean 0 and variance $\frac{1}{4n^2}\sum_{i\neq j}^{n}[E(\phi(C_1;\mathbf{Z}_i)\phi'(C_1;\mathbf{Z}_j))+E(\phi(C_i;\mathbf{Z}_1)\phi'(C_j;\mathbf{Z}_1))]$. We see that

$$E\{\phi(C_i;\mathbf{Z}_1)\phi(C_j;\mathbf{Z}_1)\}=E\{E(\phi(C_1;\mathbf{Z}_1)\phi(C_2;\mathbf{Z}_1)|\mathbf{Z}_1)\}=0$$

and from a simple calculation, we have

$$E\{\phi(C_1;\mathbf{Z}_i)\phi'(C_1;\mathbf{Z}_j)\}=E\left\{\min(B_{ji}(\beta_0),B_{ij}(\beta_0))\mathbf{Z}_i\mathbf{Z}_j'\right\},$$

where $B_{ij}(\beta_0)=G(\beta_0'\mathbf{Z}_i)(1-G(\beta_0'\mathbf{Z}_j))$. Moreover, the covariance between $\sqrt{n}P_1$ and $\sqrt{n}P_2$ is

$$\begin{aligned}\mathrm{Cov}(\sqrt{n}P_1,\sqrt{n}P_2)&=\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left\{\left[\frac{1}{2}(1-G(\beta_0'\mathbf{Z}_i))-I_{\{Y_i\geq\beta_0'\mathbf{Z}_i\}}\right]\hat{G}_n(\beta_0'\mathbf{Z}_j)\mathbf{Z}_i\mathbf{Z}_j'\right\}\\&=\frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left\{G(\beta_0'\mathbf{Z}_i)(1-G(\beta_0'\mathbf{Z}_j))\mathbf{Z}_i\mathbf{Z}_j'\right\}\end{aligned}$$

and then $\mathrm{Cov}(\sqrt{n}P_2,\sqrt{n}P_1)=\frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left\{G(\beta_0'\mathbf{Z}_i)(1-G(\beta_0'\mathbf{Z}_j))\mathbf{Z}_j\mathbf{Z}_i'\right\}$.

It follows that $E(\sqrt{n}W_n(\beta_0))^2=\Gamma_1-\Gamma_2$, where Γ_1 and Γ_2 are defined in Theorem 4. Therefore, applying the center limit theorem, the proof of this lemma is completed. □

Lemma 3. *Under conditions* (A1)–(A4)*, we have*

$$\frac{1}{n}\sum_{i=1}^{n}W_{ni}(\beta_0)W_{ni}(\beta_0)'\xrightarrow{p}\Gamma_1,$$

where Γ_1 is defined in Section 3.2.

By virtue of the Glivenko–Cantelli Theorem, Lemma 2 and the Law of Large Numbers, the proof of Lemma 3 follows immediately.

Proof of Theorem 4. Applying Lemmas 2 and 3, we can use the same method as we prove Theorem 1. □

References

[1] M.J. Lee, Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models, Springer, 1996.
[2] J. Heckman, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, Ann. Econ. Soc. Meas. 5 (1976) 475–492.
[3] J. Heckman, Sample selection bias as a specification error, Econometrica 47 (1979) 153–161.
[4] J.L. Powell, Least absolute deviations estimation for censored regression model, J. Econometrics 25 (1984) 303–325.
[5] C.R. Rao, L.C. Zhao, Asymptotic normality of LAD estimator in censored regression models, Math. Methods Statist. 2 (1993) 228–239.
[6] I.W. McKeague, S. Subramanian, Y. Sun, Median regression and the missing information principle, J. Nonparam. Statist. 13 (2001) 709–727.
[7] S. Subramanian, Median regression using nonparametric kernel estimation, J. Nonparam. Statist. 14 (2002) 583–605.
[8] X.Q. Zhou, J.D. Wang, LAD estimation for nonlinear regression models with randomly censored data, Sci. China Ser. A 48 (2005) 880–897.
[9] A. Owen, Empirical likelihood ratio confidence regions, Ann. Statist. 18 (1990) 90–120.
[10] S.X. Chen, P. Hall, Smoothed empirical likelihood confidence interval for quantiles, Ann. Statist. 21 (1993) 1166–1181.
[11] H.J. Cui, S.X. Chen, Empirical likelihood confidence region for parameter in the error-in-variables models, J. Multivariate Anal. 84 (2003) 101–115.
[12] P. Hall, B. La Scala, Methodology and algorithms of empirical likelihood, Int. Statist. Rev. 58 (1990) 109–127.
[13] J. Qin, J.F. Lawless, Empirical likelihood and general estimating equations, Ann. Statist. 22 (1994) 300–325.
[14] G.S. Qin, M. Tsao, Empirical likelihood inference for median regression models for censored survival data, J. Multivariate Anal. 85 (2003) 416–430.
[15] J. Shi, T.S. Lau, Empirical likelihood for partially linear models, J. Multivariate Anal. 72 (2000) 132–148.
[16] T. DiCicco, P. Hall, J. Romano, Empirical likelihood is Barterlett-correctable, Ann. Statist. 19 (1991) 1053–1061.
[17] S.X. Chen, H.J. Cui, On the second order properties of empirical likelihood with moment restrictions, J. Econometrics 141 (2007) 492–516.
[18] Z. Ying, S.H. Jung, L.J. Wei, Survival analysis with median regression models, J. Amer. Statist. Assoc. 90 (1995) 178–184.
[19] X.Q. Zhou, J.D. Wang, A genetic method of LAD estimation for models with censored data, Comput. Statist. Data Anal. 48 (2005) 451–466.
[20] R.H. Randles, D.A. Wolfe, Introduction to the Theory of Nonparametric Statistics, John Wiley & Sons, 1979.