Two-step estimation of semiparametric censored regression models ☆

Shakeeb Khan[a, *], James L. Powell[b]

[a] *Department of Economics, Harkness Hall, PO Box 27056, University of Rochester, Rochester, NY 14627-0156, USA*

[b] *Department of Economics, University of California at Berkeley, Berkeley, CA 94720-3880, USA*

Accepted 24 October 2000

Abstract

Root-*n*-consistent estimators of the regression coefficients in the linear censored regression model under conditional quantile restrictions on the error terms were proposed by Powell (Journal of Econometrics 25 (1984) 303–325, 32 (1986a) 143–155). While those estimators have desirable asymptotic properties under weak regularity conditions, simulation studies have shown these estimators to exhibit a small sample bias in the opposite direction of the least squares bias for censored data. This paper introduces two-step estimators for these models which minimize convex objective functions, and are designed to overcome this finite-sample bias. The paper gives regularity conditions under which the proposed two-step estimators are consistent and asymptotically normal; a Monte Carlo study compares the finite sample behavior of the proposed methods with their one-step counterparts. © 2001 Elsevier Science S.A. All rights reserved.

JEL classification: C24; C14; C13

Keywords: Censored quantile regression; Asymptotic orthogonality; Symmetrically trimmed least squares; Convex processes

☆ This paper was presented at the 1998 North American Summer Meetings of the Econometric Society.

* Corresponding author. Tel.: +1-716-275-5252; fax: +1-716-256-2309.

E-mail address: skhan@troi.cc.rochester.edu (S. Khan).

1. Introduction

The (Type I) censored regression model has received much attention in the theoretical and applied econometric literature.[1] Parametric estimators of this model assume the distribution of the error term to belong to some known parametric family. However, in contrast to the classical linear regression model, misspecification of the distribution of the error term results in the inconsistency of the estimators of the structural parameters.

The semiparametric approach relaxes the assumption of a parametric form for the error term, but imposes sufficient restrictions on its distribution to identify the structural parameters. Restrictions which have been exploited to construct estimators include independence between the errors and the regressors (Horowitz, 1986, 1988; Moon, 1989; Honoré and Powell, 1994) conditional symmetry of the error term (Powell, 1986b), and conditional median and quantile restrictions (Powell, 1984, 1986a; Nawata, 1992; Buchinsky and Hahn, 1998).

The conditional median restriction is the weakest of these, and thus the estimators proposed under this restriction are consistent under the widest class of specifications. In particular, as well as allowing for a wide range of distributions of the error terms, median- and symmetry-based estimators are also robust to very general forms of heteroskedasticity. Assuming a conditional median restriction on the error term, Powell (1984) proposed the censored least absolute deviations (CLAD) estimator, and extended it to the censored quantile regression (CQR) estimator in Powell (1986a) to allow for any quantile restriction on the error term. Under standard regularity conditions, these estimators are $\sqrt{n}$-consistent (with n denoting the sample size) and asymptotically normal, and with a consistent estimator of the limiting covariance matrix, confidence intervals and hypothesis tests could be constructed for large samples.

However, despite these favorable asymptotic properties, there are certain drawbacks regarding the implementation of the CLAD procedure in practice. The first is computational. The CLAD estimator involves the minimization of a non-convex process, and thus iterative linear programming methods (see Buchinsky, 1994 for example) are only guaranteed to converge to a local minimum.

Furthermore, despite the favorable large sample properties of the censored quantile estimators, its finite sample performance has come into question, and has been addressed in simulation studies. Paarsch (1984) compared censored least absolute deviation (CLAD) estimation to normal maximum likelihood and Heckman's 'two-step' least squares estimation. Many stylized facts

[1] See Amemiya (1985) for a list of empirical applications, and Powell (1994) for a survey on recent theoretical developments.

emerged from this study. First, the censored quantile estimator was much more efficient (in a mean-squared sense) than the parametric two-step estimator for several Monte Carlo designs. Second, unless the sample size is fairly large, or the error distribution very non-normal, the inconsistency of the (misspecified) Gaussian maximum likelihood can be small compared to its efficiency advantage (in terms of estimator variance) over the censored LAD. Finally, the censored LAD was found to have a finite sample distribution which was mean biased in the opposite direction of the well known classical least squares bias.

As will be outlined in the following section, this last result, the finite sample bias of the CLAD estimator, which was also found in the simulation study by Moon (1989), is due to an asymmetry in the sampling distribution of the coefficient estimator rather than an actual 'recentering' of its distribution away from the true parameter value. That is, the estimator is nearly median unbiased, but the distribution of the estimator of the slope coefficient is positively skewed (and the intercept estimator has a negatively skewed distribution), so the mean of the sampling distribution of the slope estimator exceeds its median. This asymmetry is less pronounced for designs with less censoring and for certain error distributions which are heavier tailed (more kurtotic) distributions than the Gaussian distribution, but appears to be present to some extent in all of the sampling experiments yet conducted. This asymmetry raises concerns about the accuracy of standard statistical inference procedures based upon an asymptotic normal approximation for the CLAD estimator, at least for small samples.

The present paper attempts to address both the computational difficulties and this finite sample problem without sacrificing the desirable asymptotic properties of the censored quantile regression strategy. A (semiparametric) two-step estimator is proposed in which the first step selects (nonparametrically or semi-parametrically) the observations with a positive value for the regression function, and the second step performs quantile regression on the selected observations.

The paper is organized as follows. In the following section, a numerical example is given to illustrate the reason for the asymmetry in the distribution of the 'one step' censored quantile estimator, and the motivation for the two-step estimator as a means of reducing this asymmetry is outlined. In Section 3, each of the two steps of the proposed estimator is described in detail. Section 4 lists the necessary regularity conditions and discusses the asymptotic properties of the two-step estimator as well as a proposed consistent estimator of the limiting covariance matrix. In Section 5, results of a Monte Carlo study for the two-step estimator are presented. The conclusion (Section 6) summarizes the results and discusses the practical implications of findings in the Monte Carlo study. Appendix A provides the proofs of the main theorems.

2. Rationale for the two-step approach

The (Type I) censored regression model can be written in the form:

$$y_i = \max(x_i'\beta_0 + \varepsilon_i, 0) \tag{2.1}$$

$$= d_i(x_i'\beta_0 + \varepsilon_i), \tag{2.2}$$

where $d_i = I[y_i > 0]$, with $I[\cdot]$ denoting the indicator function. The dependent variable y_i and the k-dimensional regression vector x_i are observed for each i, while the k-dimensional parameter vector β_0 and error term ε_i are unobserved. For this model, the censored LAD estimator, $\hat{\beta}_{\text{CLAD}}$, was defined in Powell (1984) to be the value of β which minimizes:

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \max(0, x_i'\beta)| \tag{2.3}$$

over all β in some parameter space $\mathscr{B}$. This estimation method, based upon the conditional median of y_i, was extended to Powell (1986a) to arbitrary quantiles. Estimators of the coefficients β_0 were defined as minimizers of the function

$$Q_n(\beta;\alpha) = \frac{1}{n}\sum_{i=1}^{n}\rho_\alpha(y_i - \max(0, x_i'\beta)), \tag{2.4}$$

where $\rho_\alpha(\cdot)$ is the 'check function' introduced in Koenker and Bassett (1978),

$$\rho_\alpha(z) = z[\alpha - 1[z < 0]]. \tag{2.5}$$

As discussed in the previous section, Monte Carlo studies of the CLAD and censored regression quantile estimators have indicated a finite sample bias in the means of their distributions. The reason for this bias concerns the interaction of the estimation of β_0 with the 'selection rule' $x_i'\hat{\beta} > 0$, which determines the number of observations entering into the calculation of $\hat{\beta}$. A simple numerical example will illustrate the cause of the asymmetry in the distribution of $\hat{\beta}$. Suppose $n = 4$, $\beta_0 = [0, 1]'$, and the regression vector x_i takes the four values $[1, -2]'$, $[1, -1]'$, $[1, 1]'$, $[1, 2]'$; further, suppose the error terms ε_i have a two-point distribution, taking the values $\frac{1}{2}$ and $-\frac{1}{2}$ with equal probability. Then, for the censored regression model, the vector $y \equiv [y_1, y_2, y_3, y_4]'$ of observed dependent variables will take the values $[0, 0, 0.5, 2.5]'$, $[0, 0, .5, 1.5]'$, $[0, 0, 1.5, 2.5]'$, and $[0, 0, 1.5, 1.5]'$ with equal probability, as illustrated in Fig. 1. When the vector y assumes one of the first three possible values, the censored LAD or symmetrically trimmed regression function will pass through the two data points in the positive quadrant. However, when the last value of y is observed (with $y_3 = y_4 = 1.5$), the fitted regression line will not pass through the points in the positive quadrant (as required for the unbiasedness of the estimated coefficients), since this would

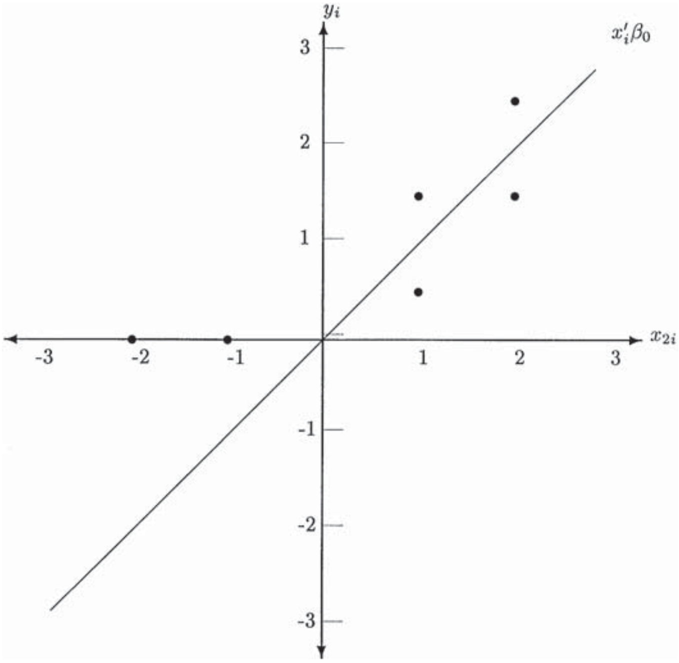

Fig. 1. Distribution of dependent variable in numerical example.

imply large negative residuals for observations 1 and 2 (i.e., for the data points at $(x_{i2}, y_i) = (-2, 0)$ and $(-1, 0)$). In other words, the observations on the $y_i = 0$ axis are 'ignored' (i.e. $\max(0, x_i'\hat{\beta}) = 0$) unless the fitted regression line is unusually flat, in which case those observations cause the fitted line to be steeper than if they were ignored. As a result, the expected value of the CLAD estimator is $[-0.25, 1.125]$ for this design, and the expected value of the symmetrically trimmed least squares estimator is $[-0.21, 1.13]$. In addition, both estimators are median unbiased, indicating an asymmetry in their sampling distributions.

Of course, this bias vanishes in large samples; the probability limit of the symmetrically censored least squares estimator is $\beta_0 = [0, 1]$ when this design is infinitely replicated, and while the censored LAD is not consistent under these conditions (because of the discrete error distribution), it would be consistent if any continuously distributed error term with zero-median and a positive density in a neighborhood of zero was used. Nevertheless, for sufficiently smooth distributions of the regressors, the bias of the two estimators, which is of probability order smaller than $n^{-1/2}$ by the asymptotic theory, can still be evident in moderately sized samples.

The 'two-step' estimator proposed in this paper is meant to correct this finite sample bias while retaining the asymptotic properties of the censored regression quantile estimator. The approach taken is somewhat similar to a parametric two-step estimator (see Heckman, 1976) which estimates the parameter vector (up to scale) using a probit estimator in the first stage. The idea here is to separate the classification of observations into $x_i'\beta_0 > 0$ and $\leqslant 0$ groups from the estimation of the relative magnitudes of the coefficients in β_0. Thus, the proposed estimation proceeds in two steps: a preliminary semiparametric or nonparametric estimator is used to determine which observations have a positive 'index', $x_i'\beta_0$, and standard quantile regression is then applied to those 'selected' observations.

To illustrate why the two-step approach should correct the mentioned bias, consider the case of censored LAD estimation. The CLAD estimator satisfies the asymptotic moment condition

$$\frac{1}{n}\sum_{i=1}^{n} 1\,[x_i'\hat{\beta} > 0]\,\mathrm{sgn}(y_i - x_i'\hat{\beta})x_i = o_\mathrm{p}(n^{-1/2}) \tag{2.6}$$

with the left-hand side being the subgradient of the function $S_n(\beta)$ evaluated at its optimizing value. As the numerical example above indicates, if the true indicator functions were known, the resulting estimator would be unbiased. Instead, the (feasible) two-step estimator would first estimate the indicator variables, only using the values of d_i and x_i. From this procedure, the fitted values, $\tilde{w}_i$ would be used to determine the observations to be included in an LAD regression by estimating the values $I[x_i'\beta_0 > 0]$. The second-step estimator using these fitted values minimizes

$$\tilde{S}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i \rho_\alpha(y_i - x_i'\beta). \tag{2.7}$$

Applied to the previous numerical example, it is clear that this two-step method would yield an unbiased estimator of β_0. The first step would always yield $\tilde{w}_i = 0$ for observations with $x_{i2} \leqslant 0$ and $\tilde{w}_i = 1$ otherwise, so minimization of $\tilde{S}_n(\beta)$ here would always amount to LAD estimation over observations in the positive quadrant. In a more realistic case, with a denser distribution of regressors, it is reasonable to expect that the first step estimates would be less sensitive to regressors with high leverage, since the weighting term $\hat{w}_i$ would only depend upon the sign, and not the magnitude, of the first-step estimator of $x_i'\beta_0$.

3. Description of the proposed estimator

The proposed estimation method will be described in detail in this section. As mentioned above, the approach involves two stages. The first-stage estimation of the indicator $1[x_i'\beta_0 > 0]$ can be conducted in one of various ways, as outlined below. The value of the first-step estimator can then be 'plugged in' the criterion function which is minimized in the second stage.

3.1. First stage estimator

We consider four possible estimators for the first stage. The first two involve estimating the binary choice model under a conditional quantile restriction, where the dependent variable is now classified as either censored ($d_i = 1$) or uncensored. Estimators for the binary choice model under this model which we consider are Manski's 'maximum score' estimator (Manski, 1975, 1985), and the 'smooth maximum score' estimator proposed by Horowitz (1992). The other two estimators we consider for the first stage involve nonparametric methods for estimation of either the conditional censoring probability or the conditional quantile of y_i. Specifically, we consider a nonparametric kernel estimator of the 'propensity score' (Rosenbaum and Rubin, 1983) and the local linear estimator of the conditional quantile function, introduced in Chaudhuri (1991a, b). The relative advantages and disadvantages of these estimators will be discussed after describing each of them in detail.

3.1.1. The maximum score and smooth maximum score estimators

Manski's maximum score estimator can be applied to the censoring indicators, d_i and the regressors x_i in the first stage to estimate the sign of the regression function. The maximum score estimator, denoted here by $\hat{\beta}_{\mathrm{MS}}$, was introduced by Manski (1975, 1985) to estimate the binary choice model under a constant conditional quantile restriction, and can be defined as any minimizing value of

$$M_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(1[d_i = 1] - 1[x_i'\beta > 0]). \tag{3.1}$$

The fact that the maximum score estimator only estimates β_0 up to scale is irrelevant in the context of the proposed estimator, since we are only interested in estimating the sign of the regression function. Furthermore, the relatively strict conditions for consistency of the maximum score estimator are not necessary for consistently estimating the sign of the regression function; in particular it is not required that one of the regressors has a continuous distribution. For the numerical example in the previous section, it is apparent that $1[x_i'\hat{\beta}_{\mathrm{MS}} > 0] \equiv 1[x_i'\beta_0 > 0]$ even though $\hat{\beta}_{\mathrm{MS}}$ is not uniquely determined in large samples.

As an alternative to the maximum score estimator, Horowitz proposed a 'smooth' version, which maximizes the smoothed objective function

$$SM_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\alpha(1[d_i = 1] - K(x_i'\beta/h_n)), \tag{3.2}$$

where $K(\cdot)$ is a smooth function in $[0,1]$ and h_n is a sequence of bandwidths converging to 0 as the sample size increases. As discussed in Horowitz (1992), this approach is computationally simpler than the maximum score estimator, and under stronger conditions than in Manski (1975, 1985), the estimator converges at a faster rate and is asymptotically normally distributed.

3.1.2. Kernel estimation of the propensity score

An alternative approach would be to nonparametrically estimate the 'propensity score', $p(x_i) = \mathrm{E}[d_i|x_i]$. Note the conditional quantile restriction on the error term implies the relationship

$$p(x_i) = \mathrm{P}[\varepsilon_i > -x_i'\beta_0|x_i] > 1-\alpha \Rightarrow x_i'\beta_0 > 0. \tag{3.3}$$

Thus if the propensity score is estimated in the first stage, observations for which its value is greater than $1-\alpha$ will be used in the second stage quantile estimation.

While there exist several methods for estimating a conditional mean function, for the regularity conditions and proofs in this paper, we focus on the Nadaraya–Watson kernel estimator of the propensity score

$$\hat{p}(x) = \frac{\sum_{i=1}^{n} K_h(x_i - x)d_i}{\sum_{i=1}^{n} K_h(x_i - x)}, \tag{3.4}$$

where $K_h(\cdot)$ denotes $h^{-d}K(\cdot/h)$, $K(\cdot)$ is a kernel function, and h is a bandwidth.

3.1.3. Local linear estimation of the conditional quantile function

Another approach would be to nonparametrically estimate the conditional α-quantile function at each observed value of the regressors. Nonparametric estimation of the conditional median function has proven useful in the estimation of several semiparametric models: see, for example Chaudhuri et al. (1997), Khan (2001), Chen and Khan (1998a, b, 2000, 2001). In the context of this model, an equivariance property of quantiles implies that the conditional quantile function, denoted $q_\alpha(\cdot)$, is of the form

$$q_\alpha(x_i) = \max(x_i'\beta_0, 0). \tag{3.5}$$

So, given a nonparametric estimator $\hat{q}_\alpha(\cdot)$ of the conditional quantile, the selection rule would keep the ith observation if $\hat{q}_\alpha(x_i) > 0$.

Several estimators for conditional quantile functions have been recently proposed in the statistics and econometrics literature, notably Stute (1986), Truong (1989), Bhattacharya and Gangopadhyay (1990), Koenker et al. (1992, 1994), and Chaudhuri (1991a, b). For the proof of the asymptotic properties of our proposed two-stage estimator, we use Chaudhuri's local polynomial estimator, noting that it will suffice to implement a local linear estimator in the context of the model we consider.

To illustrate this procedure, let $I[\cdot]$ again denote the indicator function, and let δ_n denote a bandwidth sequence used to smooth the data. A local linear estimator of the conditional quantile function at a point x simply involves quantile regression (see Koenker and Bassett, 1978) on observations which are 'close' to x. Specifically, let $\hat{\theta}_0 \in \mathbf{R}$ and $\hat{\theta}_1 \in \mathbf{R}^k$ minimize the kernel weighted objective function

$$\sum_{i=1}^{n} I[x_i \in C_n(x)]\rho_\alpha(y_i - \theta_0 - \theta_1'(x_i - x)), \tag{3.6}$$

where C_n is a sequence of cubes centered at x whose sides are of length $2\delta_n$, so

$$x_i \in C_n(x) \Leftrightarrow |x_i^{(i)} - x^{(i)}| \leqslant \delta_n, \quad i = 1, 2 \ldots k. \tag{3.7}$$

The conditional quantile estimator which will be used in the first stage will be the value $\hat{\theta}_0 = \hat{\theta}_0(x)$. The motivation for including the linear term in the objective function and estimating the nuisance parameter $(\hat{\theta}_1)$ is to achieve bias reduction of the necessary order for $\sqrt{n}$-consistency of our second stage estimator.

Remark 3.1. Any of these estimators are valid procedures for the first stage, and the asymptotic distribution of the second stage estimator is invariant to the choice of first step procedure. Each of the first stage procedures has its benefits and drawbacks. The Nadaraya–Watson propensity score estimator and the local polynomial estimator are easier to compute than the maximum score and smooth maximum estimators (the first has a closed form solution, and the second minimizes a globally convex objective function) though each must be evaluated n times. The main advantage of using maximum score in the first stage, is that it avoids the problem of determining the value of a smoothing parameter encountered in nonparametric estimation and the smooth maximum score estimator. Also both it and the smooth maximum score estimators sidestep the need to incorporate a 'trimming' function in the objective function of the second stage estimator, as is typically required for two-step estimators with nonparametric estimation in the first stage.

3.2. Second step estimator

Let $\hat{s}_i$ denote the first step estimator for the value of the variable of interest for the ith observation, which we denote by s_i (so $\hat{s}_i = x_i'\hat{\beta}_{\mathrm{MS}}$ if maximum score or smooth maximum score is used in the first stage, $\hat{s}_i = \hat{p}(x_i) - (1-\alpha)$ if the propensity score is estimated, and $\hat{s}_i = \hat{q}_\alpha(x_i)$ if the conditional quantile function is estimated). Now let $w(\cdot)$ denote a smooth 'weighting' function which assigns positive weights to those observations where $\hat{s}_i$ is greater than zero. So $w(\hat{s}_i)$ can be thought of as a smooth approximation to the indicator functions $I[x_i'\hat{\beta}_{\mathrm{MS}} > 0]$, $I[\hat{p}(x_i) > 1-\alpha]$, or $I[\hat{q}_\alpha(x_i) > 0]$, depending on which first stage procedure is used.

The second stage estimator is defined as the minimizer of

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau(x_i) w(\hat{s}_i) \rho_\alpha(y_i - x_i'\beta), \tag{3.8}$$

where $\tau(\cdot)$ is an exogenous 'trimming' function, whose properties are discussed in the next section.[2]

If the propensity score or conditional quantile function is estimated in the first stage, this estimator falls into the class of 'MINPIN' or 'Semiparametric-M' estimators for which many general results have been developed (see, for example, Andrews, 1994a, b; Newey and McFadden, 1994). The most interesting result (which will be shown to hold for this estimator) is that the second step estimator can be $\sqrt{n}$-consistent despite the slower rate of convergence of the first step estimator. It will also be shown that this second step estimator has an 'asymptotic orthogonality' property (see Andrews, 1994a). What this means is that the limiting distribution will be the same as if the true values of the function $w(s_i)$ were used in the second stage objective function instead of their estimated values. Because of this result, it makes no difference (as far as the limiting distribution is concerned) which of the proposed estimators is used in the first step.

We note also that the (second stage) objective function is (globally) convex, which is not the case for the one-step estimator. Thus, any solution found exploiting simplex methods will be a global minimizer, whereas for the CLAD estimator only a local minimum is guaranteed to be found. This uniqueness property (given the first-step estimator) is an additional practical advantage of the two-step approach.

4. Regularity conditions and asymptotics

Before proceeding with the characterization of the asymptotic properties of the proposed estimator, we now list the regularity conditions necessary for $\sqrt{n}$-consistency and asymptotic normality of the second stage estimator. We impose assumptions (which are standard in the literature) on the parameter space, the weighting function, the distribution of the regressors and the error term, and for the case where a nonparametric estimator is used in the first stage, we specify regularity conditions for the bandwidth sequence.

[2] It should be noted that $\tau(\cdot)$ is not required if the maximum score or smooth maximum score estimator is used in the first stage; that is, $\tau(\cdot) = 1$ for the semiparametric first-step estimators.

Full rank condition

FR The matrix

$$J = \mathrm{E}[\tau(x_i)w(s_i)f_{\varepsilon|x_i}(0|x_i)x_i x_i']$$

is of full rank.

Assumption on the parameter space

P1 The true parameter value β_0 is assumed to lie in the interior of $\mathscr{B}$, a convex subset of $\mathbf{R}^k$.

Assumptions on the weighting function

W1 $0 \leqslant w(\cdot) \leqslant 1$.
W2 $w(\cdot) > 0$ if and only if its argument is greater than some fixed constant $c > 0$.
W3 $w(\cdot)$ is twice differentiable, with bounded second derivative.

Assumption on the trimming function

T1 The trimming function $\tau(\cdot) : \mathbf{R}^k \to \mathbf{R}^+$ is bounded and takes the value 0 iff its argument lies outside $\mathscr{X}$, a compact subset of the support of the regressors.

Assumptions on error term and regressors

ER1 The sequence of $k+1$ dimensional vectors (ε_i, x_i) are independent and identically distributed.

ER2 The vector of regressors x_i can be partitioned as $x_i = (x_i^{\mathrm{c}}, x_i^{\mathrm{d}})$ where x_i^{c} has density with respect to Lebesgue measure and x_i^{d} is discretely distributed, with a finite number of mass points on $\mathscr{X}$.

Let $f_{X^{\mathrm{c}}|X^{\mathrm{d}}}(x^{\mathrm{c}}|x^{\mathrm{d}})$ denote the conditional density of x_i^{c} given $x_i^{\mathrm{d}} = x^{\mathrm{d}}$, and $f_{X^{\mathrm{d}}}(x^{\mathrm{d}})$ denote the probability function of x_i^{d}. Also, letting $f_X(x)$ denote $f_{X^{\mathrm{c}}|X^{\mathrm{d}}}(x^{\mathrm{c}}|x^{\mathrm{d}})f_{X^{\mathrm{d}}}(x^{\mathrm{d}})$, we assume for all $x \in \mathscr{X}$:

ER2.1 There exists a strictly positive function $f_0(\cdot)$ such that $f_{X^{\mathrm{c}}|X^{\mathrm{d}}}(x^{\mathrm{c}}|x^{\mathrm{d}}) > f_0(x^{\mathrm{d}}) > 0$.

ER2.2 $f_{X^{\mathrm{c}}|X^{\mathrm{d}}}(x^{\mathrm{c}}|x^{\mathrm{d}})$ and $p(x)f_{X^{\mathrm{c}}|X^{\mathrm{d}}}(x^{\mathrm{c}}|x^{\mathrm{d}})$ are bounded and m times continuously differentiable with bounded derivatives in x^{c}, where m is an even integer such that $m > k_{\mathrm{c}} = \dim(x_i^{\mathrm{c}})$.

ER3 $\varepsilon_i|x_i$ has αth conditional quantile $= 0$.

ER4 The conditional distribution of the latent error terms given the regressors has a density with respect to Lebesgue measure in a neighborhood of 0, denoted by $f_{\varepsilon_i|x_i}(e|x)$ which satisfies the following properties:

ER4.1 As a function of e, it is continuously differentiable and positive for e in a neighborhood of 0, and all x in the support of x_i; and

ER4.2 As a function of x, it is continuous with respect to each component of $x^{(\mathrm{c})}$ for all x in the support of x_i and all e in a neighborhood of 0.

Propensity score first step regularity conditions

PS1 The kernel function $K : R^k \to R$ used in the first step is the product of two separate kernel functions:

$$K(x) = K_{\mathrm{c}}(x^{\mathrm{c}}) * I[x^{\mathrm{d}} = 0].$$

The indicator function serves as the kernel function for the discrete regressors. The kernel function for the continuous regressors, K_{c}, has the following properties:

PS1.1 $\int K_{\mathrm{c}}(u)\,\mathrm{d}u = 1$;

PS1.2 $K_{\mathrm{c}}(\cdot)$ is 0 outside a bounded set.

PS1.3 Let $u = (u_1, u_2, \ldots u_{k_{\mathrm{c}}})$. For all integers $i_1, i_2, \ldots i_{k_{\mathrm{c}}}$ such that $\sum_{j=1}^{k_{\mathrm{c}}} i_j < m$,

$$\int K_{\mathrm{c}}(u) u_1^{i_1} u_2^{i_2} \ldots u_{k_{\mathrm{c}}}^{i_{k_{\mathrm{c}}}}\,\mathrm{d}u = 0.$$

PS2 The bandwidth sequence satisfies the conditions:

PS2.1 $\sqrt{n}(\ln n)h^{2m} \to 0$;

PS2.2 $(\ln n)n^{-1/2}h^{-k_{\mathrm{c}}} \to 0$.

Local linear first step regularity conditions

LL1 The kernel function used for the local linear estimator is the product of k_{c} uniform kernels; specifically, letting $C_n(x_i^{\mathrm{c}})$ denote the cube in $\mathbf{R}^{k_{\mathrm{c}}}$ centered at x_i^{c} with side length $2\delta_n$, the observations in the sample which are used to estimate the conditional quantile function are indexed by the set

$$S_n(x_i) = \{j : 1 \leqslant j \leqslant n, j \neq i, x_j^{(\mathrm{d})} = x_i^{\mathrm{d}}, x_j^{\mathrm{c}} \in C_n(x_i^{\mathrm{c}})\}.$$

LL2 The bandwidth sequence, δ_n is of the form:

$$\delta_n = c_0(\ln n/n)^{-\eta},$$

where c_0 is a positive constant and $0 < \eta < 1/3k_{\mathrm{c}}$.

Remark 4.1. While the list of regularity conditions are quite standard when compared to conditions found in the two-step estimation literature, we comment on some specific conditions which warrant further explanation.

1. The full rank condition imposed in Assumption FR is analogous to the condition needed in Powell (1984). Essentially, it rules out collinearity among the regressors in the support of the weighting and trimming functions.
2. Assumption P1 does not impose compactness on the parameter space. Though compactness is a necessary condition for the consistency of Powell's CLAD estimator, the convexity of our second-step objective function permits us to relax this restriction, as for the estimation method proposed by Buchinsky and Hahn (1998). While the normalization $||\beta|| = 0$ or 1 is imposed for first-step estimators maximizing the score criteria (3.1) or (3.2), this compactness restriction is eliminated in the second step.
3. Condition W2 and W3 are imposed to simplify the derivation of the asymptotic distribution of the second-step estimator $\hat{\beta}$, at the cost of ruling out use of the indicator function ($w(s) \neq I[s > 0]$). Thus, the asymptotic distribution of the two-step estimator will necessarily differ from that for the CLAD (which implicitly uses $w(s) = I[s > 0]$), though, as Buchinsky and Hahn (1998) note, appropriate choices of $w(\cdot)$ and the trimming function $\tau(\cdot)$ will yield a large-sample distribution arbitrarily close to that for CLAD.
4. Assumption T1 serves to exogenously trim the data, which by Assumption ER2.1 enables us to avoid the usual 'denominator' problem (the imprecision of the second-stage estimator when the first-stage estimator of the density function is near 0) encountered with two-step estimators, by bounding the density of the regressors away from zero. A similar approach was adopted by, for example, Buchinsky and Hahn (1998).
5. The strong smoothness assumptions on $p(x_i)$ and $f_{X^{\mathrm{c}}|X^{\mathrm{d}}}(x_i^{\mathrm{c}}|x_i^{\mathrm{d}})$ in Assumption ER2.2 are only required if the propensity score estimator is used in the first stage, making this approach less desirable. They can be weakened to the assumption of continuity if the maximum score or local linear estimator is used in the first stage.
6. Assumption PS1.2 could be relaxed to allow for kernel functions with unbounded supports, and is imposed only to simplify the proofs. Thus the method for constructing 'higher order' kernels described in Bierens (1987) would be valid in this context as well. Alternatively, Müller (1984) discusses how to construct higher order kernels with compact support.
7. The restrictions in Assumption PS2 are satisfied for a wide range of bandwidth sequences. In particular, they are satisfied for the sequence which attains the optimal rate of uniform convergence (in the sense of Stone, 1982) of the first step, $h = O((\ln n/n)^{1/(2m+k_{\mathrm{c}})})$. Similarly, the restrictions in Assumption LL2 also allow for the optimal uniform rate of convergence of the local linear estimator.

Under these conditions, the asymptotic properties of the two-step estimator can be derived. The following theorem, proven in Appendix A, characterizes its limiting distribution, which is virtually identical to that of the CLAD estimator in Powell (1984). The only difference stems from the fact, noted above, that $w(s_i)$ replaces $I[x_i'\beta_0 > 0]$ in our estimation procedure.

Theorem 4.1 (Limiting distribution of the two-step estimator). *Under the conditions imposed above* (*which vary according to the form of the first-step estimator*),

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0, J^{-1} \Lambda J^{-1}), \tag{4.1}$$

where

$$J = \mathrm{E}[\tau(x_i) w(s_i) x_i x_i' f_{\varepsilon|x_i}(0)],$$

$$\Lambda = \mathrm{E}[\tau^2(x_i) w^2(s_i) \alpha(1-\alpha) x_i x_i'].$$

We next consider estimation of the limiting variance of the two-step estimator. We propose consistent estimators for both components, J and Λ. An obvious estimator for the latter term replaces the expectation with a sample average, and the value s_i with its consistent estimator:

$$\hat{\Lambda} = \frac{1}{n} \sum_{i=1}^{n} \alpha(1-\alpha) \tau^2(x_i) w^2(\hat{s}_i) x_i x_i'. \tag{4.2}$$

The 'Hessian term', J, is more difficult to estimate due to the presence of the conditional density term. We adopt the approach taken in Pakes and Pollard (1989), and propose a numerical derivative estimator. The idea behind this approach is that J is the derivative of the function

$$G(\beta) = \mathrm{E}[\tau(x_i) w(s_i) \rho_\alpha'(y_i - x_i'\beta) x_i] \tag{4.3}$$

evaluated at β_0, where $\rho_\alpha'(\cdot)$ denotes the right derivative of the function $\rho_\alpha(\cdot)$. This suggests estimating the ith column of J, denoted J_i by

$$\hat{J}_i = \varepsilon_n^{-1}(\hat{G}_n(\hat{\beta} + \varepsilon_n u_i) - \hat{G}_n(\hat{\beta})), \tag{4.4}$$

where ε_n is a sequence of (possibly random) numbers converging to 0, u_i is a k dimensional vector whose ith element is 1, and whose remaining elements are 0, and where

$$\hat{G}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \tau(x_i) w(\hat{s}_i) \rho_\alpha'(y_i - x_i'\beta) x_i. \tag{4.5}$$

The following theorem, whose proof is also left to Appendix A establishes that the proposed estimators, and hence the estimator $\hat{J}^{-1} \hat{\Lambda} \hat{J}^{-1}$ are consistent:

Theorem 4.2 (Consistency of limiting variance estimator). *Under the conditions imposed above*,

$$\hat{\Lambda} \xrightarrow{\mathrm{p}} \Lambda. \tag{4.6}$$

Furthermore, *if the 'smoothing' parameter* ε_n *satisfies*

$$\varepsilon_n = O_{\mathrm{p}}(n^{-1/4}), \tag{4.7}$$

then

$$\hat{J} \xrightarrow{\mathrm{p}} J. \tag{4.8}$$

5. Monte Carlo results

The preceding results characterize the large-sample equivalence between the one- and two-step censored quantile regression estimators under the stated conditions. However, investigation of the two-step approaches was motivated not by a desire for favorable large sample performance, but rather as a means of attenuating the finite-sample bias observed in the one-step estimators. To determine whether two-step estimation is in fact successful in reducing this bias, Monte Carlo experiments with varying designs were conducted.

Most designs we considered are of the form

$$y_i = \max(\alpha_0 + \beta_0 x_i + \varepsilon_i, 0)$$

with $\beta_0 \in \mathbf{R}$ the parameter of interest, and x_i the single regressor. The intercept α_0 was varied across designs to keep the censoring level constant at 50%. β_0 was set to one for all designs, and the distributions of (x_i, ε_i) varied across designs to explore the relative performance of the various estimators in different settings.

The simulation experiment consisted of 801 replications for sample sizes of 50, 100, 200, and 400. Summary statistics were calculated and reported for six estimators: (1) the CLAD; (2) the two-step LAD with maximum score in the first step; (3) the two-step LAD with propensity score in the first step; (4) the two-step LAD with conditional quantile in the first step; (5) the 'infeasible' LAD, taking $w(s) = I[x_i'\beta_0 > 0]$ and $\tau(\cdot) = 1$; and (6) the Tobit maximum likelihood estimator (assuming normal, homoskedastic errors). In Tables 1–6 these estimators are referred to as CLAD, 2SLADm, 2SLADp, 2SLADq, IFLAD, and MLE, respectively.

While we have not formally verified, either in general or for the specific designs below, that first- and higher order moments exist for any of the estimators, we follow the usual tradition (e.g., Paarsch, 1984; Powell, 1986b; Buchinsky and Hahn, 1998) of reporting estimates of the mean bias and root-mean-squared-error (RMSE) for each design, along with more robust measures of location and dispersion, the median bias and median absolute deviation (MAD). We suspect that, like the numerical example given above, moments for the estimators (at least the quantile-based estimators) should be well behaved, and a comparison of mean to median bias will give a good indication of asymmetry of the sampling distribution of the estimators for finite sample sizes. The fact that, in the tables below, the MAD and RMSE are of comparable orders of magnitude are suggestive that the simulation results are not driven by extreme tail behavior of the estimators.

The simulation study was performed in GAUSS. The CLAD was calculated using the iterative linear programming method discussed in Buchinsky (1994). The second stages of the 2SLAD estimators and IFLAD used the linear programming method discussed in Buchinsky (1994). For 2SLADm, the first stage maximum score estimator was calculated using the Nelder–Meade simplex algorithm. For 2SLADp, the propensity scores were calculated with a Gaussian kernel, and the bandwidth was chosen by cross validation. For 2SLADq, the bandwidth was calculated using the rule of thumb procedure discussed on page 202 of Fan and Gijbels (1996). For each of the 2SLAD estimators, a simple indicator function was used instead of a smooth weighting function, as the latter was mainly adopted to simplify the asymptotic arguments in the proofs. The bounding constant c was set to 0.05 for all 2SLAD estimators.

Tables 1–3 report results for designs similar to those used in Powell (1986b); as mentioned there, certain aspects of these designs correspond to data configurations encountered in practice. The base design, for which results are reported in Table 1, had error terms generated as i.i.d. standard normal, and the regressor values were generated as i.i.d variates uniformly distributed on the interval $[-\sqrt{3}, \sqrt{3}]$ (resulting in a variance of 1). For this design, the CLAD does not exhibit a significant bias, and in fact has a smaller bias than its two-step counterparts, as well as the infeasible LAD estimator. For sample sizes of 50 and 100, the IFLAD estimator exhibits a significant negative bias. As illustrated in Fig. 2, this is to be expected if the regression function takes values near 0 with relatively high probability. This suggests, ironically, that for this particular design, the favorable performance of the CLAD is due to its poor ability to estimate the selection rule. The Tobit MLE estimator, which here correctly specifies the error distribution, performs the best in terms of all summary statistics, as expected.

In Tables 2 and 3, the effects of heteroskedasticity on each of the estimators are investigated. We consider cases where the scale of the error term is a monotonic function of $x_i\beta_0$. Table 2 reports results for the design where the scale of the error term is $ce^{0.75x_i\beta_0}$, where the constant c was set so the average value of the scale function was 1. For this design the results are more in favor of the two-step estimators. The CLAD exhibits a positive mean bias as expected, which is now slightly larger in magnitude than its two-step

Table 1
Simulation results for one- and two-step estimators

Design 1: Scale=1, Censoring=50%				
	Mean bias	RMSE	Med. bias	MAD
50 *obs.*				
CLAD	0.054	0.476	−0.027	0.253
2SLADm	−0.108	0.469	−0.069	0.277
2SLADp	−0.122	0.622	−0.117	0.423
2SLADq	−0.114	0.411	−0.141	0.265
IFLAD	−0.116	0.470	−0.057	0.277
MLE	0.003	0.204	−0.003	0.135
100 *obs.*				
CLAD	0.046	0.422	0.001	0.205
2SLADm	−0.067	0.324	−0.034	0.200
2SLADp	−0.057	0.418	−0.059	0.254
2SLADq	−0.088	0.276	−0.099	0.201
IFLAD	−0.066	0.276	−0.052	0.163
MLE	−0.001	0.146	−0.016	0.093
200*obs.*				
CLAD	−0.006	0.240	−0.038	0.145
2SLADm	−0.046	0.232	−0.020	0.151
2SLADp	−0.035	0.258	−0.040	0.176
2SLADq	−0.046	0.206	−0.045	0.167
IFLAD	−0.074	0.213	−0.064	0.135
MLE	0.004	0.104	0.002	0.066
400 *obs.*				
CLAD	0.010	0.171	−0.010	0.102
2SLADm	−0.035	0.165	−0.012	0.106
2SLADp	−0.021	0.180	−0.022	0.122
2SLADq	−0.029	0.162	−0.033	0.118
IFLAD	−0.040	0.140	−0.035	0.090
MLE	0.003	0.071	0.002	0.048

counterparts. The Tobit MLE, which is misspecified due to the presence of heteroskedasticity, performs very poorly at all sample sizes.

For the decreasing heteroskedasticity design, where the scale of the error term was set to $ce^{-0.75x_i\beta_0}$, the results are reversed, though not nearly as extreme. As reported in Table 3, for small sample sizes, the one-step estimators outperform their two-step counterparts, but the differences are less noticeable than those found in Table 2. One result which remains the same for this design is the poor performance of the MLE.

Tables 4–6 report results for a designs with a different distribution of x_i. Here x_i is distributed as a mixture of two normal distributions. The mixture

Table 2
Simulation results for one- and two-step estimators

Design 2: Scale $= ce^{x_i'\beta_0}$, Censoring = 50%

	Mean bias	RMSE	Med. bias	MAD
50 *obs.*				
CLAD	0.091	0.571	−0.005	0.304
2SLADm	−0.074	0.536	−0.040	0.361
2SLADp	−0.052	0.691	−0.090	0.412
2SLADq	−0.046	0.437	−0.096	0.267
IFLAD	−0.080	0.536	−0.070	0.343
MLE	0.640	0.700	0.627	0.627
100 *obs.*				
CLAD	0.059	0.435	−0.007	0.229
2SLADm	−0.029	0.378	0.000	0.246
2SLADp	−0.019	0.483	−0.028	0.281
2SLADq	−0.040	0.327	−0.064	0.200
IFLAD	−0.061	0.348	−0.058	0.225
MLE	0.630	0.657	0.626	0.626
200 *obs.*				
CLAD	0.029	0.274	0.006	0.167
2SLADm	−0.017	0.264	0.007	0.167
2SLADp	−0.010	0.302	0.002	0.199
2SLADq	−0.079	0.223	−0.081	0.144
IFLAD	−0.033	0.233	−0.010	0.155
MLE	0.623	0.637	0.620	0.620
400 *obs.*				
CLAD	0.013	0.201	−0.015	0.121
2SLADm	−0.024	0.191	−0.010	0.125
2SLADp	−0.004	0.205	−0.004	0.140
2SLADq	−0.058	0.157	−0.062	0.106
IFLAD	−0.035	0.170	−0.032	0.111
MLE	0.625	0.631	0.617	0.617

probability was set to 0.5, and the two normals were centered at 2 and −2, respectively, each with a standard deviation of 0.25. Table 4 reports results for the design where the error term ε_i was distributed as a standard normal. For this design, the benefits of the two-step approach become more apparent. The CLAD exhibits a significant higher mean bias than its two-step counterparts for sample sizes as large as 200.

In Table 5, results are reported for a design where the regressor distribution is the same mixture of normals, but the scale of the error term was set to $ce^{0.75x_i\beta_0}$, where again the constant c was chosen to set the average value of the

Table 3
Simulation results for one- and two-step estimators

Design 3: Scale = $ce^{-0.75x_i'\beta_0}$, Censoring = 50%

	Mean bias	RMSE	Med. bias	MAD
50 obs.				
CLAD	−0.012	0.236	−0.028	0.153
2SLADm	−0.051	0.255	−0.027	0.156
2SLADp	−0.093	0.311	−0.084	0.193
2SLADq	−0.057	0.270	−0.034	0.156
IFLAD	−0.067	0.241	−0.041	0.141
MLE	−0.434	0.446	−0.438	0.438
100 obs.				
CLAD	−0.009	0.164	−0.009	0.112
2SLADm	−0.032	0.172	−0.011	0.117
2SLADp	−0.041	0.191	−0.043	0.121
2SLADq	−0.010	0.184	−0.002	0.116
IFLAD	−0.050	0.169	−0.028	0.098
MLE	−0.432	0.437	−0.434	0.434
200 obs.				
CLAD	0.006	0.121	0.004	0.079
2SLADm	−0.022	0.124	−0.008	0.084
2SLADp	−0.036	0.121	−0.034	0.078
2SLADq	0.004	0.147	0.009	0.103
IFLAD	−0.025	0.103	−0.012	0.060
MLE	−0.439	0.441	−0.438	0.438
400 obs.				
CLAD	−0.004	0.090	0.016	0.060
2SLADm	−0.011	0.001	0.088	0.064
2SLADp	−0.027	0.090	−0.026	0.062
2SLADq	0.002	0.109	0.003	0.074
IFLAD	−0.024	0.076	−0.015	0.047
MLE	−0.434	0.435	−0.432	0.432

scale function to 1. Here the results are dramatically in favor of the two-step estimators. For a sample size of 50, the CLAD exhibits a positive mean bias of 50%, whereas the 2SLAD estimators have biases ranging from 1% to 3%. Furthermore, the two-step estimators outperform the CLAD in terms of mean bias at all sample sizes. As expected, the Tobit MLE performs very poorly due to the presence of heteroskedasticity. Table 6 reports results for the design where the scale function was set to $ce^{-0.75x_i\beta_0}$. Here all estimators (except the MLE) have insignificant mean biases.

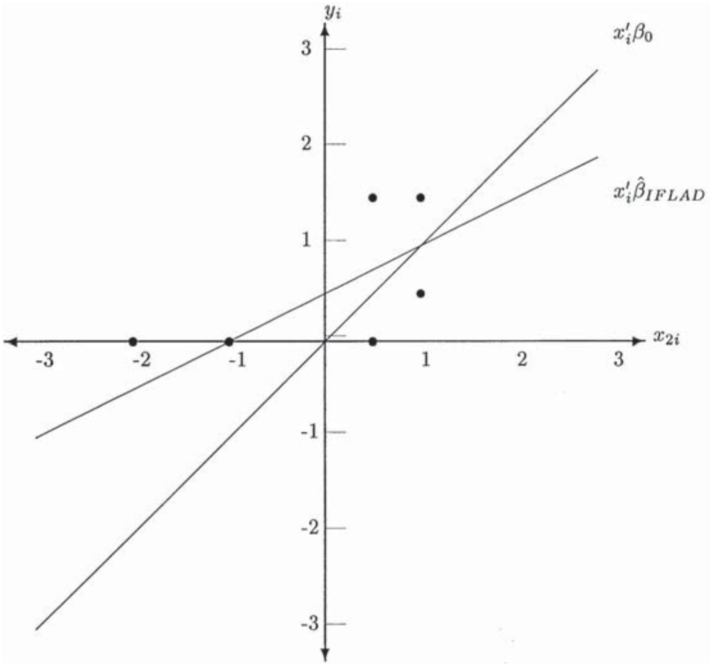

Fig. 2. Design for which the second-step estimator is downward biased.

Finally, to explore the effect of increased dimensionality on the two-step estimators, we simulated from a design with five explanatory variables. Table 7 reports results from a simulation experiment using a design considered in Buchinsky and Hahn (1998). Here, the disturbance term was homoskedastic normal with standard deviation equal to 5, and the explanatory variables were five i.i.d. standard normal random variables. The intercept term and slope coefficients were $1, 1, 0.5, -1, -0.5$, and 0.25, respectively, and the censoring point was -0.75. Sample sizes of 100, 400, and 600 were considered, and the same summary statistics are reported for 10,001 replications. The estimators compared for this design were the IFLAD,CLAD,MLE, the three two-step quantile estimators, and the two-step estimator introduced in Buchinsky and Hahn (1998) with different bandwidth selection methods, referred to in the table as BH1 and BH2.[3]

[3] For their estimator, results are reported for a cross validated (BH1) and adjusted cross validated (BH2) procedure to select the bandwidth. For 2SLADb, a Gaussian kernel was used, and the bandwidth was selected by least squares cross validation. For 2SLADc, the bandwidth was set to $kn^{-1/15}$, with k being the average value of the constant used in the first design.

Table 4
Simulation results for one- and two-step estimators

Design 4: Scale = 1, Censoring = 50%				
	Mean bias	RMSE	Med. bias	MAD
50 obs.				
CLAD	0.238	0.854	−0.054	0.446
2SLADm	−0.002	1.103	−0.054	0.692
2SLADp	0.001	1.104	−0.049	0.692
2SLADq	−0.008	1.099	−0.060	0.688
IFLAD	0.045	1.011	0.087	0.739
MLE	0.182	0.455	0.056	0.156
100 obs.				
CLAD	0.139	0.694	0.011	0.503
2SLADm	0.007	0.726	0.012	0.459
2SLADp	0.028	0.711	0.041	0.476
2SLADq	0.005	0.725	0.001	0.459
IFLAD	0.010	0.740	0.011	0.504
MLE	0.065	0.230	0.003	0.088
200 obs.				
CLAD	0.099	0.440	0.042	0.329
2SLADm	−0.005	0.517	0.002	0.342
2SLADp	0.048	0.510	0.042	0.330
2SLADq	−0.005	0.517	0.002	0.342
IFLAD	0.015	0.505	−0.027	0.339
MLE	0.029	0.135	0.007	0.054
400 obs.				
CLAD	0.020	0.323	0.000	0.232
2SLADm	−0.007	0.355	−0.022	0.242
2SLADp	−0.009	0.354	−0.024	0.241
2SLADq	−0.008	0.354	−0.021	0.242
IFLAD	0.002	0.357	0.000	0.232
MLE	0.006	0.066	−0.001	0.041

For this design, the increased dimensionality adversely affects the two-step estimators the most. 2SLADb and 2SLADc exhibit significant large biases for all sample sizes. In terms of bias, they are outperformed by CLAD, as well as the estimator of Buchinsky and Hahn, but in terms of RMSE and MAD they perform better. Of the three 2SLAD estimators, 2SLADa clearly performs the best, as would be expected when the number of regressors is large, since no nonparametric estimation is involved. For 100 observations, it even exhibits a smaller bias than the CLAD and Buchinsky–Hahn estimator.

Table 5
Simulation results for one- and two-step estimators

Design: Scale $= ce^{0.75x_i'\beta_0}$, Censoring $= 50\%$

	Mean bias	RMSE	Med. bias	MAD
50 obs.				
CLAD	0.518	1.278	0.042	0.347
2SLADm	0.036	1.831	−0.007	0.814
2SLADp	0.017	2.093	−0.067	1.320
2SLADq	0.013	2.077	−0.026	1.305
IFLAD	0.020	1.585	0.003	0.311
MLE	0.993	1.153	0.854	0.854
100 obs.				
CLAD	0.253	0.717	0.015	0.189
2SLADm	0.035	1.030	−0.009	0.190
2SLADp	0.008	1.393	−0.014	0.851
2SLADq	0.011	1.383	−0.004	0.837
IFLAD	0.007	0.758	−0.0114	0.111
MLE	0.864	0.905	0.818	0.818
200 obs.				
CLAD	0.110	0.408	0.002	0.053
2SLADm	0.003	0.521	−0.000	0.085
2SLADp	−0.007	0.985	0.026	0.656
2SLADq	−0.003	0.964	0.022	0.630
IFLAD	−0.009	0.291	−0.010	0.057
MLE	0.845	0.860	0.833	0.833
400 obs.				
CLAD	0.065	0.283	−0.002	0.039
2SLADm	−0.007	0.216	−0.005	0.051
2SLADp	−0.023	0.653	−0.028	0.449
2SLADq	−0.028	0.594	−0.007	0.343
IFLAD	−0.005	0.127	−0.002	0.039
MLE	0.821	0.827	0.813	0.813

In summary, the results of the simulation study are somewhat encouraging for the use of the two-step estimators in practice when the number of regressors are small. For some of the designs considered, they greatly outperform their one-step counterparts in terms of mean bias. For the 2SLAD estimators, it becomes clear that as the number of regressors increases, the estimator which uses maximum score in the first step is more desirable than the other two in terms of all the summary statistics.

Table 6
Simulation results for one- and two-step estimators

Design 6: Scale $= ce^{-0.75x_i'\beta_0}$, Censoring = 50%

	Mean bias	RMSE	Med. bias	MAD
50 obs.				
CLAD	0.004	0.106	−0.009	0.069
2SLADm	−0.005	0.111	−0.009	0.073
2SLADp	−0.029	0.109	−0.033	0.089
2SLADq	−0.005	0.114	−0.011	0.075
IFLAD	−0.002	0.102	0.003	0.067
MLE	−0.958	14.381	−0.725	0.842
100 obs.				
CLAD	0.001	0.070	0.002	0.046
2SLADm	0.002	0.072	0.001	0.046
2SLADp	−0.020	0.069	−0.022	0.056
2SLADq	0.002	0.072	−0.000	0.047
IFLAD	−0.002	0.069	0.001	0.044
MLE	−0.652	0.702	−0.745	0.745
200 obs.				
CLAD	−0.003	0.048	−0.003	0.032
2SLADm	−0.000	0.050	−0.001	0.034
2SLADp	−0.025	0.056	−0.022	0.043
2SLADq	−0.001	0.047	−0.001	0.031
IFLAD	−0.004	0.046	−0.003	0.031
MLE	−0.720	0.727	−0.741	0.741
400 obs.				
CLAD	0.000	0.034	−0.001	0.022
2SLADm	−0.001	0.036	−0.002	0.025
2SLADp	−0.019	0.041	−0.023	0.031
2SLADq	−0.002	0.034	−0.002	0.024
IFLAD	−0.001	0.032	−0.001	0.021
MLE	−0.742	0.743	−0.744	0.744

6. Conclusions

This paper introduces new estimation procedures for semiparametric censored regression models. The procedures involve two estimation stages, and are motivated by their globally convex second-stage objective functions and their potential for improved finite sample properties over existing one-step estimators. The new estimators are shown to have asymptotic properties which

Table 7
Simulation results for one- and two-step estimators

Design 7: Scale = 5, Censoring = 37%				
	Mean bias	RMSE	Med. bias	MAD
100 obs.				
CLAD	0.262	1.190	0.112	0.589
2SLADa	−0.171	0.602	−0.184	0.413
2SLADb	−0.448	0.774	−0.439	0.522
2SLADc	−0.323	0.639	−0.315	0.427
IFLAD	−0.213	0.600	−0.203	0.384
BH1	0.26[a]	2.24[a]	0.05[a]	1.12[a]
BH2	0.21[a]	2.24[a]	0.00[a]	1.09[a]
MLE	−0.038	0.533	−0.063	0.367
400 obs.				
CLAD	0.139	0.524	0.082	0.299
2SLADa	−0.168	0.315	−0.169	0.218
2SLADb	−0.335	0.445	−0.331	0.371
2SLADc	−0.290	0.372	−0.287	0.291
IFLAD	−0.119	0.308	−0.102	0.192
BH1	−0.19[a]	0.78[a]	−0.25[a]	0.52[a]
BH2	−0.24[a]	0.77[a]	−0.30[a]	0.51[a]
MLE	−0.070	0.262	−0.072	0.169
600 obs.				
CLAD	0.098	0.409	0.058	0.239
2SLADa	−0.162	0.268	−0.169	0.194
2SLADb	−0.298	0.382	−0.298	0.301
2SLADc	−0.293	0.348	−0.290	0.291
IFLAD	−0.105	0.252	−0.086	0.158
BH1	−0.14[a]	0.63[a]	−0.20[a]	0.45[a]
BH2	−0.16[a]	0.63[a]	−0.22[a]	0.45[a]
MLE	−0.065	0.217	−0.068	0.146

[a] Values from Table 2 in Buchinsky and Hahn (1998).

are very similar to their one-step counterparts, and the results of a simulation study indicate that they may indeed have finite sample advantages.

Acknowledgements

We are grateful to Moshe Buchinsky, Bo Honoré, Cheng Hsiao, and anonymous referees for helpful comments on earlier drafts.

Appendix A

In arguing the asymptotic properties of the estimator, we adopt shorthand notation for several of the terms. Specifically, we let $w_i, w_i', \hat{w}_i, p_i, \hat{p}_i, q_i, \hat{q}_i, f_i, \hat{f}_i, \tau_i$ denote $w(s_i), w'(s_i), w(\hat{s}_i), p(x_i), \hat{p}(x_i), q_\alpha(x_i), \hat{q}_\alpha(x_i), f_X(x_i), \hat{f}_X(x_i), \tau(x_i)$, respectively. Also, we will let $\rho_\alpha'(\cdot)$ denote the subgradient of the α check function:

$$\rho_\alpha'(z) = \alpha I[z > 0] + (\alpha - 1)I[z \leqslant 0]. \tag{A.1}$$

Finally, for a matrix a, we let $||a||$ denote $(\sum_{i,j} a_{ij}^2)^{1/2}$, where a_{ij} denotes the individual components of a.

Our strategy will be to establish the asymptotic properties of the proposed estimator in three stages. The first stage will establish the consistency of the two-step estimator. The second stage will use the consistency result to establish a higher (though slower than $\sqrt{n}$) rate of convergence for the estimator. This higher rate will then in turn be used to establish $\sqrt{n}$-consistency and asymptotic normality in the third stage.

A.1. Consistency

Consistency of the estimator will follow from standard consistency theorems for minimizers of convex processes. This will require establishing a uniform rate of convergence for the first-stage estimator and pointwise convergence of the second-stage objective to a limiting objective function which is uniquely minimized at β_0.

We first establish uniform rates consistency of the first-step estimators.

Lemma A.1 (Uniform rates for each of the first step estimators). *Under Assumptions* ER1–ER4, PS1–PS2, LL2,

$$\sup_{x_i \in \mathscr{X}} |\hat{s}_i - s_i| = o_\mathrm{p}(n^{-1/4}). \tag{A.2}$$

Proof. For propensity score estimation in the first stage, the kernel regularity conditions satisfy the assumptions of Lemma 8.10 in Newey and McFadden (1994), and so

$$\sup_{x_i \in \mathscr{X}} |\hat{p}_i - p_i| = O_\mathrm{p}((\ln n)^{1/2}(nh^{k_c})^{-1/2} + h^m) \tag{A.3}$$

and thus is $o_\mathrm{p}(n^{-1/4})$ by the Assumptions ER2.2 and PS2.

For either maximum score estimator in the first stage, we appeal to the results of Manski (1985), Cavanagh (1987) and Kim and Pollard (1990), Horowitz (1992) which establish the cube root consistency of the maximum score estimator, and faster rate for the smooth maximum score estimator. The fourth root consistency of the first-step estimator $x_i'\hat{\beta}_{\text{MS}}$ then follows trivially.

For the local linear estimator, recall that c denotes the lower bound of the support of w_i; we show the two results:

$$\sup_{q_i \geqslant c/2, x_i \in \mathcal{X}} |\hat{q}_i - q_i| = o_{\text{p}}(n^{-1/4}) \tag{A.4}$$

and

$$P\left(\sup_{q_i \leqslant c/2, x_i \in \mathcal{X}} |\hat{q}_i - q_i| \geqslant c/2\right) \to 0 \quad \text{at an exponential rate} \tag{A.5}$$

From Lemma 4.3a in Chaudhuri et al. (1997) we have

$$\sup_{q_i \geqslant c/2, x_i \in \mathcal{X}} |\hat{q}_i - q_i| = O_{\text{p}}\left(\sqrt{\frac{\log n}{n\delta_n^{k_c}}}\right) \tag{A.6}$$

so the first result follows from Assumption LL2. The second result for the local linear estimator has been proved in Lemma 2 in Chen and Khan (2000). □

We next show the limiting second-stage objective function is uniquely minimized at β_0.

Lemma A.2 (Identification). *Under Assumptions* W2, FR, ER3, ER4.1 *the limiting objective function*,

$$\text{E}[\tau_i w_i \rho_\alpha(y_i - x_i'\beta)] \tag{A.7}$$

is uniquely minimized at β_0.

Proof. If $w_i > 0$, since the expected α 'check function' is minimized at its conditional αth quantile, and the conditional quantile of ε_i is uniquely 0 by Assumptions ER3, ER4.1, we have that the function

$$\text{E}[\tau_i w_i \rho_\alpha(y_i - x_i'\beta) \,|\, x_i, w_i > 0] \tag{A.8}$$

is minimized at $x_i'\beta_0$. For β_0 to uniquely minimize the objective function, we require both $\text{P}(w_i > 0) > 0$ and, for $\beta \neq \beta_0$,

$$\text{P}(x_i'\beta_0 \neq x_i'\beta | w_i > 0) > 0. \tag{A.9}$$

Both these conditions follow by Assumption FR. □

We can now show consistency under the assumptions needed for the previous two lemmas, and Assumptions P1,W3:

Theorem A.1 (consistency).

$$\hat{\beta} \xrightarrow{\mathrm{p}} \beta_0. \tag{A.10}$$

Proof. A mean value expansion of $\hat{w}_i$ around w_i, the uniform consistency of the first-step estimator, and the LLN imply:

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i \hat{w}_i \rho_\alpha(y_i - x_i'\beta) \xrightarrow{\mathrm{p}} \mathrm{E}[\tau_i w_i \rho_\alpha(y_i - x_i'\beta)] \tag{A.11}$$

for each β. This limiting function is uniquely minimized at β_0 by Lemma A.2. The theorem thus follows from a standard consistency theorem for convex minimizers (see, for example, Theorem 2.7 in Newey and McFadden, 1994). □

A.2. 4th-root consistency

The consistency of the estimator can be used to establish a higher rate of convergence. Here, we show 4th-root consistency, which will be used in the next section to establish $\sqrt{n}$-consistency and asymptotic normality. We first use consistency to establish the following equicontinuity condition:

Lemma A.3. Assume Assumptions P1, W1–W2, T1, ER1, ER4.1 *hold. For each* $\varepsilon > 0$ *and* $\eta > 0$ *there exists a* $\delta > 0$ *such that*

$$\limsup \mathrm{P}\left(\sup_{||\beta-\beta_0||<\delta} \left|\left| \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau_i w_i(\rho_\alpha'(y_i - x_i'\beta_0) - \rho_\alpha'(y_i - x_i'\beta))x_i \right.\right.\right.$$
$$\left.\left. - \mathrm{E}[\tau_i w_i \rho_\alpha'(y_i - x_i'\beta)x_i]\right|\right| > \eta \Bigg) < \varepsilon. \tag{A.12}$$

Proof. We first prove the two following lemmas, which establish, respectively, that the subgradient belongs to a Euclidean class, and that it is $\mathscr{L}^2$ continuous at β_0. □

Lemma A.4 (Euclidean property). *Under Assumptions* W1, T1, *the functional space*

$$\{\tau_i w_i \rho_\alpha'(y_i - x_i'\beta)x_i : \beta \in \mathscr{B}\} \tag{A.13}$$

is Euclidean with respect to a constant envelope.

Proof. Clearly, for all β, the function $\tau_i w_i \rho_\alpha'(y_i - x_i'\beta)x_i$ is bounded since $\tau(\cdot) = 0$ if x_i lies outside a compact set. Also, the class of functions $\rho_\alpha'(y_i - x_i'\beta)$, being piece-wise linear, is Euclidean with constant envelope by Example 2.12 in Pakes and Pollard (1989). The other component $\tau_i w_i x_i$ is trivially Euclidean, so the result follows by Lemma 2.14 (ii) of Pakes and Pollard. □

Lemma A.5 ($\mathscr{L}^2$ continuity at β_0). *Under Assumptions* P1, W1, W2, T1, ER4.1, *if* $\beta \to \beta_0$,

$$E[||\tau_i w_i(\rho'_\alpha(y_i - x'_i\beta) - \rho'_\alpha(y_i - x'_i\beta_0))x_i||^2] \to 0. \tag{A.14}$$

Proof. Noting that $\tau_i w_i x_i$ is bounded, it will suffice to show that as $\beta \to \beta_0$,

$$E[I[\tau_i w_i > 0](\rho'_\alpha(y_i - x'_i\beta) - \rho'_\alpha(y_i - x'_i\beta_0))^2] \to 0. \tag{A.15}$$

Multiplying the expression inside the above expectation by $I[y_i=0]+I[y_i > 0]$, it will suffice to show

$$E[I[\tau_i, w_i > 0](\rho_\alpha(-x'_i\beta) - \rho_\alpha(-x'_i\beta_0))^2] \to 0 \tag{A.16}$$

and

$$E[I[\tau_i, w_i > 0](\rho_\alpha(\varepsilon_i + x'_i(\beta - \beta_0)) - \rho_\alpha(\varepsilon_i))^2] \to 0. \tag{A.17}$$

To show (A.16), note that $I[w_i > 0]\rho_\alpha(-x'_i\beta_0) = (\alpha - 1)$, and that for all $x_i \in \mathscr{X}$, since $w_i > 0$ implies that $x'_i\beta_0 \geqslant c^*$ for a small constant c^*, $I[w_i > 0, -x'_i\beta > 0] \to 0$ and $I[w_i > 0, -x'_i\beta \leqslant 0] \to 1$, so $I[w_i > 0](\rho_\alpha(-x'_i\beta) - \rho_\alpha(-x'_i\beta_0)) \to 0$ for each $x_i \in \mathscr{X}$. (A.16) follows from the dominated convergence theorem. To show (A.17), note that $|\rho_\alpha(\varepsilon_i + x'_i(\beta - \beta_0)) - \rho_\alpha(\varepsilon_i))|$ is bounded above by $2I[|\varepsilon_i| \leqslant ||x_i||\,||\beta - \beta_0||]$. Thus by Assumption ER4.1, and the fact that $||x_i||$ is bounded in the support of τ_i, shows that

$$E[\tau_i w_i|\rho_\alpha(\varepsilon_i + x'_i(\beta - \beta_0)) - \rho_\alpha(\varepsilon_i))|] = O(||\beta - \beta_0||). \tag{A.18}$$

This shows (A.17), so we have shown (A.15).

Lemma A.3 now follows directly by Lemma 2.17 of Pakes and Pollard (1989), which also requires Assumption ER1. □

We can now establish the following linear representation, which relates the rate of convergence of the estimator to the rate of convergence of the 'infeasible' asymptotic first order condition,

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i. \tag{A.19}$$

Lemma A.6 (Linear representation). *Under Assumptions* W1, W2, T1, ER1, ER3, ER4, *for* $\gamma \in (0, 1/2]$, *if*

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i = o_p(n^{-\gamma}), \tag{A.20}$$

then

$$\hat{\beta} - \beta_0 = J^{-1}\frac{1}{n}\sum_{i=1}^{n}\tau_i w_i \rho'_\alpha(\varepsilon_i)x_i + o_{\mathrm{p}}(n^{-\gamma}), \tag{A.21}$$

where $J = \mathrm{E}[\tau_i w_i f_{\varepsilon_i|x_i}(0)x_i x_i']$.

Proof. We first establish the following result concerning the derivative of the expected value of the first order condition:

Lemma A.7. Under Assumptions W1, W2, T1, ER3, ER4.1, *if* $\hat{\beta}\xrightarrow{\mathrm{p}}\beta_0$,

$$\frac{\partial}{\partial\beta}\mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x_i'\hat{\beta})x_i]\xrightarrow{\mathrm{p}} - J. \tag{A.22}$$

Proof. Note that for either of the first-step estimators, $w_i > 0$ implies there exists a small constant c^* such that $x_i'\beta_0 \geqslant c^*$. So for β sufficiently close to β_0, we have

$$\mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x_i'\beta)x_i]$$

$$= \mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x_i'\beta)x_i(I[y_{=}0] + I[y_i > 0])] \tag{A.23}$$

$$= \mathrm{E}[\tau_i w_i \rho'_\alpha(\varepsilon_i)I[y_i > 0]x_i] + (\alpha - 1)\mathrm{E}[\tau_i w_i x_i] \tag{A.24}$$

So to evaluate the derivative with respect to β, one only need do so for the left-hand side of the above equation. Note we have by the law of iterated expectations:

$$\mathrm{E}[\tau_i w_i \rho'_\alpha(\varepsilon_i)I[y_i > 0]x_i]$$

$$= \alpha\mathrm{E}[\tau_i w_i(1 - F_{\varepsilon\,|\,X}(x_i'(\beta - \beta_0))x_i] \tag{A.25}$$

$$+ (\alpha - 1)\mathrm{E}[\tau_i w_i(F(x_i'(\beta - \beta_0) - F(-x_i'\beta_0))x_i]. \tag{A.26}$$

So by the dominated convergence theorem we have

$$\frac{\partial}{\partial\beta}\mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x_i'\beta)x_i]$$

$$= - \alpha\mathrm{E}[\tau_i w_i f_{\varepsilon X}(x_i'(\beta - \beta_0))x_i x_i'] \tag{A.27}$$

$$+ (\alpha - 1)\mathrm{E}[\tau_i w_i f_{\varepsilon X}(x_i'(\beta - \beta_0))x_i x_i'] \tag{A.28}$$

$$= - \mathrm{E}[\tau_i w_i f_{\varepsilon X}(x_i'(\beta - \beta_0))x_i x_i'], \tag{A.29}$$

which is continuous at $\beta = \beta_0$ by Assumption ER4.1 and the dominated convergence theorem. Thus by the consistency of $\hat{\beta}$, (A.22) follows from the continuous mapping theorem. □

Now, by expanding the moment condition

$$\mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x'_i\beta_0)] = 0 \tag{A.30}$$

around $\hat{\beta}$, by the previous lemma we get

$$\hat{\beta} - \beta_0 = (-J^{-1} + o_{\mathrm{p}}(1))\mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i]. \tag{A.31}$$

We next decompose $-\mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i]$ into

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\tau_i w_i(\rho'_\alpha(y_i - x'_i\hat{\beta}))) - \rho'_\alpha(y_i - x'_i\beta_0))x_i - \mathrm{E}[\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i]\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\tau_i w_i \rho'_\alpha(y_i - x'_i\beta_0)x_i\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i.\end{aligned}$$

The conclusion of Lemma A.6 now follows since the first term in the decomposition is $o_{\mathrm{p}}(n^{-1/2})$ by the consistency of $\hat{\beta}$ and Lemma A.3, the second term is $O_{\mathrm{p}}(n^{-1/2})$ by an ordinary central limit theorem, and the third term is $o_{\mathrm{p}}(n^{-\gamma})$ by assumption. □

An immediate consequence of this representation is the faster rate of convergence for $\hat{\beta}$.

Theorem A.2 (4th root consistency). *Under Assumptions* FR, P1, W1–W3, T1, ER1–ER4, PS1–PS2, LL2,

$$\hat{\beta} - \beta_0 = o_{\mathrm{p}}(n^{-1/4}). \tag{A.32}$$

Proof. A mean value expansion of $\hat{w}_i$ around w_i in the asymptotic first order condition implies:

$$o_{\mathrm{p}}(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n}\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i + R_n. \tag{A.33}$$

By Lemma A.1, $R_n = o_{\mathrm{p}}(n^{-1/4})$, so by Lemma A.6 and an ordinary central limit theorem we have

$$\begin{aligned}\hat{\beta} - \beta_0 &= O_{\mathrm{p}}(n^{-1/2}) + o_{\mathrm{p}}(n^{-1/4})\\ &= o_{\mathrm{p}}(n^{-1/4}) \quad \square\end{aligned}$$

A.3. $\sqrt{n}$-consistency and asymptotic normality

We can now proceed to the main theorem, characterizing the limiting distribution of the estimator. Two preliminary lemmas are required. The first shows that an 'interaction' term is asymptotically negligible, and the second lemma establishes that the first stage of the estimator does not affect the asymptotic variance of the second-stage estimator.

Lemma A.8. Under Assumptions P1, T1, W2, W3, ER1–ER4, PS1–PS2, LL2,

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i(\hat{w}_i - w_i)(\rho'_\alpha(y_i - x_i'\beta_0) - \rho'_\alpha(y_i - x_i'\hat{\beta}))x_i = o_p(n^{-1/2}). \tag{A.34}$$

Proof. Note Lemma A.1 and a mean value expansion of $\hat{w}_i$ around w_i it will suffice to show

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i w_i'|\rho'_\alpha(y_i - x_i'\beta_0) - \rho'_\alpha(y_i - x_i'\hat{\beta})|\,||x_i|| = o_p(n^{-1/4}). \tag{A.35}$$

Decompose the sum into

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i w_i'|\rho'_\alpha(y_i - x_i'\beta_0) - \rho'_\alpha(y_i - x_i'\hat{\beta})|\,||x_i||I[y_i = 0] \tag{A.36}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\tau_i w_i'|\rho'_\alpha(y_i - x_i'\beta_0) - \rho'_\alpha(y_i - x_i'\hat{\beta})|\,||x_i||I[y_i > 0]. \tag{A.37}$$

To show (A.37) is $o_p(n^{-1/4})$, we note that regardless of which estimator is used in the first stage, by Assumption W2, we can find a small positive constant c^*, such that $w_i > 0$ implies $x_i'\beta_0 \geqslant c^*$. Let K_1 denote $\sup_{x_i \in \mathscr{X}}||x_i||$ and let $\varepsilon^* = c^*/K_1$. By the bounds on $w_i', \tau_i x_i$, it follows that the matrix norm of (A.37) is bounded above by constant times

$$I[||\hat{\beta} - \beta_0|| \geqslant \varepsilon^*]$$

so (A.37) is $o_p(n^{-1/4})$ by Theorem A.2.

To show (A.38) is $o_p(n^{-1/4})$, by the bounds on $w_i', \tau_i x_i$ it will suffice to show that

$$\sum_{i=1}^{n} I[|\varepsilon_i| \leqslant K_1||\hat{\beta} - \beta_0||] = o_p(n^{-1/4}). \tag{A.38}$$

By Assumption ER4.1, $\mathrm{E}[I[|\varepsilon_i| \leqslant K_1||\beta - \beta_0||]$, evaluated at $\beta = \hat{\beta}$, is $o_p(n^{-1/4})$ by Theorem A.2. So letting γ_n denote a sequence of numbers converging slowly enough to 0, by Theorem A.2, it remains to show

$$\sup_{||\beta-\beta_0|| \leqslant \gamma_n} \sum_{i=1}^{n} I[|\varepsilon_i| \leqslant K_1||\beta - \beta_0||] - \mathrm{E}[I[|\varepsilon_i| \leqslant K_1||\beta - \beta_0||]]$$

$$= o_p(n^{-1/2}). \tag{A.39}$$

This follows by Lemma 2.17 of Pakes and Pollard, as the class of functions is Euclidean for the envelope $F \equiv 1$ by Example 2.12 in Pakes and Pollard

(1989), and $\mathscr{L}^2$ continuity follows by the same argument used in Lemma A.5. This shows (A.37) and hence (A.34). □

Lemma A.9. Under Assumptions W2–W3, T1, ER1–ER4, PS1–PS2, LL2,

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i(\hat{w}_i - w_i)\rho'_\alpha(y_i - x'_i\beta_0)x_i = o_{\mathrm{p}}(n^{-1/2}). \tag{A.40}$$

Proof with propensity score estimated in the first step. Let $z_i = \rho'_\alpha(y_i - x'_i\beta_0)$. The propensity score can be decomposed as

$$p_i = \frac{p_i^\dagger}{f_i}, \tag{A.41}$$

where $p_i^\dagger = p_i f_i$. This will be analytically convenient because the Nadaraya–Watson kernel estimator is used in the first step. A mean value expansion gives us

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i(\hat{w}_i - w_i)z_i x_i = \frac{1}{n}\sum_{i=1}^{n}\tau_i w'_i(\hat{p}_i - p_i)z_i x_i + r_n, \tag{A.42}$$

where $r_n = o_{\mathrm{p}}(n^{-1/2})$ by Lemma A.1. The first term on the right-hand side can be decomposed by linearizing the difference $\hat{p}_i - p_i$ in terms of its numerator and denominator:

$$\frac{1}{n}\sum_{i=1}^{n}\tau_i w'_i(\hat{p}_i - p_i)z_i x_i = R_1 + R_2 + R_3 + R_4, \tag{A.43}$$

$$R_1 = \frac{1}{n}\sum_{i=1}^{n}\tau_i w'_i z_i x_i f_i^{-1}(\hat{p}_i^\dagger - p_i^\dagger), \tag{A.44}$$

$$R_2 = \frac{1}{n}\sum_{i=1}^{n}\tau_i w'\, p_i z_i x_i\, p_i^\dagger f_i^{-2}(\hat{f}_i - f_i), \tag{A.45}$$

$$R_3 = \frac{1}{n}\sum_{i=1}^{n}\tau_i w'_i z_i x_i f_i^{-1}\hat{f}_i^{-1}(\hat{f}_i - f_i)(\hat{p}_i^\dagger - p_i^\dagger), \tag{A.46}$$

$$R_4 = \frac{1}{n}\sum_{i=1}^{n}\tau_i w'_i z_i x_i f_i^{-2}\hat{f}_i^{-1}(\hat{f}_i - f_i)^2. \tag{A.47}$$

The uniform consistency result in Lemma A.1, and the result that $\sup_{x_i\in\mathscr{X}}|\hat{f}_i| = O_{\mathrm{p}}(1)$ imply that R_3, R_4 are $o_{\mathrm{p}}(n^{-1/2})$. R_1 can be decomposed as follows:

$$R_1 = R_{11} + R_{12},$$

$$R_{11} = \frac{1}{n}\sum_{i=1}^{n}\tau_i w'_i z_i x_i f_i^{-1}(\hat{p}_i^\dagger - \bar{p}_i^\dagger),$$

$$R_{12} = \frac{1}{n}\sum_{i=1}^{n}\tau_i w'_i z_i x_i f_i^{-1}(\bar{p}_i^\dagger - p_i^\dagger),$$

where $\bar{p}_i^\dagger = \mathrm{E}[\hat{p}_i^\dagger]$ Now, plugging in the form of the Nadaraya–Watson estimator (which leaves out the own observation term), into R_{11} yields the following U-statistic:

$$R_{11} = \frac{1}{n(n-1)} \sum_{i \neq j} \tau_i w_i' z_i x_i f_i - 1(d_j K_h(x_i - x_j) - \bar{p}_i^\dagger). \tag{A.48}$$

We denote the 'kernel' of this U-statistic by $\mathscr{F}_n(\xi_i, \xi_j)$, where $\xi_i = (d_i, x_i', \varepsilon_i)'$. We note by a change of variables and the bounds on $\tau_i x_i$, w_i', z_i, f_i^{-1}, K that $\mathrm{E}[||\mathscr{F}_n(\xi_i, \xi_j)||^2 = O(h^{-k_c})$ which is $o(n)$ by Assumption PS2.2. Also, by the quantile restriction, we have $\mathrm{E}[\mathscr{F}_n(\xi_i, \xi_j)|\xi_j] = 0$, and by the definition of $\bar{p}_i^\dagger$, we have $\mathrm{E}[\mathscr{F}_n(\xi_i, \xi_j)|\xi_i] = 0$. Thus by Lemma 3.1 in Powell et al. (1989), (see also Ahn and Powell, 1993), $R_{11} = o_\mathrm{p}(n^{-1/2})$. Turning attention to R_{12}, we note that by the quantile restriction, its expected value is 0. Furthermore, by the bounds on $\tau_i x_i$, w_i', z_i, f_i^{-1} the term in the summation of R_{12} is bounded above by a constant times $||\bar{p}_i^\dagger - p_i^\dagger||$. Next noting that $p_i^\dagger$ is bounded on $\mathscr{X}$, the dominated convergence theorem implies by Assumption PS2.1 that $\bar{p}_i^\dagger \to p_i^\dagger$ for all $x_i \in \mathscr{X}$. Also, by the bound on K, another application of the dominated convergence theorem implies by Assumption PS2.1 that

$$\mathrm{E}[||\bar{p}_i^\dagger - p_i^\dagger||^2] \to 0. \tag{A.49}$$

Therefore, $n^{1/2} R_{12} = o_\mathrm{p}(1)$ by Chebyshev's inequality, since both its mean and variance converge to 0. This establishes that $R_1 = o_\mathrm{p}(n^{-1/2})$. The same arguments can be used to show that $R_2 = o_\mathrm{p}(n^{-1/2})$. □

Proof with maximum score used in the first step. Let $\tilde{\beta}$ denote either the maximum score or smooth maximum score estimator, and let $\beta_0^\dagger$ denote β_0 after a suitable scale normalization. It needs to be established that

$$\frac{1}{n} \sum_{i=1}^{n} \tau_i w_i' x_i' (\tilde{\beta} - \beta_0^\dagger) \rho_\alpha'(\varepsilon_i) x_i = o_\mathrm{p}(1/\sqrt{n}). \tag{A.50}$$

By the cube-root consistency of either maximum score estimator, (Manski, 1985; Cavanagh, 1987; Kim and Pollard, 1990; Horowitz, 1992), for any $\delta > 0$, let γ_n be a sequence of numbers which is $O(n^{-1/4})$; we have

$$\begin{aligned} &\mathrm{P}\left(\left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tau_i w_i' \rho_\alpha'(\varepsilon_i) x_i' (\tilde{\beta} - \beta_0^\dagger) x_i \right|\right| > \delta \right) \\ &\quad = \mathrm{P}\left(\left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tau_i w_i' \rho_\alpha'(\varepsilon_i) x_i' (\tilde{\beta} - \beta_0^\dagger) x_i \right|\right| > \delta, \right. \\ &\qquad \left. ||\tilde{\beta} - \beta_0^\dagger|| < \gamma_n \text{ or } ||\tilde{\beta} - \beta_0^\dagger|| \geqslant \gamma_n \right) \end{aligned} \tag{A.51}$$

$$\leqslant \mathrm{P}\left(\sup_{||\beta^\dagger-\beta_0^\dagger||\leqslant\gamma_n}\left|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \tau_i w_i' \rho_\alpha'(\varepsilon_i)x_i'(\beta^\dagger-\beta_0^\dagger)x_i\right|\right| > \delta\right)$$

$$+\mathrm{P}(||\tilde{\beta}-\beta_0^\dagger|| \geqslant \gamma_n). \tag{A.52}$$

The second probability in the above summation converges to 0 by the cube root consistency of the maximum score estimator. To show the first probability can be made arbitrarily small, we first note that

$$\mathrm{E}[||\tau_i w_i' \rho_\alpha'(\varepsilon_i)x_i'(\beta^\dagger-\beta_0^\dagger)x_i||^2] \leqslant K||\beta-\beta_0||, \tag{A.53}$$

where K is a constant, so the left-hand side of the above inequality converges to 0 as $\beta \to \beta_0$. Furthermore, the class of functions $(\tau_i w_i' \rho_\alpha'(\varepsilon_i)x_i'\beta^\dagger x_i\colon \beta^\dagger \in \mathscr{B}^\dagger)$ is Euclidean for a constant envelope by Example 2.12 in Pakes and Pollard. Thus (A.52) goes to 0 by Lemma 2.17 in Pakes and Pollard (1989). □

Proof with local linear estimation in first step. We first 'plug in' the local Bahadur representation for the local linear estimator established in Chaudhuri (1991a), Chaudhuri et al. (1997). Following the same steps used in the proof of Lemma 5 of Chen and Khan (2000), we can express this term as a second order U-statistic plus an asymptotically negligible remainder term:

$$\frac{1}{n(n-1)}\sum_{i\neq j}\tau_i w_i' f_{\varepsilon,X}(0,x_i)^{-1}\delta_n^{-k_c} \tag{A.54}$$

$$(I[y_j \leqslant q_j]-\alpha)I[j \in S_n(x_i)]\rho_\alpha'(\varepsilon_i)x_i + o_\mathrm{p}(n^{-1/2}). \tag{A.55}$$

Let $\mathscr{F}_n(\xi_i,\xi_j)$ denote the 'kernel' of this U-statistic, where $\xi_i \equiv (y_i, x_i')'$. Note that

$$\mathrm{E}[\mathscr{F}_n(\xi_i,\xi_j)\,|\,\xi_i] = \mathrm{E}[\mathscr{F}_n(\xi_i,\xi_j\,|\,\xi_j] = \mathrm{E}[\mathscr{F}_n(\xi_i,\xi_j)] = 0. \tag{A.56}$$

Also, by the proof in Lemma 5 in Chen and Khan (2000), we have

$$\mathrm{E}[||\mathscr{F}_n(\xi_i,\xi_j)||^2] = O(\delta_n^{-k_c}) = o(n), \tag{A.57}$$

where the second equality follows from Assumption LL2. So by Lemma 3.1 in Powell et al. (1989), this U-statistic is $o_\mathrm{p}(n^{-1/2})$. □

We can now prove Theorem 4.1 under Assumptions FR, P1, W1–W3, T1, ER1–ER4, PS1–PS2, LL2. We decompose the asymptotic first order condition,

$$o_\mathrm{p}(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \tau_i \hat{w}_i \rho_\alpha'(y_i - x_i'\hat{\beta})x_i \tag{A.58}$$

in the following manner:

$$\begin{aligned} o_{\mathrm{p}}(n^{-1/2}) = {} & \frac{1}{n}\sum_{i=1}^{n}\tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i \\ & + \frac{1}{n}\sum_{i=1}^{n}\tau_i(\hat{w}_i - w_i)\rho'_\alpha(y_i - x'_i\beta_0)x_i \\ & + \frac{1}{n}\sum_{i=1}^{n}\tau_i(w_i - \hat{w}_i)(\rho'_\alpha(y_i - x'_i\beta_0) - \rho'_\alpha(y_i - x'_i\hat{\beta}))x_i. \end{aligned}$$

Thus by Lemmas A.8 and A.9 we have the asymptotically equivalent first order condition:

$$o_{\mathrm{p}}(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n} \tau_i w_i \rho'_\alpha(y_i - x'_i\hat{\beta})x_i. \tag{A.59}$$

By Lemma A.6, Theorem 1 easily follows from the Lindeberg–Levy central limit theorem. □

A.4. Consistent asymptotic variance estimator

We now prove Theorem 4.2. Note that by Lemma A.1, we have

$$\hat{\Lambda} = \frac{1}{n}\sum_{i=1}^{n} (\alpha(1-\alpha))\tau^2(x_i)w^2(s_i)x_i x'_i + o_{\mathrm{p}}(n^{-1/4}) \tag{A.60}$$

and so consistency of $\hat{\Lambda}$ then follows by the law of large numbers. For establishing the consistency of $\hat{J}$, let

$$G_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau_i w_i \rho'_\alpha(y_i - x'_i\beta)x_i. \tag{A.61}$$

It follows by Lemma A.1 that

$$||\hat{J}_i - \varepsilon_i^{-1}(G_n(\hat{\beta} + \varepsilon_n u_i)) - G_n(\hat{\beta}))|| = \varepsilon_n^{-1} o_{\mathrm{p}}(n^{-1/4}) \tag{A.62}$$

and is thus negligible by the assumption on ε_n. We next show

$$||G_n(\hat{\beta} + \varepsilon_n u_i) - G(\hat{\beta} + \varepsilon_n u_i) - G_n(\beta_0)|| = o_{\mathrm{p}}(n^{-1/2}) \tag{A.63}$$

and

$$||G_n(\hat{\beta}) - G(\hat{\beta}) - G_n(\beta_0)|| = o_{\mathrm{p}}(n^{-1/2}). \tag{A.64}$$

To show (A.64), we note by the $\sqrt{n}$-consistency of $\hat{\beta}$, and the fact that $\varepsilon_n = O_{\mathrm{p}}(n^{-1/4})$ that it will suffice to show that

$$\sup_{||\beta - \beta_0|| \leqslant n^{-1/4}} (G_n(\beta) - G(\beta) - G_n(\beta_0)) = o_{\mathrm{p}}(n^{-1/2}). \tag{A.65}$$

This follows by Lemma 2.17 in Pakes and Pollard (1989), as the Euclidean property for a constant envelope follows from Lemma A.4 and $\mathscr{L}^2$ continuity follows from Lemma A.5. The same argument can be used to show (A.64). Thus, we have

$$\begin{aligned}||\varepsilon_i^{-1}(G_n(\hat\beta+\varepsilon_n u_i)-G(\hat\beta+\varepsilon_n u_i)+G_n(\hat\beta)-G(\hat\beta))|| &= \varepsilon_n^{-1}o_{\mathrm{p}}(n^{-1/2})\\ &= o_{\mathrm{p}}(n^{-1/4}). \qquad (A.66)\end{aligned}$$

Finally, note that for ε_n close enough to 0,

$$||G(\hat\beta+\varepsilon_n u_i)-G(\beta_0+\varepsilon_n u_i)-G(\hat\beta)|| = \varepsilon_n^{-1}O_{\mathrm{p}}(n^{-1/2}) = O_{\mathrm{p}}(n^{-1/4}) \qquad (A.67)$$

by the differentiability of $G(\cdot)$ in a neighborhood of β_0 and the $\sqrt{n}$-consistency of $\hat\beta$.

We have thus shown that

$$\hat J_i = \varepsilon_n^{-1}(G(\beta_0+\varepsilon_n u_i)-G(\beta_0))+o_{\mathrm{p}}(1), \qquad (A.68)$$

which establishes the desired result. □

References

Ahn, H., Powell, J.L., 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics 58, 3–29.

Amemiya, T., 1985. Advanced Econometrics. Cambridge, MA, Harvard University Press.

Andrews, D.W.K., 1994a. Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica 64, 43–72.

Andrews, D.W.K., 1994b. Empirical process methods in econometrics. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam.

Bhattacharya, P.K., Gangopadhyay, A.K., 1990. Kernel and nearest neighbor estimation of a conditional quantile. Annals of Statistics 18, 1400–1415.

Bierens, H.J., 1987. Kernel estimators of regression functions. In: Bewley, T.F. (Ed.), Advances in Econometrics, Fifth World Congress, Vol. 1. Cambridge University Press, Cambridge.

Buchinsky, M., 1994. Methodological issues in quantile regression. Ph.D. Dissertation, Department of Economics, Harvard University, Cambridge, MA.

Buchinsky, M., Hahn, J., 1998. An alternative estimator for the quantile regression model. Econometrica 66, 653–672.

Cavanagh, C.L., 1987. Limiting behavior of estimators defined by optimization. Manuscript, Department of Economics, Harvard University, Cambridge, MA.

Chaudhuri, P., 1991a. Nonparametric quantile regression. Annals of Statistics 19, 760–777.

Chaudhuri, P., 1991b. Global nonparametric estimation of conditional quantiles and their derivatives. Journal of Multivariate Analysis 39, 246–269.

Chaudhuri, P., Doksum, K., Samarov, A., 1997. On average derivative quantile regression. Annals of Statistics 25, 715–744.

Chen, S., Khan, S., 1998a. Semiparametric estimation of a semilinear censored regression model. Manuscript, Department of Economics, University of Rochester, Rochester, NY.

Chen, S., Khan, S., 1998b. Estimation of non–stationary censored regression panel data model. Manuscript Department of Economics, University of Rochester, Rochester, NY.

Chen, S., Khan, S., 2000. Estimating censored regression models in the presence of nonparametric multiplicative heteroskedasticity. Journal of Econometrics 98, 283–316.

Chen, S., Khan, S., 2001. Semiparametric estimation of a partially linear censored regression model. Econometric Theory 17, 567–590.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator of such models. Annals of Economic and Social Measurement 15, 475–492.

Honoré, B.E., Powell, J.L., 1994. Pairwise difference estimators of censored and truncated regression models. Journal of Econometrics 64, 241–278.

Horowitz, J.L., 1986. A distribution-free least squares estimator for linear censored regression models. Journal of Econometrics 32, 59–84.

Horowitz, J.L., 1988. Semiparametric M-estimation of censored linear regression models. Advances in Econometrics 7, 45–83.

Horowitz, J.L., 1992. A smoothed maximum score estimation for the binary response model. Econometrica 60, 505–531.

Khan, S., 2001. Two stage rank estimation of quantile index models. Journal of Econometrics 100, 319–355.

Kim, J., Pollard, D., 1990. Cube root asymptotics. Annals of Statistics 18, 191–219.

Koenker, R., Bassett Jr., G.S., 1978. Regression quantiles. Econometrica 46, 33–50.

Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.

Koenker, R., Portnoy, S., Ng, P., 1992. Nonparametric estimation of a conditional quantile function. In Dodge, Y. (Ed.), Proceedings of the conference on L_1-statistical Analysis and Related Methods. North–Holland, Amsterdam.

Manski, C.F., 1975. Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics 3, 205–228.

Manski, C.F., 1985. Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. Journal of Econometrics 27, 205–228.

Moon, C.-G., 1989. A Monte Carlo comparison of semiparametric Tobit estimators. Journal of Applied Econometrics 4, 361–382.

Müller, H.G., 1984. Smooth optimum kernel estimators of densities, regression curves and modes. Annals of Statistics 12, 766–774.

Nawata, K., 1992. Robust estimation based on group-adjusted data in censored regression models. Journal of Econometrics 43, 337–362.

Newey, W.K., McFadden, D., 1994. Estimation and hypothesis testing in large samples. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam.

Paarsch, H.J., 1984. A Monte Carlo comparison of estimators for censored regression models. Journal of Econometrics 24, 197–213.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1058.

Powell, J.L., 1984. Least absolute deviations estimation of the censored regression model. Journal of Econometrics 25, 303–325.

Powell, J.L., 1986a. Censored regression quantiles. Journal of Econometrics 32, 143–155.

Powell, J.L., 1986b. Symmetrically trimmed least squares estimation of Tobit models. Econometrica 54, 1435–1460.

Powell, J.L., 1994. Estimation of semiparametric models. In: Engle, R.F, McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam.

Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1404–1430.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Annals of Statistics 10, 1040–1053.

Stute, W., 1986. Conditional empirical processes. Annals of Statistics 14, 638–647.

Truong, Y., 1989. Asymptotic properties of kernel estimates based on local medians. Annals of Statistics 17, 606–617.