경제학석사 학위논문

Smoothed Empirical Likelihood Methods for Censored Quantile Regression

2013년 8월

서울대학교 대학원

경제학부 경제학 전공

문 철

Smoothed Empirical Likelihood Methods for Censored Quantile Regression

Chul Moon
Department of Economics
The Graduate School
Seoul National University

Abstract

This article verifies the efficiency of the empirical likelihood method to estimate the parameters of the censored quantile regression models suggested by Whang (2003) via simulation. We smooth the simple estimating equation in a censored quantile regression model with a nonparametric kernel function for higher-order refinements. We show that the confidence region based on the smoothed empirical likelihood estimator, known to be the first-order equivalent to the standard censored quantile estimator, has coverage error of order $O\left(n^{-1}\right)$. Monte Carlo experiments suggest that the Bartlett corrected smoothed empirical likelihood method performs well in small samples, and it provides more accurate and computationally efficient results than the commonly used (smoothed) bootstrap methods. Moreover, simulation results show that the proposed confidence region has better finite sample performance than the confidence interval obtained from the un-corrected smoothed empirical likelihood estimation, which are consistent with the argument of Whang (2003) that Bartlett correction can reduce the coverage error of smoothed empirical likelihood confidence region to order $O\left(n^{-2}\right)$.

Keywords: Empirical Likelihood, Censored Quantile Regression, Smoothing, Bartlett Correction
Student Number: 2011-20172

Contents

1 Introduction

The quantile regression model was first introduced by Koenker and Basset (1978, 1982) and has been widely used in econometrics and survival analysis. When the dependent variable is subject to censoring, modelling the conditional quantile instead of the conditional mean offers advantages. The conditional quantile is often identifiable under weaker distributional assumptions whereas additional distributional assumptions are required to identify the conditional mean. (see Portnoy, 2003)

There is a big literature in the estimation of parameters in the censored quantile regression (CQR) model. Powell (1984, 1986) focused on a fixed censoring, where censoring values C_i are observable. Fixed censoring often occurs in social surveys where data are up to ceiling effect. For random censoring cases, which are common in many survival analysis, Portnoy (2003), Peng and Huang (2008), and Wang and Wang (2009), Huang (2010) developed estimating methods. This paper focuses on the fixed censoring case. Under the regularity conditions, the Powell's CQR estimator is $\sqrt{n}$-consistent to the true parameter and has an asymptotic normal distribution. However, the first-order approximation might provide inaccurate results in many applications.

The CQR estimation poses the computational complexity that is partly due to the non-convex and piecewise linear distance function which includes multiple local minima. The optimization algorithm for the standard CQR estimation has been suggested by many researchers: A modified reduced-gradient algorithm by Womersley (1986), an interior point approach by Koenker and Park (1994), an emulation algorithm (EA) by Pinkse (1993), and a three-step algorithm by Chernozhukov and Hong (2002). EA only can converge to global minima by checking every critical point, while the other algorithms used to converge to local minima. However, EA requires heavy computational complexity. We use the algorithm of Stengos and Wang (2007) that has less computational load than the other methods.

The reason that the estimation of variance for the CQR estimation is known to be difficult stems from the unsmoothness of the estimating function. Chen and Hall (1993), Horowitz (1998) and Whang (2006) used the kernel smoothing methods to approximate the estimating function. Under the kernel smoothing, with the condition of the bandwidth parameter converging to zero with a proper rate, the smoothed estimating function is known to be asymptotically equivalent to the estimating equation. In the CQR estimation, the variance-covariance matrix depends on $f(0|X)$, unknown conditional error density function. Buchinsky (1995) and Hahn (1995) applied the bootstrap methods to construct confidence intervals for quantile regressions. However, the standard bootstrap method cannot be directly applied to obtain higher-order refinements of the confidence region, because the Edgeworth expansion usually cannot be applied to unsmooth functions. Horowitz (1998) considers a median regression model and demonstrates that the smoothed censored least absolute deviation estimator is asymptotically equivalent to the standard CQR estimator. He shows that the bootstrap method could achieve asymptotic refinements on the smoothed censored quantile regression (SCQR) estimator and it is corrected to have order of $O(n^{-\gamma})$, where $\gamma < 1$ but close to 1. He suggests that results could also be applied to coverage probabilities of confidence regions.

This paper mainly focuses on an empirical likelihood method (EL) for estimating the parameters of the CQR models. It is shown that smoothed empirical likelihood (SEL) estimator is first-order asymptotically equivalent to the standard CQR estimator. Whang (2003) derives the finite sample properties that establishes the higher order properties of smoothed unconditional EL confidence regions of the CQR model. This paper investigates the finite sample properties of Whang (2003) via simulation studies. To achieve higher order development of the EL confidence region, smoothing functions are implemented to the estimating equations of CQR parameters. Chen and Hall (1993) show that smoothed confidence intervals for quantiles with no covariates are Bartlett correctable, and Whang (2006) extends it to the quantile regression model.

Simulation results indicate that the Bartlett corrected SEL estimation performs well, which supports the theory suggested by Whang (2003). The results is consistent with theories of Whang (2003) that confidence regions based on the uncorrected SEL estimator have coverage error of order $O(n^{-1})$ and Bartlett-corrected SEL confidence regions have coverage error up to order $O(n^{-2})$. Moreover, the coverage probability of SEL is not sensitive to the choice of bandwidth at the range we considered and Bartlett-corrected SEL provides accurate estimations among the other compared methods. Also, its coverage error decreases faster than the other methods as the number of observations increases.

The rest of this article is organized as follows. Section 2 and Section 3 summarize theories investigated by Whang (2003). Section 2 defines the SEL estimator in CQR models and discusses their asymptotic properties. Section 3 contains definitions of confidence regions and coverage accuracy for SEL and higher-order analysis via Bartlett correction. Section 4 explains estimating methods which are used in simulations and compares their performances. Section 5 makes concluding remarks and suggests some possible extensions.

2 Models

2.1 Estimators

Consider a fixed right censoring model,

$$Y_i = \min\left[C_i, X_i'\beta_0 + U_i\right] \quad i = 1, ..., n,$$

where a dependent variable Y_i, a $K \times 1$ vector of regressors X_i, and a censoring value C_i is observed while a $K \times 1$ parameter vector β_0, and an error U_i are unobserved. The error satisfies $P[U_i \leq 0|X_i] = q$ a.s. for $0 < q < 1$.

The CQR estimator $\hat{\beta}_{CQ}$ of β_0 solves, for all β on a parameter space B,

$$\min_{\beta} H_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_q\left(Y_i - \min\{C_i, X_i'\beta\}\right),$$

where $\rho_q(x) = [q - 1(x \leq 0)]x$ and $1(\cdot)$ is an indicator function. $\hat{\beta}_{CQ}$ satisfies the first-order condition (Powell, 1984)

$$\frac{1}{n}\sum_{i=1}^{n}\left[1(Y_i \leq X_i'\hat{\beta}_{CQ}) - q\right] 1\left(X_i'\hat{\beta}_{CQ} < C_i\right) X_i = o_p(n^{-1/2}).$$

To motivate our estimator, an unbiased estimating function g is defined as

$$\begin{aligned} &E\left[g(Y_i, X_i, \beta_0)\right] = 0, \text{ where} \\ &g(Y_i, X_i, \beta) = \left[1(Y_i \leq X_i'\beta) - q\right] 1\left(X_i'\beta < C_i\right) X_i. \end{aligned}$$

The CQR estimator $\hat{\beta}_{CQ}$ can be estimated by the empirical likelihood method. However, it is difficult to achieve high-order refinements because the estimating equation g contains an indicator function, which is not differentiable. In this paper, g is smoothed with the nonparametric kernel function K that satisfies assumptions on section 2.2. Let $G(x) = \int_{-\infty}^{x} K(u)du$ and $G_h(x) = G(x/h)$. Then, the estimating equation which is smoothed with the kernel function is

$$Z_i(\beta) = (G_h(Y_i - X_i'\beta) - q)G_h(X_i'\beta - C_i)X_i.$$

Empirical likelihood can be applied to this framework by maximizing

$$L = \prod_{i=1}^{n} dF(x_i) = \prod_{i=1}^{n} p_i,$$

subject to restrictions on $p_i = dF(x_i) = Pr(X = x_i)$

$$p_i \geq 0, \ \ \sum_{i=1}^{n} p_i = 1, \ \ \sum_{i=1}^{n} p_i Z_i(\beta) = 0.$$

The maximum value can be found via Kuhn-Tucker methods. Let

$$H = \sum_{i=1}^{n} \log p_i + \lambda \left(1 - \sum_{i=1}^{n} p_i\right) + t' \sum_{i=1}^{n} p_i Z_i(\beta) + \sum_{i=1}^{n} \mu_i p_i,$$

where λ, μ_i and $t = (t_1, t_2, ..., t_k)'$ are Kuhn-Tucker multipliers. Taking derivatives with respect to p_i, we have

$$p_i = \frac{1}{n} \cdot \frac{1}{1 + t'Z_i(\beta)}.$$

With the restrictions on p_i above,

$$0 = \sum_{i=1}^{n} p_i Z_i(\beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + t'Z_i(\beta)} Z_i(\beta).$$

Then, the (profile) empirical log-likelihood function for β_0 is now defined to be

$$L_h(\beta) = \prod_{i=1}^{n} \left\{ \frac{1}{n} \cdot \frac{1}{1 + t' Z_i(\beta)} \right\}.$$

Because $\prod_{i=1}^{n} p_i$ is maximized for $p_i = 1/n$ (see Qin and Lawless, 1994), the empirical log-likelihood ratio is

$$l_h(\beta) = \sum_{i=1}^{n} \log[1 + t(\beta)' Z_i(\beta)].$$

By definition, the SEL estimator $\hat{\beta}_E$ of β_0 minimizes $l_h(\beta)$ for $\beta \in B$.

2.2 Assumptions and additional notation

2.2.1 Assumptions

The assumptions are as follows:

Assumption 1. $\{(Y_i, X_i) : i = 1, ..., n\}$ are i.i.d random vectors.

Assumption 2. β_0 is an interior point of the compact parameter space B, which is a subset of $\mathbb{R}^K$.

Assumption 3. X_i has bounded support, $P(X_i'\beta_0 = 0) = 0$, and $E[1(X_i'b > \varepsilon)]X_i X_i'$ is positive definite for some $\varepsilon > 0$ and all b in a neighborhood of β_0.

Assumption 4.
(a) $F(0|x) = q$ for almost every x.
(b) For all u in a neighborhood of C_i and almost every x, $f(u|x)$ exists, is bounded away from zero, and is r times continuously differentiable with respect to u.

Assumption 5.
(a) $K(\cdot)$ is bounded and compactly supported on [-1, 1].

(b) For some constant $C^K \neq 0$, $K(\cdot)$ is an rth-order kernel, that is,

$$\int u^j K(u)du = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \leq j \leq r-1 \\ C^K & \text{if } j = r. \end{cases}$$

(c) Let $\tilde{G}(u) = ([G(u)], [G(u)]^2, ..., [G(u)]^{L+1})'$ for some $L \geq 1$, where $G(u) = \int_{-\infty}^{u} K(v)dv$. For any $\theta \in \mathbb{R}^{L+1}$ satisfying $\|\theta\| = 1$, there is a partition of [-1,1], $-1 = a_0 < a_1 < ... < a_{L+1} = 1$ such that $\theta'\tilde{G}(u)$ is either strictly positive or strictly negative on (a_{l-1}, a_l) for $l = 1, ..., L+1$.

Assumption 6. A smoothing parameter h satisfies
(a) $nh^{2r} \to 0$
(b) $nh/\log n \to \infty$ as $n \to \infty$.

Assumptions 1-5 are similar to assumptions that are used in Horowitz (1998); Assumption 3 is modified for the censored model. Assumptions 1-5(b) define the model and ensure that β_0 is identified. Also, based on these assumptions, asymptotic normality of β_0 and the Taylor expansion for higher-order asymptotic approximation can be derived. Assumption 5(c) is used to establish a modified form of Cramér's condition in Edgeworth expansion of $l_h(\beta_0)$. Assumption 6 states the rate of convergence of bandwidth h compared to the rate of divergence of n. Assumption 6(a) implies that kernel smoothing parameters need to converge to zero fast enough as $n \to \infty$. It is required that if $h \propto n^{-\kappa}$, where $\frac{1}{2r} < \kappa < 1$ for $r \geq 2$. On the other hand, part (b) indicates that h should not be too small. This condition is required to maintain smoothness of $l_h(\beta_0)$ to derive Cramér's condition in the Edgeworth series analysis.

2.2.2 Additional notation

For further discussion, additional notation is required. We let

$$\lambda = \Omega_n^{1/2} t \text{ and } W_i = \Omega_n^{-1/2} Z_i \quad \text{for } i = 1, ..., n,$$

where $t = t(\beta_0)$, $Z_i = Z_i(\beta_0)$ and $\Omega_n = EZ_iZ_i'$.

Also, let W_i^j denote the jth component of W_i and define

$$\alpha^{j_1\cdots j_k} = EW_i^{j_1}\ldots W_i^{j_k}, \quad \bar{A}^{j_1\cdots j_k} = n^{-1}\sum_{i=1}^{n} W_i^{j_1}\ldots W_i^{j_k},$$
$$\text{and } A^{j_1\cdots j_k} = \bar{A}^{j_1\cdots j_k} - \alpha^{j_1\cdots j_k}$$

In particular, $\alpha^{j_kj_l} = \delta^{j_kj_l}$, where $\delta^{j_kj_l}$ is the Kronecker delta.

Finally, the Einstein summation convention is used (i.e. α^{iikk} and α^{ikm}) for the convenience of expression and calculation. The rules of the summation convention are: (i) In any term in an equation, an index can appear at most twice. (ii) Repeated indices (dummy index) are implicitly summed over. (iii) If an index appears only once (free index), the same index must appear only once in other terms. For example, the summation convention $\alpha^{ij}\alpha^i$ is same as $\sum_i \alpha^{ij}\alpha^i$. Let $\bar{Q}$ be the vector of all distinct first $L+1$ order multivariate centered moments of W_i that

$$\bar{Q} = \left(A^1, ..., A^K, A^{11}, ..., A^{KK}, ..., A^{11\cdots 1}, ..., A^{KK\cdots K}\right)' \equiv \frac{1}{n}\sum_{i=1}^{n} Q_i.$$

Here, Q_i includes elements such as

$$\left\{(G(-U_i/h) - q)(G(X_i'\beta_0 - C_i/h))\right\}^{|\nu|} W_i^{\nu_1}\cdots W_i^{\nu_k}$$
$$\text{for } 1 \leq k \leq L+1, \text{ where } |\nu| = \nu_1 + \cdots + \nu_k$$

so that it covers all terms of $l_h(\beta_0)$.

2.3 Asymptotic properties

Powell (1984, 1986) show $\sqrt{n}$-consistency and asymptotically normality of $\hat{\beta}_{CQ}$. Asymptotic equivalence of the SEL and CQR estimator are established and the asymptotic distribution of SEL estimator is derived.

Theorem 1 *Under Assumptioins 1-5(b) and 6(a), as $n \to \infty$, we have*

$$(a)\ \sqrt{n}\left(\hat{\beta}_E - \hat{\beta}_{CQ}\right) = o_p(1),$$
$$(b)\ \sqrt{n}\left(\hat{\beta}_E - \beta_0\right) \to N\left(0, V_0\right),$$
where $V_0 = D_0^{-1} T_0 D_0^{-1}, D_0 = E\left[X_i' X_i f\left(0|X_i\right) 1\left(X_i'\beta < C_i\right)\right],$
$T_0 = E\left[X_i' X_i 1\left(X_i'\beta < C_i\right)\right]$

(a) implies that under the assumptions of section 2.2, CQR and SEL estimators are asymptotically equivalent. Therefore, using (a) and asymptotic normality of $\hat{\beta}_E$ shown by Powell (1984, 1986), (b) can be derived.

3 Confidence regions and coverage accuracy

3.1 Smoothed empirical likelihood confidence regions

Asymptotic equivalence and the distribution of estimators can be used to construct confidence regions for β_0. First, the SEL confidence region for $\beta_0 \in \mathbb{R}^K$ is defined as

$$I_{SEL} = \{\beta : l_h(\beta) \leq c\},$$

where $c > 0$ determines its coverage probability $P(\beta_0 \in I_{SEL}) = P(l_h(\beta_0) \leq c)$. Theorem 2 establishes the asymptotic distribution of $l_h(\beta_0)$.

Theorem 2 *Suppose assumptions 1-5(b) and 6(a) hold. Then, as $n \to \infty$, we have,*

$$l_h(\beta_0) \xrightarrow{d} \chi^2_K.$$

If $c = c_\alpha$ is chosen from χ^2_K distribution as $P(\chi^2_K \leq c_\alpha) = \alpha$, then Theorem 2 implies that the asymptotic coverage probability of the SEL confidence region I_{SEL} will be α. Therefore, as $n \to \infty$,

$$P(\beta_0 \in I_{SEL}) = P(l_h(\beta_0) \leq c_\alpha) = \alpha + o(1).$$

The higher order properties of SEL confidence region is established in Theorem 3. Whang (2003) uses an Edgeworth expansion of the distribution of $l_h(\beta_0)$ to show that the asymptotic coverage error of the SEL confidence region has an order $O(n^{-1})$.

Theorem 3 *Suppose $c = c_\alpha$ is given as Theorem 2, and assumptions 1-6 hold. If it is further to be assumed that $\sup_n nh^r < \infty$, as $n \to \infty$,*

$$P(\beta_0 \in I_{SEL}) = \alpha + O(n^{-1}).$$

3.2 Bartlett corrected smoothed empirical likelihood confidence regions

With appropriate h, coverage error of the SEL confidence region has the order $O(n^{-1})$. This relatively low order partly stems from the difference between the mean of $l_h(\beta_0)$ and χ^2_K distribution. (i.e. $E[l_h(\beta_0)] \neq K$) Therefore, error can be diminished by adjusting $l_h(\beta_0)$ to have the correct mean. This method is known as the Bartlett correction. As shown by DiCiccio et al. (1991), the empirical likelihood method for constructing confidence intervals is Bartlett-correctable. It is established that Bartlett correction can further reduce the coverage error to order $O(n^{-2})$. From the Taylor expansion of $n^{-1}l_h(\beta_0)$, if $nh^{2r} \to 0$,

$$\begin{aligned} &E[l_h(\beta_0)] = K(1 + n^{-1}b) + o(n^{-1}), \ \text{ where} \\ &b = K^{-1}(\alpha^{iikk}/2 - \alpha^{ikm}\alpha^{ikm}/3) \end{aligned}$$

From the result above, a confidence region corrected with the Bartlett factor b can be considered as

$$I^b_{SEL} = \{\beta : l_h(\beta) \leq c(1 + n^{-1}b)\}.$$

However, the Bartlett factor b is not observed in practice, and has to be estimated. Whang (2003) suggested two estimated Bartlett factors, $\hat{b}$ and $\tilde{b}$, for censored quantile regression models. Let $\hat{\beta}$ is $\sqrt{n}$-consistent estimator of β_0 such as SEL estimator $\hat{\beta}_E$ or usual CQR estimator $\hat{\beta}_{CQ}$.

First, $\hat{b}$ is defined to be

$$\hat{b} = K^{-1}(\hat{\alpha}^{iikk}/2 - \hat{\alpha}^{ikm}\hat{\alpha}^{ikm}/3), \text{ where}$$

$$\begin{aligned}
\hat{\alpha}^{iikk} &= n^{-1}\sum_{j=1}^{n}\hat{\varepsilon}_j^4\left(X_j'\hat{\Omega}_n^{-1}X_j\right)^2,\\
\hat{\alpha}^{ikm} &= n^{-1}\sum_{j=1}^{n}\hat{\varepsilon}_j^3\hat{\omega}_{ni}^{-1/2}X_j\hat{\omega}_{nk}^{-1/2}X_j\hat{\omega}_{nm}^{-1/2}X_j,\\
\hat{\alpha}^{ikm}\hat{\alpha}^{ikm} &= n^{-2}\sum_{j=1}^{n}\sum_{l=1}^{n}\hat{\varepsilon}_j^3\hat{\varepsilon}_l^3\left(X_j'\hat{\Omega}_n^{-1}X_l\right)^3,\\
\hat{\Omega}_n &= n^{-1}\sum_{j=1}^{n}\hat{\varepsilon}_j^2X_jX_j', \text{ and } \hat{\varepsilon}_j = \left(G_h(X_j'\hat{\beta} - Y_j) - q\right)G_h(X_j'\hat{\beta} - C_i)
\end{aligned}$$

and $\hat{\omega}_{ni}^{-1/2}$ is the ith row of $\hat{\Omega}_n^{-1/2}$. The SEL confidence region corrected with $\hat{b}$ is defined to be

$$I_{SEL}^{\hat{b}} = \{\beta : l_h(\beta) \leq c(1 + n^{-1}\hat{b})\}.$$

Also, the other suggested estimated Bartlett factor $\tilde{b}$ is defined with

$$\begin{aligned}
\alpha^{iikk} &= q^{-1}(1-q)^{-1}(1-3q+3q^2)E\left[(X_j'S_0X_j)^2\right] + O(h),\\
\alpha^{ikm} &= q^{-1/2}(1-q)^{-1/2}(1-2q)E\left[(s_i^{-1/2}X_j)(s_k^{-1/2}X_j)(s_m^{-1/2}X_j)\right] + O(h),
\end{aligned}$$

where $S_0 = E[1(X_i'\beta_0 < C_i)X_iX_i']$ and $s_i^{-1/2}$ denote the ith row of $S_0^{-1/2}$. If it is suggested using $\tilde{S} = n^{-1}\sum_{j=1}^{n}\tilde{X}_j\tilde{X}_j'$ and $\tilde{X}_j = 1(X_j'\hat{\beta} < C_j)X_j$ for $j = 1, ..., n$, the Bartlett factor $\tilde{b}$ can be found as

$$\begin{aligned}
\tilde{b} = K^{-1}\Bigg[&\frac{1}{2}\cdot q^{-1}(1-q)^{-1}(1-3q+3q^2)\left\{n^{-1}\sum_{j=1}^{n}(\tilde{X}_j'\tilde{S}^{-1}\tilde{X}_j)^2\right\}\\
&-\frac{1}{3}\cdot q^{-1}(1-q)^{-1}(1-2q)^2\left\{n^{-2}\sum_{j=1}^{n}\sum_{l=1}^{n}\tilde{X}_j'\tilde{S}^{-1}\tilde{X}_j)^3\right\}\Bigg],
\end{aligned}$$

where $\hat{\beta}$ is defined same as above. Then the SEL confidence region rescaled with the Bartlett factor $\tilde{b}$ is

$$I_{SEL}^{\tilde{b}} = \{\beta : l_h(\beta) \leq c(1 + n^{-1}\tilde{b})\}.$$

The estimation of $\tilde{b}$ is computationally simpler than the estimation of $\hat{b}$ because $\tilde{b}$ is approximated with terms that do not depend on bandwidth h.

Theorem 4 below shows that the SEL confidence region adjusted with b has coverage error of order $O(n^{-2})$ and $\hat{b}$ also decreases its coverage error of the same order as b. However, $\tilde{b}$ reduces asymptotic coverage accuracy only to be order $O(n^{-1}h)$. It can be derived using $\tilde{b} = b + O(n^{-1/2}) + O(h)$.

Theorem 4 *Define $c = c_\alpha$ as above. Suppose Assumptions above hold. If it is assumed further that* $\sup_n n^3 h^{2r} < \infty$*, then as* $n \to \infty$*,*

$$\begin{aligned}
&(a)\ \ P(\beta_0 \in I_{SEL}^{b}) = \alpha + O(n^{-2}) \\
&(b)\ \ P(\beta_0 \in I_{SEL}^{\hat{b}}) = \alpha + O(n^{-2}) \\
&(c)\ \ P(\beta_0 \in I_{SEL}^{\tilde{b}}) = \alpha + O(n^{-1}h).
\end{aligned}$$

4 Simulation

In this section, we conduct a simulation study to compare the numerical performance of parameter estimating methods: Censored Quantile Regression (CQR), Bootstrap Censored Quantile Regression (BCQR), Smoothed Quantile Regression (SCQR), unsmoothed Empirical Likelihood (EL), and Smoothed Empirical Likelihood (SEL) and Bartlett corrected with Bartlett factor $\tilde{b}$ and $\hat{b}$ SEL (i.e. SEL1 and SEL2).

4.1 Compared Models in Simulation

4.1.1 Censored Quantile Regression

In this paper, the CQR estimator $\hat{\beta}_{CQ}$ is estimated using the algorithm proposed by Stengos and Wang (2007). The algorithm reduces computational loads in the

CQR estimation problem. The method uses the distance function $\Theta(\beta)$ defined as

$$\Theta(\beta) = \sum_{i=1}^{n} \{q \cdot 1(d_i > 0) + (1-q) \cdot 1(d_i < 0)\} \cdot |d_i|,$$

where $d_i = Y_i - \max(X_i'\beta, C_i)$, and $q \in (0,1)$ is the quantile. $\Theta(\beta)$ is piecewise linear, non-convex, and has local minima. The suggested algorithm is designed to find the minimum point by comparing the values of critical points. It requires only $O(k \times n^2)$ operations with n observations and k regressors, which are simpler than other algorithms.

Powell (1984, 1986) suggests an asymptotic covariance matrix $\hat{V}_0$ for V_0 as

$$\begin{aligned}
\hat{V}_0 &= \hat{D}_0^{-1}\hat{T}_0\hat{D}_0^{-1}, \text{ where} \\
\hat{D}_0 &= \frac{2}{nh}\sum_{i=1}^{n} X_i'X_i K\left(\frac{Y_i - X_i'\hat{\beta}_{CQ}}{h}\right) 1(X_i'\hat{\beta}_{CQ} < C_i), \\
\hat{T}_0 &= \frac{1}{n}\sum_{i=1}^{n} X_i'X_i 1(X_i'\hat{\beta}_{CQ} < C_i),
\end{aligned}$$

Therefore, the confidence regions based on $\hat{\beta}_{CQ}$ is

$$I_{CQR} = \left\{\beta : n\left(\hat{\beta}_{CQ} - \beta\right)' \hat{V}_0^{-1}\left(\hat{\beta}_{CQ} - \beta\right) \leq c_\alpha\right\}$$

where c_α is a α-quantile of the χ_2^2 distribution. In this simulation, the second order kernel function $K_1(u) = \left(\frac{15}{16}\right)(1-u^2)1(|u| \leq 1)$ is used. (see Horowitz, 1998)

4.1.2 Bootstrap Censored Quantile Regression

We consider the confidence region for BCQR by

$$I_{BCQR} = \left\{\beta : n\left(\hat{\beta}_{CQ} - \beta\right)' V^{*-1}\left(\hat{\beta}_{CQ} - \beta\right) \leq c_\alpha\right\}$$

where V^* is a bootstrap estimator of V_0 (see Efron, (1979, 1982) and Buchinsky (1995)) defined as below.

$$V^* = \frac{n}{B}\sum_{b=1}^{B}(\hat{\beta}^*_{CQb} - \bar{\hat{\beta}}^*_{CQ})(\hat{\beta}^*_{CQb} - \bar{\hat{\beta}}^*_{CQ})',$$

where $\bar{\hat{\beta}}^*_{CQ} = (1/B)\Sigma_{b=1}^{B}\hat{\beta}^*_{CQb}$ and $\{\hat{\beta}^*_{CQb} : b = 1, ..., B\}$ are B estimates of β_0, calculated in the bootstrap samples from the estimation data $\{(Y_i, X_i) : i = 1, ..., n\}$. The bootstrap method has an advantage in that it is not necessary to choose the optimal bandwidth h. However, the bootstrap estimation requires the heavier computational load than the kernel-based estimation.

4.1.3 Smoothed Censored Quantile Regression

Let $\tilde{\beta}_S$ be a SCQR estimator of β_0. It solves

$$\min_{\beta \in B} \tilde{H}_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} g_{sc}(Y_i, X_i, C_i, h, \beta) \quad \text{where}$$

$$g_{sc}(y, x, c, h, \beta) = \left\{(y - x'\beta)\left[\tilde{K}\left(\frac{y - x'\beta}{h}\right) - q\right] - y\right\}\tilde{K}\left(\frac{x'\beta - c}{h} - 2\right).$$

where $\tilde{K}$ is the integral of a fourth-order kernel for the nonparametric density estimation (Müler, 1984) given as

$$\tilde{K}(u) = \begin{cases} 0 & \text{if } u < -1, \\ 0.5 + \frac{105}{64}\left[u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7\right] & \text{if } |u| \le 1, \\ 1 & \text{otherwise} \end{cases}$$

The smoothed version of $1(X_i'\beta > C_i)$ on the censored quantile regression model is $\tilde{G}\left(\frac{X_i'\beta - C_i}{h}\right)$. However, to prevent asymptotic bias $\tilde{G}\left(\frac{X_i'\beta - C_i}{h} - 2\right)$ is used instead. Under conditions given by Horowitz (1998), $\sqrt{n}(\hat{\beta}_{CQ} - \tilde{\beta}_S) =$

$o_p(1)$. V_0 is estimated consistently by

$$\tilde{V} = \tilde{D}_n(\tilde{\beta}_S)^{-1}\tilde{T}_n\tilde{D}_n(\tilde{\beta}_S)^{-1}, \text{ where}$$
$$\tilde{D}_n(\tilde{\beta}_S) = (nh)^{-1}\sum_{i=1}^{n} X_i'X_i\tilde{K}^{(1)}\left(\tfrac{Y_i - X_i'\tilde{\beta}_S}{h}\right)1(Y_i > 0),$$
$$\tilde{T}_n(\tilde{\beta}_S) = (n)^{-1}\sum_{i=1}^{n}\left[\partial g_c(Y_i, X_i, h, \tilde{\beta}_S)/\partial\beta\right]\left[\partial g_c\left(Y_i, X_i, h, \tilde{\beta}_S\right)/\partial\beta\right]'$$

where $\tilde{K}^{(1)}(\cdot)$ denote the first derivative of $\tilde{K}(\cdot)$. Therefore, the confidence region based on SCQR estimator is

$$I_{SCQR} = \left\{\beta : n\left(\tilde{\beta}_S - \beta\right)'\tilde{V}^{-1}\left(\tilde{\beta}_S - \beta\right) \leq \tilde{c}_\alpha\right\}.$$

The critical value $\tilde{c}_\alpha$ is obtained from the following bootstrap analogue.
(i) Generate a bootstrap sample $\{(Y_i^*, X_i^*) : i = 1, ..., n\}$ by sampling from estimation data.
(ii) Compute the SCQR estimate $\tilde{\beta}_S^*$ using the same algorithm used in the CQR estimation, and calculate its variance estimate $\tilde{V}^*$. Derive $S_n^* = n(\tilde{\beta}_S^* - \tilde{\beta}_S)'\tilde{V}^{*-1}(\tilde{\beta}_S^* - \tilde{\beta}_S)$.
(iii) Estimate the bootstrap distribution of S_n^* by the empirical distribution that is obtained by repeating (i)-(ii) with the bootstrap iteration number B times.
(iv) The bootstrap critical value $\tilde{c}_\alpha$ is estimated by taking α-quantile of this empirical distribution.

4.2 Setup

We consider a right censored model

$$Y_i = \min\left[C_i, X_i'\beta_0 + U_i\right], \text{ for } i = 1, ..., n$$

where $X_i = (1, X_{2i})'$, $\beta_0 = (\beta_{01}, \beta_{02})'$ is a 2×1 parameter vector whose true value is $\beta_0 = (1, 1)'$. In this paper, three different scenarios for generating random error U_i are considered:

∘ Scenario 1: Student t distribution with 3 degrees of freedom rescaled to have variance 2

∘ Scenario 2: Heteroskedastic $U_i = 0.25(1 + X_{2i})V_i$, where $V_i \sim N(0, 1)$

∘ Scenario 3: χ^2 distribution with 3 degrees of freedom recentered to have median zero but skewed

Scenario 1 and 2 are the same as Horowitz (1998), and Scenario 3 is used by Chen and Hall (1993).

In this paper, coverage probabilities of confidence regions of β_0 are presented. C_i is given by 0 and q is given as median (i.e. $q = 1/2$). The second-order kernel $K(u) = (\frac{3}{4})(1 - u^2)1(|u| \leq 1)$ is used to smooth empirical likelihood.

In cases with computing CQR, SCQR and SEL confidence regions, it is required to choose a bandwidth h. This paper considers a rule of thumb $h = c_0\, n^{\gamma}$ in our simulations and take $\gamma \in [-0.16, -0.32, -0.48, -0.64, -0.80]$. We take $c_0 = 0.06$ for CQR, $c_0 = 1.0$ for BCQR and SCQR, and $c_0 = 3.5$ for SEL. However, as will be seen from the simulation results, coverage error probabilities of the SEL confidence regions show small differences over wide variations of c_0 and γ values. We set the number of simulation repetition as $5,000$ for CQR, EL and SEL methods. For BCQR and SCQR estimations, the number of repetition is $1,000$ due to the heavy computation with bootstrapping. The number of bootstrap repetitions is restricted to $B = 100$. Simulations are conducted with five different sample sizes $n \in [20, 30, 40, 50, 60]$. The Intel Core i5 2.3GHz with 4GB memory computer is used under the same conditions.

4.3 Results

Tables 1-3 summarize simulation results for Scenario 1, 2, and 3, respectively. Results show that the coverage probabilities of the CQR confidence regions are relatively poor and very sensitive to the choice of bandwidth h. For example, for Scenario 1 with $n = 40$, the coverage probability is 0.856 when $\sigma = 0.16$ whereas the coverage probability is 0.330 when $\sigma = 0.80$.

On the other hand, coverage probabilities of BCQR, SCQR, and SEL confidence regions are relatively stable across different error cases. However, the SEL

Table 1: Estimated true coverage probabilities of α-level confidence region (Scenario 1)

n	$-\gamma$	CQR	BCQR	SCQR	EL	SEL	SEL1	SEL2
$\alpha = 0.90$								
20	0.16	0.576	0.948	0.913	0.874	0.818	0.818	0.837
	0.32	0.421	0.948	0.926	0.874	0.839	0.839	0.858
	0.48	0.301	0.948	0.934	0.874	0.852	0.852	0.868
	0.64	0.399	0.940	0.945	0.874	0.862	0.862	0.878
	0.80	0.363	0.938	0.952	0.874	0.867	0.867	0.880
40	0.16	0.856	0.950	0.899	0.889	0.869	0.869	0.882
	0.32	0.669	0.946	0.930	0.889	0.882	0.882	0.892
	0.48	0.487	0.942	0.943	0.889	0.893	0.893	0.904
	0.64	0.389	0.944	0.945	0.889	0.893	0.893	0.901
	0.80	0.330	0.944	0.946	0.889	0.893	0.893	0.902
60	0.16	0.945	0.940	0.911	0.896	0.891	0.891	0.899
	0.32	0.803	0.932	0.923	0.896	0.894	0.894	0.904
	0.48	0.601	0.920	0.927	0.896	0.897	0.897	0.903
	0.64	0.454	0.932	0.935	0.896	0.897	0.897	0.904
	0.80	0.375	0.920	0.932	0.896	0.897	0.897	0.901
$\alpha = 0.95$								
20	0.16	0.612	0.970	0.954	0.923	0.873	0.873	0.893
	0.32	0.449	0.982	0.968	0.923	0.894	0.894	0.908
	0.48	0.325	0.974	0.969	0.923	0.906	0.907	0.919
	0.64	0.404	0.968	0.974	0.923	0.914	0.914	0.926
	0.80	0.366	0.976	0.977	0.923	0.920	0.920	0.930
40	0.16	0.875	0.966	0.956	0.944	0.935	0.935	0.941
	0.32	0.702	0.964	0.971	0.944	0.943	0.943	0.948
	0.48	0.522	0.968	0.978	0.944	0.944	0.944	0.949
	0.64	0.403	0.964	0.974	0.944	0.943	0.943	0.949
	0.80	0.336	0.964	0.977	0.944	0.945	0.945	0.950
60	0.16	0.950	0.954	0.951	0.944	0.944	0.944	0.947
	0.32	0.823	0.958	0.954	0.944	0.949	0.949	0.953
	0.48	0.632	0.962	0.962	0.944	0.949	0.949	0.951
	0.64	0.474	0.960	0.967	0.944	0.947	0.947	0.951
	0.80	0.384	0.966	0.976	0.944	0.949	0.949	0.952

Table 2: Estimated true coverage probabilities of α-level confidence region (Scenario 2)

n	$-\gamma$	CQR	BCQR	SCQR	EL	SEL	SEL1	SEL2
$\alpha = 0.90$								
20	0.16	0.546	0.924	0.894	0.885	0.830	0.831	0.852
	0.32	0.406	0.932	0.920	0.885	0.851	0.851	0.869
	0.48	0.312	0.922	0.937	0.885	0.863	0.863	0.879
	0.64	0.383	0.934	0.938	0.885	0.870	0.871	0.887
	0.80	0.356	0.934	0.947	0.885	0.876	0.878	0.894
40	0.16	0.828	0.934	0.912	0.897	0.885	0.885	0.897
	0.32	0.625	0.926	0.938	0.897	0.891	0.891	0.900
	0.48	0.455	0.928	0.945	0.897	0.895	0.895	0.905
	0.64	0.370	0.928	0.943	0.897	0.894	0.895	0.900
	0.80	0.321	0.924	0.949	0.897	0.895	0.895	0.901
60	0.16	0.936	0.930	0.886	0.903	0.889	0.889	0.896
	0.32	0.794	0.928	0.904	0.903	0.891	0.891	0.898
	0.48	0.591	0.934	0.910	0.903	0.891	0.891	0.899
	0.64	0.445	0.928	0.921	0.903	0.894	0.894	0.897
	0.80	0.394	0.922	0.923	0.903	0.896	0.896	0.901
$\alpha = 0.95$								
20	0.16	0.582	0.962	0.951	0.938	0.890	0.891	0.907
	0.32	0.432	0.966	0.963	0.938	0.905	0.905	0.918
	0.48	0.327	0.966	0.972	0.938	0.919	0.920	0.931
	0.64	0.389	0.960	0.973	0.938	0.926	0.926	0.937
	0.80	0.362	0.952	0.977	0.938	0.931	0.931	0.940
40	0.16	0.852	0.958	0.956	0.948	0.935	0.935	0.941
	0.32	0.660	0.960	0.968	0.948	0.943	0.943	0.948
	0.48	0.486	0.960	0.975	0.948	0.944	0.944	0.949
	0.64	0.386	0.960	0.978	0.948	0.943	0.943	0.949
	0.80	0.326	0.956	0.977	0.948	0.945	0.945	0.950
60	0.16	0.944	0.966	0.953	0.951	0.942	0.942	0.947
	0.32	0.815	0.968	0.957	0.951	0.948	0.948	0.953
	0.48	0.625	0.966	0.964	0.951	0.947	0.947	0.950
	0.64	0.466	0.976	0.970	0.951	0.948	0.948	0.951
	0.80	0.402	0.960	0.965	0.951	0.948	0.948	0.951

Table 3: Estimated true coverage probabilities of α-level confidence region (Scenario 3)

n	$-\gamma$	CQR	BCQR	SCQR	EL	SEL	SEL1	SEL2
$\alpha = 0.90$								
20	0.16	0.419	0.888	0.930	0.877	0.842	0.842	0.859
	0.32	0.350	0.876	0.951	0.877	0.861	0.861	0.877
	0.48	0.302	0.894	0.953	0.877	0.866	0.866	0.884
	0.64	0.419	0.886	0.965	0.877	0.867	0.867	0.885
	0.80	0.388	0.896	0.957	0.877	0.872	0.872	0.885
40	0.16	0.598	0.914	0.928	0.891	0.874	0.874	0.884
	0.32	0.456	0.920	0.934	0.891	0.885	0.885	0.894
	0.48	0.385	0.902	0.937	0.891	0.891	0.891	0.898
	0.64	0.379	0.898	0.945	0.891	0.891	0.891	0.898
	0.80	0.358	0.916	0.948	0.891	0.891	0.891	0.901
60	0.16	0.763	0.904	0.927	0.888	0.882	0.882	0.889
	0.32	0.578	0.894	0.932	0.888	0.890	0.890	0.896
	0.48	0.446	0.886	0.933	0.888	0.892	0.892	0.896
	0.64	0.410	0.912	0.944	0.888	0.891	0.892	0.896
	0.80	0.396	0.894	0.941	0.888	0.875	0.895	0.900
$\alpha = 0.95$								
20	0.16	0.435	0.940	0.973	0.932	0.900	0.901	0.914
	0.32	0.358	0.924	0.983	0.932	0.915	0.915	0.924
	0.48	0.305	0.938	0.981	0.932	0.922	0.922	0.933
	0.64	0.420	0.932	0.985	0.932	0.925	0.926	0.937
	0.80	0.388	0.938	0.986	0.932	0.927	0.927	0.937
40	0.16	0.634	0.944	0.966	0.944	0.933	0.933	0.939
	0.32	0.475	0.952	0.972	0.944	0.941	0.941	0.947
	0.48	0.394	0.952	0.970	0.944	0.943	0.943	0.948
	0.64	0.382	0.948	0.980	0.944	0.944	0.944	0.948
	0.80	0.360	0.948	0.977	0.945	0.941	0.941	0.947
60	0.16	0.782	0.938	0.960	0.945	0.934	0.934	0.939
	0.32	0.604	0.936	0.970	0.945	0.942	0.942	0.945
	0.48	0.464	0.952	0.975	0.945	0.944	0.944	0.948
	0.64	0.418	0.950	0.967	0.945	0.945	0.945	0.948
	0.80	0.398	0.944	0.971	0.945	0.944	0.944	0.948

Figure 4.1: Estimated coverage error (SCQR and SEL2)

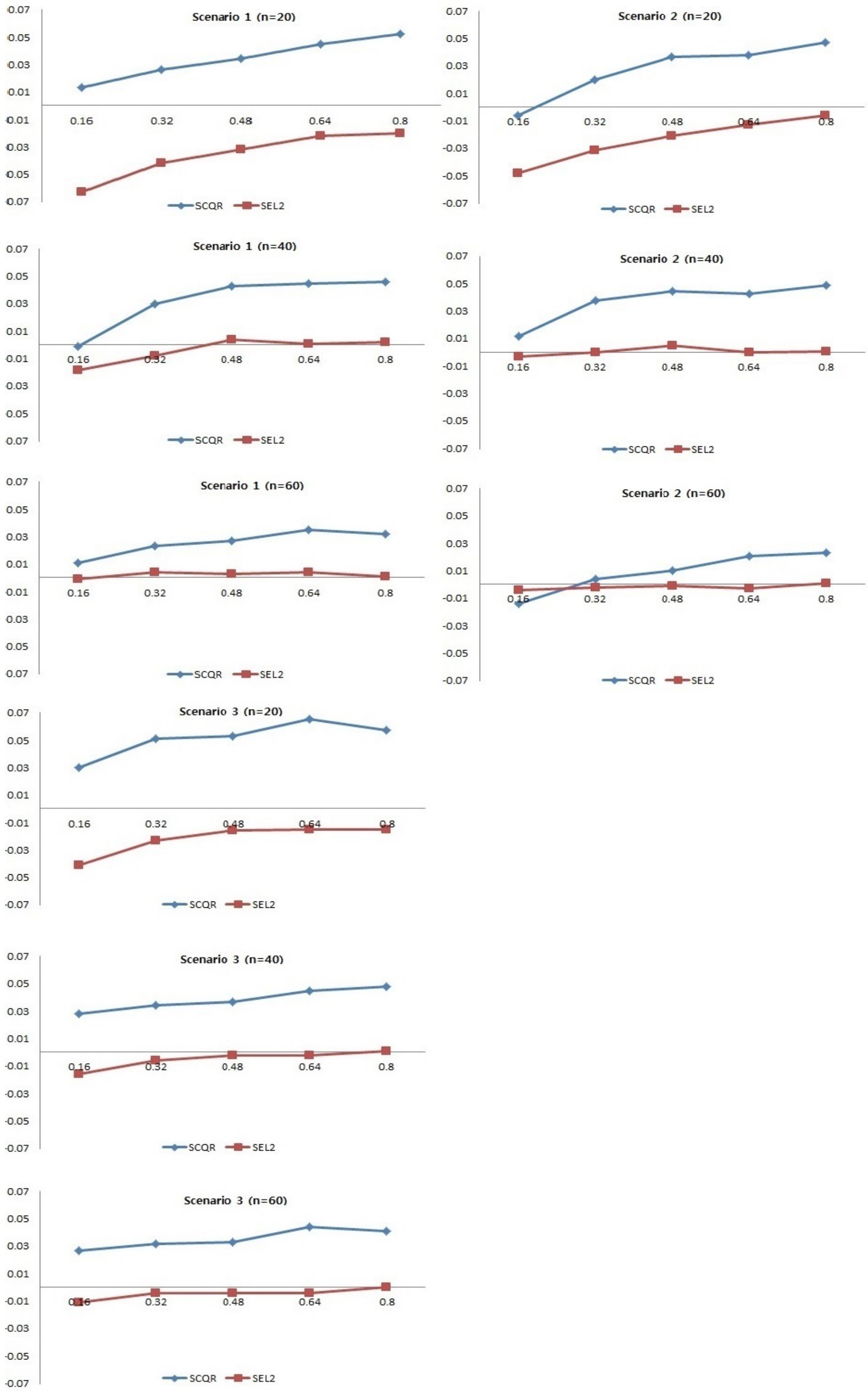
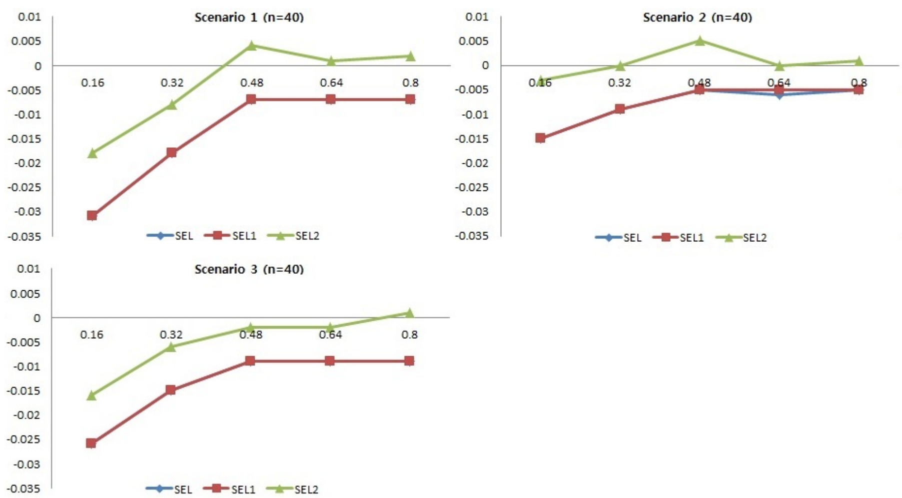


Table 4: Average estimation time per 100 iteration (seconds)

n	CQR	BCQR	SCQR	EL	SEL	SEL1	SEL2
20	4.540	1001	138.6	4.320	4.890	4.770	4.790
30	12.60	2373	148.1	11.58	12.56	13.49	13.23
40	24.92	5003	155.7	24.34	26.06	25.63	24.78
50	41.64	7620	173.6	43.18	39.83	43.69	43.28
60	62.18	11620	195.0	63.12	64.39	62.65	61.72

Figure 4.2: Estimated coverage error (SEL, SEL1, and SEL2)



confidence region is less sensitive to the bandwidth h than BCQR and SCQR confidence regions, especially for $n \geq 40$.

The increasing number of sample size reduces coverage errors of confidence regions for all methods. However, the SEL confidence region Bartlett corrected with $\hat{b}$ (SEL2) outperforms in most cases especially when n is relatively large. Figure 4.1 indicates that errors of SEL2 decrease faster than SCQR as the sample size increases. This confirms Theorem of Whang (2003) that the SEL2 confidence region coverage error has order of $O(n^{-2})$, which is higher than SCQR coverage error order of $O(n^{-a})$ for $a < 1$.

The unsmoothed EL method shows similar or better performance than SEL or SEL1. The SEL confidence region with no Bartlett correction (SEL) shows almost the same performance as the confidence region of Bartlett corrected with $\tilde{b}$ (SEL1). SEL2 shows better performance than SEL and SEL1 (see Figure 4.2). This result suggests that implementing smoothing equations is not necessary unless researchers want to achieve higher-order improvements using Bartlett correction. Also, it verifies Theorem 4 in which the order of coverage errors of SEL and SEL1 are similar ($O(n^{-1})$ and $O(n^{-1}h)$ each) whereas order of coverage accuracy of SEL2 is higher ($O(n^{-2})$).

Table 4 demonstrates the estimation of each compared methods. It is estimated based on Scenario 1, and time is estimated of average seconds per 100 iteration. BCQR requires the longest time whereas CQR, EL, SEL, SEL1, and SEL2 show similar estimation time. SCQR also needs relatively longer time than CQR, EL, and SEL based methods. However, the amount of additional time required as n increase is almost similar to other methods except BCQR.

5 Concluding Remarks and Extensions

This paper verifies finite sample properties of Whang (2003) that consider the smoothed empirical likelihood-based method on the censored quantile regression model. We have shown that SEL confidence regions achieve the first-order asymptotic properties. Simulation results show that the Bartlett-corrected SEL confidence region has higher-order refinements, which are better than the refinements based on the bootstrap methods.

In a future study, one can compare the SEL methods and the random censoring methods. Koenker (2008) compared a standard Powell Estimator for fixed censoring with random censoring methods of Portnoy (2003) and Peng and Huang (2008). It would be interesting to examine Bartlett corrected SEL and random censoring methods especially for small sample cases. In addition, there are several directions to extend simulation studies. Experiments could be conducted under different error types. This paper only considers error structures which are aimed to analyze the quantile regression model. However, other error types can be implemented in the simulation study, such as error cases of Chernozhukov and Hong (2002), and Pang, et al. (2010) designed for the CQR model. Also, simulations under different censoring values and quantiles could be considered.

References

[1] Buchinsky, M. (1995) Estimating the asymptotic covariance matrix for quantile regression models: A Monte Carlo study, *Journal of Econometrics* 68, 303338.

[2] Chen, S.X. & P. Hall (1993) Smoothed empirical likelihood confidence intervals for quantiles. *Annals of Statistics* 21, 1166-1181.

[3] Chernozhukov, V. and Hong, H. (2002) Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association* 97, 872-882.

[4] DiCiccio, T.J., P. Hall, and J.P. Romano (1991) Empirical likelihood is Bartlett-correctable. *Annals of Statistics* 19, 1053-1061.

[5] DiCiccio, T.J., P. Hall, and J.P. Romano (1988) Bartlett adjustment for empirical likelihood. *Technical Report* 298, Dept. Statistics, Stanford Univ.

[6] Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 14, 1431-1452.

[7] Hahn, J. (1995) Bootstrapping quantile regression estimators. *Econometric Theory* 11, 105-121.

[8] Horowitz, J.L. (1998) Bootstrap methods for median regressions models. *Econometrica* 66, 1327-1351.

[9] Huang, Y. (2010) Quantile calculus and censored regression. *Annals of Statistics* 38, 1607-1637.

[10] Koenker, R. (2008) Censored quantile regression redux. *Journal of Statistical Software* 27, 1-25.

[11] Koenker, R. and Basset, G. (1978) Regression quantiles. *Econometrica* 46, 33-50.

[12] Koenker, R and B.J. Park (1996) An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics* 71, 265-283.

[13] Müller, H.G. (1984) Smooth optimum kernel estimators of densities, regression curve and models. *The Annals of Statistics* 12(2), 766-774.

[14] Owen, A.B. (1988) Empirical Likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237-249.

[15] Owen, A.B. (1990) Empirical Likelihood ratio confidence regions. *Annals of Statistics* 18, 90-120.

[16] Peng, L. and Huang, Y. (2008) Survival analysis with quantile regression models. *Journal of the American Statistical Association* 98, 1001-1012.

[17] Pinkse, C.A.P (1993) On the computation of semiparametric estimates in limited dependent variable models. *Journal of Econometrics* 58, 185-205.

[18] Portnoy, S. (2003) Censored regression quantiles. *Journal of the American Statistical Association* 103, 637-649

[19] Powell, J.L. (1984) Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25, 303-325.

[20] Powell, J.L. (1986) Censored regression quantiles. *Journal of Econometrics* 32, 143-155.

[21] Stengos and Wang, (2007) An algorithm for censored quantile regressions. *Economics Bulletin*, Vol. 3, No. 1 pp 1-9

[22] Wang, H. and Wang, L. (2009) Locally weighted censored quantile regression. *Journal of the American Statistical Association* 104, 1117-1128.

[23] Whang, Y.J. (2003) Smoothed empirical likelihood methods for quantile regression models. *Working Paper*.

[24] Whang, Y.J. (2006) Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory* 22, 173-205.

[25] Womersley, R.S., (1986), Censored Discrete Linear l_1 Approximation. *SIAM Journal of Scientific and Statistical Computing* 7, 105-122.

APPENDIX

Lemma 1. *Under Assumptions 1-5(b) and 6(a), as* $n \to \infty$,

$$(a)\ EZ_i(\beta_0) = (-h)^r \left(\frac{1}{r!}\right) C_K F(Y_i|X_i) E\left[X_i f^{(r-1)}(0|X_i)\right] + o(h^r)$$
$$(b)\ EZ_i(\beta_0)Z_i(\beta_0)' = q(1-q)S_0 + o(1)$$
$$(c)\ E\frac{\partial Z_i(\beta_0)}{\partial \beta'} = D_0 + o(1).$$

where $S_0 = E[1(X_i'\beta_0 < C_i)X_iX_i']$ *and* $D_0 = E[f(0|X_i)1(X_i'\beta_0 < C_i)X_iX_i']$.

Proof of Lemma 1. By a change of variables, we have

$$\begin{aligned} EZ_i(\beta_0) &= E\left[\left\{\int [F(-uh|X_i) - F(0|X_i)]K(u)du\right\}\right. \\ &\qquad \left. \times \left\{\int F(Y_i - uh|X_i)K(u)du\right\} X_i\right]. \end{aligned}$$

By applying a Taylor expansion to the each equation in the integral, (a) can be derived. Similarly, parts (b) and (c) can be derived by using

$$\begin{aligned} &EZ(\beta_0)Z(\beta_0)' \\ &= q(1-q)E[1(X_i'\beta_0 < C_i)X_iX_i'] \\ &+ 2E\left\{\int [F(-uh|X_i) - F(0|X_i)][G(-u) - q]\left[1(X_i'\beta_0 < C_i)\right] K(u)duX_iX_i'\right\}, \end{aligned}$$

and

$$
\begin{aligned}
&E\frac{\partial Z(\beta_0)}{\partial \beta'} \\
&= E[f(0|X_i)1(X_i'\beta_0 < C_i)X_iX_i'] \\
&+ E\left[\int [f(-uh|X_i) - f(0|X_i)]K(u)du \cdot 1(X_i'\beta_0 < C_i)X_iX_i'\right] \\
&+ E\left[\int \{f(-uh|X_i) \cdot (G((X_i'\beta_0 - C_i)/h) - 1(X_i'\beta_0 < C_i)) \cdot K(u)du\}X_iX_i'\right].
\end{aligned}
$$

Lemma 2. *Suppose Assumptions 1-5(b) and 6(a) hold. Then,*
(a) There exists a $K \times 1$ vector $\hat{\beta}_E \in int(B)$ such that $l_h(\beta)$ attains its minimum value at $\hat{\beta}_E$
(b) $\hat{\beta}_E$ satisfies $t(\hat{\beta}_E) = 0$ and $Q_n(\hat{\beta}_E) = 0$, where $Q_n(\beta) = n^{-1}\sum_{i=1}^n Z_i(\beta)$ with probability 1 as $n \to \infty$.

Proof of Lemma 2. Lemma 2 can be proved using Lemma 2 and Lemma 3 of Whang (2006).

Proof of Theorem 1. By Lemma 1 and the weak law of large numbers, we have $\partial Q_n(\beta_0)/\partial\beta' \overset{p}{\to} D_0$. Letting $G_{ni} \equiv [G(-U_i/h) - 1(U_i \leq 0)][G(X_i'\beta/h)]$ and rearranging terms, we have

$$
\begin{aligned}
\sqrt{n}Q_n(\beta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n [G(-U_i/h) - q]\left[G\left(X_i'\beta_0 - C_i/h\right)\right]X_i \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n [1(U_i \leq 0) - q]\left[1(X_i'\beta_0 < C_i)\right]X_i \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n [1(U_i \leq 0) - q]\left[G\left(\frac{X_i'\beta_0 - C_i}{h}\right) - 1(X_i'\beta_0 < C_i)\right]X_i \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n [G_{ni}X_i - EG_{ni}X_i] + \sqrt{n}EG_{ni}X_i
\end{aligned}
$$

Here, the third term is $O_p(h^{1/2})$ and hence $o_p(1)$ since, for each $\varepsilon > 0$,

$$
\begin{aligned}
& P\left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [G_{ni} X_i - E G_{ni} X_i] \right\| > \epsilon \right) \\
& \leq \epsilon^{-2} E \left[\left\{ G\left(\frac{-U_i}{h} \right) - 1(U_i \leq 0) \right\}^2 \left\{ G\left(\frac{X_i' \beta - C_i}{h} \right) \right\}^2 \right] \|X\|^2 \\
& \leq C \cdot P(-h \leq U \leq h) = O(h).
\end{aligned}
$$

Also, the last term is $o(1)$ since using Assumption 6(a), we have $\sqrt{n} E G_{ni} X_i = \sqrt{n} E Z_i(\beta_0) = O(n^{1/2} h^r) \to 0$.

Therefore, a Taylor series expansion of $Q_n(\hat{\beta}_E)$ about β_0 yields,

$$
\sqrt{n} \left(\hat{\beta}_E - \beta_0 \right) = -D_0^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [1(U_i \leq 0) - q] \left[1(X_i' \beta_0 < C_i) \right] X_i + o_p(1).
$$

Now applying a Bahadur representation of the censored quantile regression to the equation above leads to the result of Theorem 1.

Proof of Theorem 2. Theorem 2 can be proved with same arguments of Whang (2006, Theorem 2) after checking following conditions. Re-write $l_h(\beta_0)$ as

$$
l_h(\beta_0) = 2 \sum_{i=1}^{n} \log \left(1 + \lambda' W_i \right),
$$

where λ satisfies

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

Let $\lambda \equiv \lambda(\beta_0)$ denote the solution of the equation above. It is easy to see that $n^{-1} \sum W_i W_i' \xrightarrow{p} E W_i W_i' = I_K$ by a weak law of large numbers, $n^{-1} \sum W_i = O_p(n^{-1/2} + h^r)$, and $\max_i \|W_i\| = O_p(1)$.

$\lambda = O_p(n^{-1/2} + h^r)$. Also, by Lemma 1(a), we can check $\alpha^j = O(h^r)$,

$\bar{A}^j = A^j + \alpha^j = O_p(n^{-1/2} + h^r)$, $A^{jk} = O_p(n^{-1/2})$, and $\bar{A}^{j_1 \cdots j_k} = O_p(1)$ for $k \geq 3$.

Lemma 3. *Let* t *be a vector that has the same dimension with* Q*. Define* $I(t,h) = E\{\exp[it'Q]\}$*, where* $i = \sqrt{-1}$*. Under assumptions 1-6, we have, for each* $\varepsilon > 0$*, some* $C > 0$*, all* t *satisfying* $||t|| > \varepsilon$*, and all sufficient small* h*,*

$$|I(t,h)| < 1 - Ch.$$

In Lemma 3, the modified version of the Cramér's condition for the Edgeworth expansion is established.

Proof of Lemma 3. The proof of Lemma 3 is similar to Horowitz (1998, Lemma 9) and Whang (2006, Lemma 4). G satisfies $|G(v)| = 0$ or 1 if $|v| \geq 1$. Let $\delta^- = G(v)$ if $v \leq -1$ and $\delta^+ = G(v)$ if $v \geq 1$.

$$\begin{aligned}
&I(t,h) \\
&= E\left[\exp(it'Q)\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(i\sum_{r=0}^{L+1}[G(-U/h)G((X'\beta_0 - C_i)/h)]^r t_r' g_r(X)\right) f(U|X)dUdP(X) \\
&= I_1(t,h) + I_2(t,h),
\end{aligned}$$

where

$$I_1(t,h) =$$
$$E\left[F(-h|X)\exp\left(i\sum_{r=0}^{L+1} t_r' g_r(X)\delta_r^-\right) + \{1 - F(h|X)\}\exp\left(i\sum_{r=0}^{L+1} t_r' g_r(X)\delta_r^+\right)\right],$$

and

$$I_2(t,h) =$$
$$\int_{-\infty}^{\infty}\int_{-h}^{h} \exp\left(i\sum_{r=0}^{L+1}[G(-U/h)G((X'\beta_0 - C_i)/h)]^r t_r' g_r(X)\right) f(U|X)dUdP(X).$$

Here, $g_r(X)$ denote the product of elements of X which is the r-th order polynomial $[G(-U/h)]^r$ in the expansion of $t'Q$.

First, consider $I_1(t,h)$. For h sufficiently small, by Assumption 4(b) and Taylor series expansion,

$$\begin{aligned} |I_1(t,h)| &\leq \left| F(-h|X)\exp\left(i\sum_{r=0}^{L+1} t_r' g_r(X)\delta_r^-\right) \right. \\ &\quad \left. + \{1-F(h|X)\}\exp\left(i\sum_{r=0}^{L+1} t_r' g_r(X)\delta_r^+\right) \right| \\ &\leq F(-h|X) + 1 - F(h|X) \\ &\leq 1 - Ef(0|X)h \end{aligned}$$

Now, consider $I_2(t,h)$. By a change of varibles,

$$I_2(t,h) =$$
$$h\int_{-\infty}^{\infty}\int_{-1}^{1} -\exp\left(i\sum_{r=0}^{L+1}[G(u)G((X'\beta_0 - C_i)/h)]^r t_r' g_r(X)\right) f(-uh|X)dudP(X).$$

Given $\varepsilon > 0$, for sufficiently small h, there exists $\delta(\varepsilon) > 0$ such that

$$\int_{-\infty}^{\infty}\int_{-1}^{1} |f(-uh|X) - f(0|X)|\, du\, dP(X) \leq 2\delta(\varepsilon)Ef(0|X).$$

Then

$$|I(t,h)| \leq 1 - hEf(0|X)(1-2\delta(\varepsilon)) + |I_3(t,h)|,$$

where

$$I_3(t,h) =$$
$$-h\int_{-\infty}^{\infty}\int_{-1}^{1} \exp\left(i\sum_{r=0}^{L+1}[G(u)G((X'\beta_0 - C_i)/h)]^r t_r' g_r(X)\right) f(0|X)dudP(X).$$

There are $\eta > 0$ and $\gamma_1 < 1$ such that

$$2\int_{||x||\leq\eta} f(0|X)dP(X) = \gamma_1 Ef(0|X).$$

With a partition of $[-1, 1]$ that satisfies Assumption 5(c) and using an argument similar to Horowitz (1998, Proof of Lemma 9), we have

$$\sup_{||t||>\varepsilon} \int_{-1}^{1} \left| \exp\left(i\sum_{r=0}^{L+1} [G(u)G((X'\beta_0 - C_i)/h) \right) t_r' g_r(X) \right| = C_1$$

for some $C_1 < 1$. Combining above yields, for $\varepsilon > 0$ and for all $h > 0$ sufficiently small,

$$\sup_{||t||>\varepsilon} |I(t, h)| \leq 1 - hEf(0|X)(1 - 2\delta(\varepsilon) - \gamma_2) = 1 - C(\varepsilon)h$$

where $\gamma_2 = [\gamma_1 + (1 - \gamma_1)C_1(\varepsilon)] < 1$. This establishes Lemma 3.

Proof of Theorem 3. Under the modified Cramér's condition proven at Lemma 3, the proof of Theorem 3 can be proven using the same argument with Whang (2006, Theorem 3).

Proof of Theorem 4. From the Edgeworth expansion for the distribution of $l_h(\beta_0)$, for any $c > 0$,

$$\begin{aligned}
&P\left(l_h(\beta_0) \leq c(1 + n^{-1}b)\right) \\
&= P(\chi_K^2 \leq c(1 + n^{-1}b)) - (n\alpha^i\alpha^i K^{-1} + n^{-1}b)c(1 + n^{-1}b)g_K\left[c(1 + n^{-1}b)\right] \\
&\quad + O(n^{-2}) + o(nh^{2r}).
\end{aligned}$$

where g_K is the density of the χ_K^2 distribution. Therefore, $g_K\left[c(1 + n^{-1}b)\right] = g_K(c) + O(n^{-1})$ and $P(\chi_K^2 \leq c(1 + n^{-1}b)) = P(\chi_K^2 \leq c) + cn^{-1}bg_K(c) +$

$O(n^{-2})$. By combining above equations and Lemma 1(a),

$$
\begin{aligned}
&P\left(l_h(\beta_0) \le c(1+n^{-1}b)\right) \\
&= P(\chi^2_K \le c) - cn\alpha^i\alpha^i K^{-1} g_K(c) + O(n^{-2}) + o(nh^{2r}) \\
&= P(\chi^2_K \le c) - c \cdot nh^{2r} \left(\frac{1}{r!}\right)^2 C_K^2 (F(Y_i|X_i))^2 (\zeta' S^{-1}\zeta) q^{-1}(1-q)^{-1} K^{-1} g_K(c) \\
&\quad + O(n^{-2}) + o(nh^{2r}),
\end{aligned}
$$

since $n\alpha^i\alpha^i = \frac{1}{n}(nh^r)^2 \frac{1}{r!} C_K^2 (F(Y|X))^2 (\zeta' S^{-1}\zeta) q^{-1}(1-q)^{-1}$, where $\zeta = E[X \cdot f^{(r-1)}(0|X)]$. Here, since $nh^{2r} \to 0$ and $\sup_n n^3 h^{2r} < \infty$, for all $c > 0$,

$$
P\left(l_h(\beta_0) \le c(1+n^{-1}b)\right) = P(\chi^2_K \le c) + O(n^{-2}).
$$

The proof of Theorem 4(a) is completed by taking $c = c_\alpha$ that satisfies $P(\chi^2_K \le c_\alpha) = \alpha$. The proof of Theorem 4(b) and (c), $\hat{b}$ and $\tilde{b}$ are used instead of b, can be verified as Whang (2003) and using the fact that $\hat{b} = b + O_p(n^{-1/2})$, and $\tilde{b} = b + O(n^{-1/2}) + O(h)$.

국문초록

경험적 우도 방법을 활용한 중도절단회귀모형 추정

문 철

경제학부 경제학 전공

서울대학교 대학원

본 논문에서는 중도절단회귀모형의 모수 추정을 위한 경험적 우도 방법(Whang, 2003)의 유용성을 시뮬레이션을 통하여 검증해 보았다. 우리는 고차 정제를 위하여 비모수 커널 함수를 이용하여 중도절단회귀모형 추정 함수를 평활화하였다. 본 연구에서는 중도절단회귀 추정량과 일차 동등하다고 알려진 평활화된 경험적 우도 추정량으로 구한 신뢰 구간이 포함오차 차수 $O(n^{-1})$를 가짐을 보였다. 몬테 카를로 실험은 바틀렛 보정된 평활화된 경험적 우도 방법이 작은 표본에서 좋은 하였으며, 일반적으로 사용되는 부트스트랩 방법보다 더 정확한 결과를 도출함을 나타낸다. 또한, 시뮬레이션 결과는 평활화된 경험적 우도 방법이 비평활화된 경험적 우도 방법보다 더 나은 결과를 도출한다는 것을 확인하였다. 이는 바틀렛 보정이 평활화된 경험적 우도 방법 신뢰 구간의 포함 오차 차수를 $O(n^{-2})$ 로 줄인다는 Whang (2003)의 이론과 부합하는 결과를 보여준다.

주요어 : 경험적 우도, 중도절단회귀모형, 평활화, 바틀렛 보정

학 번 : 2011-20172