

# Efficient Semiparametric Estimation of Censored and Truncated Regressions via a Smoothed Self-Consistency Equation

Stephen R. Cosslett<sup>1</sup>  
Department of Economics  
Ohio State University

Revised: September 2003

## Abstract

An asymptotically efficient likelihood-based semiparametric estimator is derived for the censored regression (tobit) model, based on a new technique for estimating the density function of the residuals in semiparametric models that involve an underlying partially observed regression. Smoothing the self-consistency equation for the nonparametric maximum likelihood estimator of the distribution of the residuals yields an integral equation, which in some cases can be solved explicitly. The resulting estimated density is smooth enough to be used in a practical implementation of the profile likelihood estimator, but is still sufficiently close to the nonparametric maximum likelihood estimator to allow estimation of the semiparametric efficient score. The parameter estimates obtained by solving the estimated score equations are then asymptotically efficient. Details of the method and proofs are given specifically for the censored regression (tobit) model, with a summary of analogous results for the case of random censoring and for truncated regression. Simulation results are also presented.

---

<sup>1</sup> I would like to thank Songnian Chen, H. L. Koul, and three anonymous referees for helpful comments. An earlier version of this paper was presented at the North American Summer Meeting of the Econometric Society, June 2002.

## 1. Introduction

This paper presents an asymptotically efficient semiparametric estimator for the censored regression (tobit) model, based on the profile likelihood approach together with a new technique for estimating the unknown distribution function of partially observed residuals. The error distribution is estimated by solving an integral equation, derived from the self-consistency equation for the corresponding nonparametric maximum likelihood estimator (MLE). The estimated density is smooth enough to be plugged into the parametric likelihood function; it yields a score function that is asymptotically equivalent to the semiparametric efficient score, leading to asymptotically efficient parameter estimates.

Consider a semiparametric model with an underlying latent variable  $y^*$  generated by the linear model  $y^* = x\beta_0 + \varepsilon$ , and an observed dependent variable  $y$  with likelihood  $\ell(y, x\beta, f)$ . Here  $\ell$  is a known function of the index  $x\beta$  and the density function  $f$  of the error terms  $\varepsilon$ . One standard approach, associated with the profile likelihood method of Severini and Wong (1992), is to first estimate the density function  $f(\cdot|\beta)$  of the residuals  $e = y^* - x\beta$  by some suitable estimator  $\tilde{f}(\cdot|\beta)$ ; substitute this estimator for  $f$  in the likelihood; and then estimate  $\beta$  either by maximizing the log likelihood or by solving the corresponding score equation. Under some regularity and convergence-rate conditions on  $\tilde{f}$ , (see, for example, van der Vaart, 1998, p.391), the score  $S(\beta, \tilde{f}(\cdot|\beta))$  will be asymptotically equivalent to the efficient score  $S(\beta, f(\cdot|\beta))$ . In that case the estimator  $\hat{\beta}$  will achieve the semiparametric efficiency bound. A natural estimator of  $f(\cdot|\beta)$  in this context is the nonparametric MLE, but this may be irregular, typically with discrete mass points, and the desirable asymptotic properties of the estimated score function may not survive the necessary smoothing and trimming operations.

The estimator proposed here is based on the “self-consistency” equation for the MLE of  $f$  (Efron, 1967; see also Turnbull, 1976, and Tsai and Crowley, 1985). Informally, an observed residual  $e_i$  contributes a point mass  $n^{-1}\delta(\varepsilon - e_i)$  to the empirical density; if it is not observable, then its contribution is distributed in proportion to its density conditional on the observed variables,  $n^{-1}p(\varepsilon|y_i, x_i\beta, f(\cdot|\beta))$ . Replacing  $f$  by its estimate  $\hat{f}$  then gives

$$\hat{f}(u|\beta) = \sum_i p(u|y_i, x_i\beta, \hat{f}(\cdot|\beta)),$$

which can be integrated to give a self-consistency equation for the estimated distribution function  $\hat{F}(\cdot|\beta)$ . In a number of common applications, the solution can be expressed in closed form and is equal to the MLE of  $F$ .

In the present approach, instead of directly solving for  $\hat{F}$ , the self-consistency equation is first smoothed over the index  $x_i\beta$  (by a conventional kernel method). This results in an integral equation for the smoothed estimator  $\tilde{F}$ . In general, this integral equation could be solved for  $\tilde{F}$  by the EM algorithm, or by some direct numerical method of solution. However, in a class of relatively simple semiparametric models, which includes binary choice, tobit, truncated regression, the two-equation censored regression model, and endogenously stratified regression with two strata, the integral equation can be solved to give an explicit solution for  $\tilde{F}$  and its derivative  $\tilde{f}$ . The estimated log likelihood is then  $\log \tilde{L}(\beta) = \log L(\beta, \tilde{f}(\cdot|\beta))$ , and the resulting first-order conditions  $\tilde{S}(\beta) = 0$  are solved for  $\hat{\beta}$ .

In two of these cases, this approach leads to efficient semiparametric estimators that are already known: the Klein and Spady (1993) estimator for binary choice, and the semiparametric maximum likelihood estimator of Ai (1997) for the two-equation censored regression model (type 2 tobit).

This paper formulates the estimator and derives its asymptotic properties for a specific case, the tobit model  $y = \max(0, x\beta + \varepsilon)$  (i.e., left-censored regression with a fixed censoring point), where the errors are i.i.d. with unknown probability density function  $f(\varepsilon)$  and are independent of the regressors. Results for some other models are summarized in the appendix. Key properties needed for this approach to work are, first, that the solution of the self-consistency equation is in fact a consistent estimator of  $F$ , and secondly, that trimming affects neither the expected value of the score function nor the asymptotic orthogonality between the estimators of  $\beta$  and  $f$  (as expressed by the moment condition given in equation A.22 below); in general it would be difficult to characterize the set of models for which these conditions hold.

For censored regression, the MLE of  $F$  is the Kaplan-Meier estimator (Kaplan and Meier, 1958).<sup>2</sup> It underlies a number of semiparametric estimators for the censored regression model, including the  $M$ -function approach of Buckley and James (1979) and its modifications by Lai and Ying (1991), Ritov (1990), and others. Related work in the econometric literature on the tobit model includes the SGLS (semiparametric generalized least squares) estimators of Horowitz (1986, 1988) and Ichimura (1993).<sup>3</sup> However, these estimators do not achieve the efficiency bound. In related work on the censored regression model with random censoring, Lai and Ying (1992) proposed using a rank

---

<sup>2</sup> For a discussion of the Kaplan-Meier estimator, see for example Pagan and Ullah (1999, p.325).

<sup>3</sup> Previous likelihood-based semiparametric estimators of the tobit model with an estimated  $f$  include those of Duncan (1986) and Fernandez (1986).

estimator, weighted by a preliminary smoothed consistent estimator of  $d \log(f/F)/d\varepsilon$  to achieve the semiparametric efficiency bound; more recently Kim and Lai (2000), using asymptotic results derived by Lai and Ying (1994), developed an asymptotically efficient adaptive  $M$ -estimator for random censoring, using spline estimation of the appropriate weight factor  $d \log f / d\varepsilon$ .<sup>4</sup>

Tobit estimators based on weaker assumptions about the relationship between the errors and the regressors include those of Powell (1984) for the case where the conditional median of the error terms is zero, and Newey (2001) (and other references cited there) for the case of conditional moment restrictions. Powell's estimator can be used when an initial consistent estimator is required.

The next section derives the smoothed version of the Kaplan-Meier estimator and shows how it is used to set up the estimated score function  $\tilde{S}(\beta)$  that defines the semiparametric estimator  $\hat{\beta}$ . Section 3 outlines the convergence of  $\tilde{S}(\beta)$  to the efficient score  $S(\beta)$ , and then brings in the relevant trimming functions needed to establish uniform convergence. Finally, the asymptotic properties of  $\hat{\beta}$  are established by showing that it is asymptotically equivalent to the solution of  $S(\beta) = 0$ . Important steps are to verify that trimming has no effect on consistency or on the asymptotic orthogonality between the estimators of  $\beta$  and  $f$ , which implies that only the terms of second order in the kernel estimation errors have to converge fast enough not to interfere with the asymptotic variance of  $\hat{\beta}$ . This can be done without the need for either sample splitting or bias-reducing kernels. Concluding remarks are given in Section 4.

Appendix A contains technical assumptions and proofs of the propositions stated in the main text. Since the proofs closely follow those developed in previous research on kernel-based semiparametric estimators, particularly by Ichimura and Lee (1991), Klein and Spady (1993), and Ai (1997), only the essential steps are given here. The final two sections of the Appendix A summarize the corresponding estimators for the case of random censoring and the case of truncated regression. Appendix B gives the results of a small simulation study to investigate how closely the estimator approaches the efficiency bound for samples of realistic size, and a comparison with the adaptive  $M$ -estimator of Kim and Lai (2000).

---

<sup>4</sup> Efficient estimation of censored regression via an  $M$ -estimator with adaptive weights was first proposed by Ritov (1986).

## 2. Semiparametric Likelihood Function

### *A smoothed version of the Kaplan-Meier estimator*

The starting point is the “self-consistency” equation for the nonparametric maximum likelihood estimator of the distribution function  $F(\varepsilon)$  of a censored random variable  $\varepsilon$  from a sample of  $n$  observations  $(u_i, c_i, d_i)$ , where  $u_i = \max(\varepsilon_i, c_i)$  and  $d_i = 1(\varepsilon_i > c_i)$ . Assign a weight  $n^{-1}$  to each observation. For uncensored observations, the weight is concentrated at the observed point  $u_i$ . For censored observations, suppose that the weight is distributed according to the probability density of the unobserved point  $\varepsilon_i$ , i.e.,  $1(\varepsilon \leq c_i)f(\varepsilon)/F(c_i)$ , where  $f$  is the corresponding density function. That would give the following formal representation of the empirical density function,

$$\hat{f}(\varepsilon|\beta) = \frac{1}{n} \sum_{i=1}^n \{1(d_i = 1) \delta(\varepsilon - u_i) + 1(d_i = 0) 1(u_i > \varepsilon) f(\varepsilon|\beta) / F(u_i|\beta)\}. \quad (2.1)$$

Integrating (2.1) with respect to  $\varepsilon$  gives the empirical distribution function,

$$\hat{F}(\varepsilon) = \frac{1}{n} \sum_{i=1}^n \{1(u_i \leq \varepsilon) + 1(d_i = 0) 1(u_i > \varepsilon) F(\varepsilon) / F(u_i)\} \quad (2.2)$$

To make this operational, replace the unknown  $F$  on the right-hand side by its estimate  $\hat{F}$ , giving

$$\hat{F}(\varepsilon) = \frac{1}{n} \sum_{i=1}^n \left\{ 1(u_i \leq \varepsilon) + 1(d_i = 0) 1(u_i > \varepsilon) \hat{F}(\varepsilon) / \hat{F}(u_i) \right\} \quad (2.3)$$

As is well known, the solution of this equation is the Kaplan-Meier product-limit estimator, which is the nonparametric maximum likelihood estimator of  $F$ .

To obtain an estimate of  $F$  more suitable for use in semiparametric estimation, we smooth the terms on the right-hand side over the observed values  $u_i$ . In general, a function  $\gamma(z_i)$  of an observation  $z_i$  is “smoothed” by replacing it by

$$\tilde{\gamma}(z_i) = \frac{1}{h_n} \int dv \gamma(v) K\left(\frac{v - z_i}{h_n}\right) \quad (2.4)$$

with a suitable kernel function  $K$  and bandwidth  $h_n$ . Then equation (2.3) becomes

$$\tilde{F}(\varepsilon) = \frac{1}{n} \sum_{i=1}^n \left\{ \bar{K}\left(\frac{\varepsilon - u_i}{h_n}\right) + 1(d_i = 0) \tilde{F}(\varepsilon) \frac{1}{h_n} \int_{\varepsilon}^{\infty} dv \frac{1}{\tilde{F}(v)} K\left(\frac{v - u_i}{h_n}\right) \right\} \quad (2.5)$$

where

$$\bar{K}(u) = \int_{-\infty}^u dv K(v) \quad (2.6)$$

The integral equation (2.5) is linear in  $1/\tilde{F}(u)$ , and can readily be solved to give

$$\tilde{F}(\varepsilon) = \exp\left(-\int_{\varepsilon}^{\infty} dv \tilde{g}(v)/\tilde{G}(v)\right) \quad (2.7)$$

where

$$\begin{aligned} \tilde{g}(u) &= \frac{1}{nh_n} \sum_j 1(d_j = 1) K\left(\frac{u-u_j}{h_n}\right) \\ \tilde{G}(u) &= \frac{1}{n} \sum_j \bar{K}\left(\frac{u-u_j}{h_n}\right) \end{aligned} \quad (2.8)$$

The integral in (2.7) always exists because by construction the right-hand tail decreases at the same rate as the tail of the kernel function. The main difference between this and other smoothed versions of the Kaplan-Meier estimator is that the smoothing is applied to the underlying equation (2.3), which is then solved for  $\tilde{F}$ , as opposed to first solving (2.2) and then applying smoothing techniques to the solution  $\hat{F}$ .

### ***Estimated likelihood function for censored regression***

Returning to the censored regression equation, we can use this method to estimate the probability distribution  $F(u|\beta)$  of the latent residuals

$$e_i(\beta) = y_i^* - x_i\beta = \varepsilon_i - x_i(\beta - \beta_0) \quad (2.9)$$

with  $c_i = -x_i\beta$ ,  $d_i = 1(y_i > 0)$ , and  $u_i = u_i(\beta) = y_i - x_i\beta$ . This gives the estimator

$$\tilde{F}(u|\beta) = \exp\left(-\int_u^{\infty} dv \tilde{g}(v|\beta)/\tilde{G}(v|\beta)\right) \quad (2.10)$$

and the corresponding probability density function

$$\tilde{f}(u|\beta) = \left(\tilde{g}(u|\beta)/\tilde{G}(u|\beta)\right) \exp\left(-\int_u^{\infty} dv \tilde{g}(v|\beta)/\tilde{G}(v|\beta)\right) \quad (2.11)$$

where

$$\begin{aligned}\tilde{g}(u|\beta) &= \frac{1}{nh_n} \sum_j 1(y_j > 0) K\left(\frac{u - (y_j - x_j\beta)}{h_n}\right) \\ \tilde{G}(u|\beta) &= \frac{1}{n} \sum_j \bar{K}\left(\frac{u - (y_j - x_j\beta)}{h_n}\right)\end{aligned}\tag{2.12}$$

(When these functions are evaluated at  $u = u_i$ , the terms with  $j = i$  are dropped.) Using  $\tilde{f}(\cdot|\beta)$  and  $\tilde{F}(\cdot|\beta)$  in place of  $f$  and  $F$  in the tobit likelihood function then gives

$$\begin{aligned}\log \tilde{L}_n(\beta) &= \sum_{i=1}^n \left\{ 1(y_i > 0) \left( \log \tilde{g}(u_i(\beta)|\beta) - \log \tilde{G}(u_i(\beta)|\beta) \right) \right. \\ &\quad \left. - \int_0^\infty dv \tilde{g}(v + u_i(\beta)|\beta) / \tilde{G}(v + u_i(\beta)|\beta) \right\}\end{aligned}\tag{2.13}$$

An estimator of  $\beta$  could be defined by maximizing this objective function, but it will be more convenient to define  $\hat{\beta}$  as the solution of the likelihood equations  $\tilde{S}_n(\beta) = 0$ .<sup>5</sup> Since an initial consistent estimator, say  $\tilde{\beta}$ , is available, numerical solution of the likelihood equations can be started at  $\beta = \tilde{\beta}$ , and the theoretical analysis of the asymptotic properties of  $\hat{\beta}$  can be restricted to a neighborhood of  $\beta_0$ .

The score function is given by

$$\tilde{S}_n(\beta) = \sum_{i=1}^n \tilde{s}(x_i, y_i, \beta)\tag{2.14}$$

with

$$\tilde{s}(x, y, \beta) = 1(y > 0) \tilde{m}_1(u(\beta), x, \beta) - \int_0^\infty dv \tilde{m}_2(u(\beta) + v, x, \beta)\tag{2.15}$$

where  $u(\beta) = y - x\beta$  and

$$\tilde{m}_1(u(\beta), x, \beta) = \frac{d}{d\beta} \log \left( \frac{\tilde{g}(u(\beta)|\beta)}{\tilde{G}(u(\beta)|\beta)} \right)\tag{2.16}$$

$$\tilde{m}_2(u(\beta), x, \beta) = \frac{d}{d\beta} \left( \frac{\tilde{g}(u(\beta)|\beta)}{\tilde{G}(u(\beta)|\beta)} \right)\tag{2.17}$$

In this notation,  $d/d\beta$  represents the total derivative with respect to  $\beta$ , i.e.,

---

<sup>5</sup> As with other semiparametric estimators of this type, we rely on solving the score equations rather than maximizing the likelihood function because of the difficulty in trimming the log likelihood function while retaining its desirable asymptotic properties.

$$\frac{dg(u(\beta) | \beta)}{d\beta} = \frac{\partial g(u(\beta) | \beta)}{\partial u} \frac{\partial u(\beta)}{\partial \beta} + \frac{\partial g(u(\beta) | \beta)}{\partial \beta} \quad (2.18)$$

and thus

$$\frac{d\tilde{g}(u_i(\beta) | \beta)}{d\beta} = \frac{1}{nh_n^2} \sum_{j=1}^n 1(y_j > 0) (x_j - x_i) K' \left( \frac{(y_i - x_i\beta) - (y_j - x_j\beta)}{h_n} \right) \quad (2.19)$$

$$\frac{d\tilde{G}(u_i(\beta) | \beta)}{d\beta} = \frac{1}{nh_n} \sum_{j=1}^n (x_j - x_i) K \left( \frac{(y_i - x_i\beta) - (y_j - x_j\beta)}{h_n} \right) \quad (2.20)$$

where  $K'(u) = dK(u)/du$  and the terms with  $j = i$  are dropped.

### 3. Properties of the Estimator

#### *A related artificial likelihood function for censored regression*

To interpret the estimator  $\hat{\beta}$ , consider first the functions

$$g(u | \beta) = \int dx h(x) f(u + x[\beta - \beta_0]) 1(x\beta > -u) \quad (3.1)$$

$$G(u | \beta) = \int dx h(x) F(u + x[\beta - \beta_0]) 1(x\beta > -u) \quad (3.2)$$

where  $h(x)$  is the density of  $x$ . The function  $g(u | \beta) / \Pr(y > 0)$  is the density of the residuals  $u(\beta)$  conditional on  $y > 0$ , while  $G(u | \beta)$  is the (unconditional) distribution function of  $u(\beta)$ .<sup>6</sup>

By calculating the means and variances of the kernel estimates  $\tilde{g}(u | \beta)$  and  $\tilde{G}(u | \beta)$  in the usual way, and letting  $h_n \rightarrow 0$  and  $nh_n \rightarrow \infty$  as  $n \rightarrow \infty$ , we see that  $\tilde{g}(u | \beta)$  and  $\tilde{G}(u | \beta)$  are consistent estimators of  $g(u | \beta)$  and  $G(u | \beta)$  (for fixed  $u$  and  $\beta$ , not both equal to zero). Therefore,  $\log \tilde{L}_n(\beta)$  can be viewed as an estimate of the artificial log likelihood function

$$\log \bar{L}_n(\beta) = \sum_{i=1}^n \left\{ 1(y_i > 0) (\log g(u_i(\beta) | \beta) - \log G(u_i(\beta) | \beta)) - \int_0^\infty dv g(v + u_i(\beta) | \beta) / G(v + u_i(\beta) | \beta) \right\} \quad (3.3)$$

(which is of course not a feasible objective function, because the functions  $g$  and  $G$  depend on the unknown parameter  $\beta_0$  as well as the unknown density  $f$ ). To verify that the integral in (3.3) exists, we note that  $g(u | \beta) \leq \partial G(u | \beta) / \partial \beta$ . Applying the bounded

---

<sup>6</sup> Note that  $g$  is not the derivative of  $G$ .

convergence theorem to the integral in (3.2) shows that  $G(\infty | \beta) = 1$ . It then follows that the integral in (3.3) is bounded by  $\log G(u_i | \beta)$ .

Let  $\bar{\beta}$  be the “estimator” that results from solving the likelihood equations  $S_n(\beta) = 0$ , where  $S_n(\beta)$  is the score function corresponding to the log likelihood in (3.3). At  $\beta = \beta_0$  we have

$$g(u | \beta_0) / G(u | \beta_0) = f(u) / F(u) = d \log F(u) / du \quad (3.4)$$

and therefore  $\bar{L}_n(\beta_0) = L_n(\beta_0)$ , the likelihood of the data-generating process. It follows that under standard regularity conditions (which follow from Assumptions 1–7 in the appendix) for classical maximum likelihood estimation, the “estimator”  $\bar{\beta}$  is consistent and asymptotically normal with asymptotic variance equal to the limiting variance of  $n^{-1/2} S_n(\beta_0)$ . Let

$$S_n(\beta) = \sum_{i=1}^n s(x_i, y_i, \beta) \quad (3.5)$$

with  $s$  defined by an equation analogous to (2.15) in terms of  $m_1$  and  $m_2$ , where

$$m_1(u(\beta), x, \beta) = (d / d\beta) \log (g(u(\beta) | \beta) / G(u(\beta) | \beta)) \quad (3.6)$$

$$m_2(u(\beta), x, \beta) = (d / d\beta) (g(u(\beta) | \beta) / G(u(\beta) | \beta)) \quad (3.7)$$

Evaluating the functions  $g$ ,  $G$  and their derivatives at  $\beta = \beta_0$  (with  $\beta_0 \neq 0$ ),<sup>7</sup>

$$g(u | \beta_0) = f(u)[1 - H(-u | \beta_0)], \quad G(u | \beta_0) = F(u)[1 - H(-u | \beta_0)]$$

$$\begin{aligned} \frac{dg(u | \beta_0)}{d\beta} &= f'(u)[1 - H(-u | \beta_0)] \{E[X | X\beta_0 > -u] - x\} \\ &\quad + f(u)h(-u | \beta_0) \{E[X | X\beta_0 = -u] - x\} \end{aligned}$$

$$\begin{aligned} \frac{dG(u | \beta_0)}{d\beta} &= f(u)[1 - H(-u | \beta_0)] \{E[X | X\beta_0 > -u] - x\} \\ &\quad + F(u)h(-u | \beta_0) \{E[X | X\beta_0 = -u] - x\} \end{aligned}$$

where  $h(\cdot | \beta)$  is the marginal probability density function of  $x\beta$  and  $H(\cdot | \beta)$  is the corresponding distribution function. After some calculations, we find that

---

<sup>7</sup> A different formulation is needed in the case  $\beta_0 = 0$  because the marginal density function  $h(u | \beta)$  is singular at  $\beta = 0$ . Details are given in Appendix A.3.

$$s(x, y, \beta_0) = \left( \frac{f(u)}{F(u)} - \frac{f'(u)}{f(u)} \right) (x - E[X | X\beta_0 > -u]) 1(y > 0) \\ + \int_u^\infty dv \frac{d}{dv} \left( \frac{f(v)}{F(v)} \right) (x - E[X | X\beta_0 > -v]) \quad (3.8)$$

(where  $u = y - x\beta_0$ ), which is the same as the efficient semiparametric score (Cosslett, 1987; see also Bickel *et al.*, 1993). The asymptotic variance of  $\bar{\beta}$  is therefore equal to the semiparametric efficiency bound  $V_*$ , which is given (in the present notation) by

$$V_*^{-1} = \int du f(u) \left[ \frac{d}{du} \log \left( \frac{f(u)}{F(u)} \right) \right]^2 [1 - H(-u | \beta_0)] \text{var}[x | x\beta_0 > -u] \quad (3.9)$$

### ***Alternative score function***

The artificial log likelihood  $\bar{L}_n(\beta)$  in (3.3) has the following peculiar feature. If we drop the second term and use the simplified log likelihood

$$\log \bar{L}_{1,n}(\beta) = \sum_{i=1}^n \{1(y_i > 0) (\log g(u_i(\beta) | \beta) - \log G(u_i(\beta) | \beta))\} \quad (3.10)$$

with the corresponding score function

$$S_{1,n}(\beta) = \sum_{i=1}^n m_1(u_i(\beta), x_i, \beta) 1(y_i > 0) \quad (3.11)$$

then the resulting “estimator”  $\bar{\beta}_{(1)}$ , i.e., the solution of  $S_{1,n}(\beta) = 0$ , is also consistent with asymptotic variance equal to the semiparametric efficiency bound. One might therefore question the role of the integral term in (3.3), and instead use an estimate of the alternative score function (3.11). However, the important properties of the efficient score at  $\beta = \beta_0$  are, not only that its variance is equal to the asymptotic information bound (3.9), but also that it is asymptotically uncorrelated with the score function corresponding to variation in  $f$ . That in turn allows  $\hat{\beta}$  to have the same asymptotic variance despite the presence of the estimated functions  $\hat{g}$  and  $\hat{G}$ . More specifically, the orthogonality condition represented by (A.22) (in the Appendix) holds for the estimated score function (2.14) but not for the estimated version of the score function in (3.11).

### ***Trimming functions***

The goal, of course, is to show that the kernel-based estimator  $\hat{\beta}$  and the artificial estimator  $\bar{\beta}$  are asymptotically equivalent, using methods developed for kernel-based

semiparametric estimators by Klein and Spady (1993), Ichimura and Lee (1991), Ai (1997), and others. The key step is to show that the various kernel estimates in  $\tilde{s}_n(x, y, \beta)$  and their ratios converge uniformly in  $(x, y)$  and  $\beta$  to the corresponding functions in  $s_n(x, y, \beta)$ . In order to do this, as with other kernel-based semiparametric estimators of this type, trimming functions are needed to bound the denominator terms and, for technical reasons, the magnitudes of  $x$  and  $\varepsilon$ . Let  $t(u)$  be a trimming function such that

$$t(u) = \begin{cases} 1 & \text{for } u \geq 1 \\ 0 & \text{for } u \leq 0 \end{cases}$$

To make  $\tau$  twice differentiable with continuous second derivative, a suitable bridging function for  $0 \leq u \leq 1$  is  $t(u) = u^3(6u^2 - 15u + 10)$ .

(i) Consider a term in the score function of the form  $N/D$ , where the numerator  $N$  is a bounded function with some structure that we do not want to lose inside a trimming function, while the denominator  $D$  is positive but not bounded away from zero. Then to give the denominator a shrinking lower bound of order  $b_n$ , the factor  $1/D$  is replaced by the trimmed function

$$\tau_1(D) = \frac{1}{D} t([D - b_n]/b_n). \quad (3.12)$$

The first and second derivatives of  $\tau_1(D)$  are of order  $b_n^{-2}$  and  $b_n^{-3}$  respectively.<sup>8</sup>

(ii) The range of integration is restricted to an expanding interval of order  $u_n$  by multiplying the integrand by the trimming function

$$\tau_2(u) = t(u_n - |u|), \quad (3.13)$$

where  $u$  is the integration variable.

(iii) If we need to allow for the case  $\beta_0 = 0$ , then  $u$  is also restricted to lie outside a shrinking neighborhood of the point  $u = 0$ . A suitable trimming function is

$$\tau_3(u) = t([|u| - \mu_n]/\mu_n), \quad (3.14)$$

where  $u$  is the residual in  $m_1$  or the integration variable in  $m_2$ .

(iv)  $x$  and  $\varepsilon$  are restricted to an expanding set  $W_n$  with volume bounded by a power of  $n$  (as in Ai, 1997). Under the assumptions leading to equation (A.5) in the appendix, all observations fall within  $W_n$  with probability approaching 1, and so this restriction does not affect convergence in probability.

---

<sup>8</sup> A similar trimming factor is used by Ai (1997).

In order for needed cancellations to occur, all the terms in the score function have to have a common trimmed denominator. At the expense of a rather complicated score function, we therefore use the following denominator and trimming factor

$$\begin{aligned} D(u|\beta) &= g(u|\beta)G(u|\beta)^2 \\ \tau(u|\beta) &= \tau_1[D(u|\beta)]\tau_2(u), \end{aligned} \quad (3.15)$$

with the additional factor  $\tau_3(u)$  if needed.  $\tilde{D}(u|\beta)$  and  $\tilde{\tau}(u|\beta)$  are defined similarly in terms of the estimated functions  $\tilde{g}$  and  $\tilde{G}$ . In fact, the relative rates of convergence of  $b_n$  and  $u_n$  turn out to be immaterial to the proofs of consistency and asymptotic efficiency (see Condition 2 in Appendix A). Since the denominator terms  $D(u|\beta)$  and  $\tilde{D}(u|\beta)$  necessarily converge to zero for large  $|u|$ , it can always be arranged for  $\tau_1(D)$  to dominate  $\tau_2(u)$ , and so make the second trimming factor unnecessary.<sup>9</sup>

The trimmed estimated score function is

$$\tilde{S}_n^*(\beta) = \sum_{i=1}^n \tilde{s}^*(x_i, y_i, \beta) \quad (3.16)$$

where

$$\tilde{s}^*(x, y, \beta) = \tilde{m}_1^*(u(\beta), x, \beta) 1(y > 0) + \int_0^\infty dv \tilde{m}_2^*(u(\beta) + v, x, \beta) \quad (3.17)$$

$$\tilde{m}_1^*(u, x, \beta) = \tilde{m}_1(u, x, \beta) \tilde{D}(u|\beta) \tilde{\tau}(u|\beta) \quad (3.18)$$

$$\tilde{m}_2^*(u, x, \beta) = \tilde{m}_2(u, x, \beta) \tilde{D}(u|\beta) \tilde{\tau}(u|\beta) \quad (3.19)$$

The score function  $S_n^*(\beta)$  and its components  $s^*(x, y, \beta)$ ,  $m_1^*(u, x, \beta)$  and  $m_2^*(u, x, \beta)$  are defined analogously in terms of  $g$  and  $G$ . The estimator  $\hat{\beta}$  is the solution of  $\tilde{S}_n^*(\beta) = 0$ .<sup>10</sup>

For this approach to work, the trimming must be done in a way that does not affect the expected value of the score at  $\beta = \beta_0$ .<sup>11</sup> If not, trimming may produce asymptotic bias in the estimator, depending on the rates of convergence of  $b_n$  and  $u_n$ . The following result is derived in the appendix:

*Proposition 1.* With  $s^*$  as defined above,  $E[s^*(X, Y, \beta_0) | \varepsilon] = 0$ .

---

<sup>9</sup> This does not, of course, have any bearing on the merits of trimming the integral in a practical implementation of the estimator.

<sup>10</sup> Under standard regularity conditions, a consistent solution exists with probability approaching 1 as  $n \rightarrow \infty$ . If there is no solution, then instead  $\hat{\beta}$  minimizes  $\tilde{S}_n^*(\beta)' \tilde{S}_n^*(\beta)$ .

<sup>11</sup> Lai and Ying, 1991, discuss this problem in connection with the modified Buckley-James estimator.

*Asymptotic equivalence of  $\hat{\beta}$  and  $\bar{\beta}$ .*

The following results show that  $\hat{\beta}$  and  $\bar{\beta}$  are asymptotically equivalent, and therefore that  $\hat{\beta}$  is consistent and asymptotically efficient.

*Proposition 2. Under Assumptions 1–10 and Conditions 1–2 (given in Appendix A.1), and if  $\beta_0 \neq 0$ ,*

- (a)  $\frac{1}{n} \left( \tilde{S}_n^*(\beta) - S_n(\beta) \right) \xrightarrow{P} 0$  uniformly in  $\beta \in B$
- (b)  $\frac{1}{n} \left( \frac{d\tilde{S}_n^*(\beta)}{d\beta'} - \frac{dS_n(\beta)}{d\beta'} \right) \xrightarrow{P} 0$  uniformly in  $\beta \in B$
- (c)  $\frac{1}{\sqrt{n}} \left( \tilde{S}_n^*(\beta_0) - S_n(\beta_0) \right) \xrightarrow{P} 0$

The proof is given in Appendix A.2. Proposition 3, which is stated and proved in Appendix A.3, gives the analogous results for the case  $\beta_0 = 0$ .

Part (a) of Proposition 2 shows that  $\hat{\beta}$  and  $\bar{\beta}$  have the same probability limit, so  $\hat{\beta}$  is consistent. The usual first-order series expansions of  $\tilde{S}_n(\hat{\beta})$  and  $S_n(\bar{\beta})$  in  $(\beta - \beta_0)$  give

$$\sqrt{n} (\hat{\beta} - \beta_0) = - \left( \frac{d\tilde{S}_n^*(\beta_*)}{d\beta'} \right)^{-1} \tilde{S}_n^*(\beta_0)$$

(where  $\beta_*$  is between  $\hat{\beta}$  and  $\beta_0$ ), and similarly for  $\sqrt{n} (\bar{\beta} - \beta_0)$ . Parts (b) and (c) of Proposition 2 then show that  $\sqrt{n} (\hat{\beta} - \bar{\beta}) \xrightarrow{P} 0$ , i.e., that  $\hat{\beta}$  and  $\bar{\beta}$  have the same asymptotic distribution. Each result is derived in two steps. In part (a), for example, one first works with the trimmed functions to show that  $n^{-1} \left( \tilde{S}_n^*(\beta) - S_n^*(\beta) \right)$  converges, using existing results on uniform convergence of kernel estimates such as  $\tilde{g}(u|\beta)$  at rates slightly slower than  $n^{-1/2}$ . Then one shows that the trimming does not matter asymptotically, i.e., that  $n^{-1} \left( S_n^*(\beta) - S_n(\beta) \right)$  also converges. Part (b) is derived similarly. In the corresponding first step of part (c), where a faster rate of convergence is needed, quadratic and higher-order terms such as  $(\tilde{g} - g)^2$  can be made to converge faster than  $n^{-1/2}$ , but linear terms such as  $\tilde{g} - g$  cannot. For these terms, one can show convergence by rewriting the sample means of kernel estimates as  $U$ -statistics, and then applying standard results on the asymptotic properties of  $U$ -statistics. Because these linear terms have expected value zero, convergence is fast enough not to require the use of bias-reducing kernels. Details are given in Appendix A.2.

## 4. Conclusion

The self-consistency equation for the nonparametric MLE of the distribution of a partially observed random variable can be converted into an integral equation by kernel smoothing. Its solution can be regarded as a smoothed version of the MLE that is more tractable when used in semiparametric estimation. In some applications, one can then construct an estimator of the efficient score and obtain parameter estimates that achieve the asymptotic semiparametric efficiency bound. In this paper, the derivation and proofs are worked out in detail for the specific case of the censored regression (tobit) model with independent errors. The resulting asymptotically efficient semiparametric estimator  $\hat{\beta}$  is a solution of the score equation  $\tilde{S}_n^*(\beta) = 0$  (defined by equations 3.16–3.19), which is based on the classical tobit likelihood function and a smoothed version of the Kaplan-Meier estimator.

It is somewhat more complicated than other estimators of this type, since evaluation of the score function involves an integral of a ratio of kernel estimates, but the computational problem of a one-dimensional numerical quadrature in evaluating the objective function can be handled relatively easily. The simulations in Appendix B show satisfactory performance in a tobit model with sample sizes of 100 and 400 and with several different error distributions. The simulations also show that the estimator can handle the case  $\beta_0 = 0$ . Further studies on bandwidth selection and on the use of alternative kernels may help to improve the small-sample properties of the estimator.

The approach presented here can also be used to derive efficient semiparametric estimators for other problems, such as truncated regression (Appendix A.5) and endogenously stratified regression with two strata.<sup>12</sup> In more general cases (for example, endogenous stratification with more than two strata), it is likely that the analog of the integral equation (2.5) will not have an explicit solution. While the corresponding estimator can be implemented by numerical solution of the integral equation, the derivation of its asymptotic properties will be more challenging.

## Appendix A

### A.1. Assumptions

The following assumptions are intended to show that there is a reasonable set of conditions under which the proposed estimator has the desired asymptotic properties, i.e., consistency and asymptotic normality, with asymptotic variance equal to the

---

<sup>12</sup> Details of those estimators will be presented elsewhere.

semiparametric efficiency bound. The goal is to allow a relatively direct proof of the results, rather than to find the least restrictive conditions.

Assumptions 1–7 provide standard regularity conditions for establishing the classical result that  $\bar{\beta}$ , the solution of  $S_n(\beta) = 0$  (in a neighborhood of the initial consistent estimator) is consistent and asymptotically normal, with asymptotic variance given by the inverse of (3.9). They are also used in showing that the solution of the trimmed score equation  $\tilde{S}_n(\beta) = 0$  is asymptotically equivalent to  $\bar{\beta}$ .

*Assumption 1.*  $\beta \in B$ , an open subset of  $R^k$ .

*Assumption 2.*  $x \in R^k$  and  $\varepsilon \in R$  are independent random variables with density functions  $h(x)$  and  $f(\varepsilon)$ . The regression function  $x\beta$  is a continuous random variable for all  $\beta \in B$  (except of course for  $\beta = 0$ , if  $0 \in B$ ).

*Assumption 3.* The data consists of a random sample  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , with  $y_i = \max(0, x_i\beta_0 + \varepsilon_i)$ .

*Assumption 4.* An initial  $\sqrt{n}$ -consistent estimator of  $\beta$  is available.

This is not restrictive, because there are a number of  $\sqrt{n}$ -consistent semiparametric estimators of the tobit model in the literature, such as the least absolute deviations estimator of Powell (1984). As a result, the parameter space  $B$  can be replaced by a neighborhood of  $\beta_0$ .

*Assumption 5.* The functions  $g(u|\beta)$  and  $G(u|\beta)$  defined in (3.1) are twice continuously differentiable with respect to  $u$  and  $\beta$  (for  $\beta \neq 0$ ).

Note that since we have explicit expressions for  $g(u|\beta)$ ,  $G(u|\beta)$  and their derivatives, i.e.,

$$\begin{aligned} \partial g(u|\beta) / \partial u &= E[f'(u + X[\beta - \beta_0]) | X\beta > -u][1 - H(-u|\beta)] \\ &\quad + E[f(-X\beta_0) | X\beta = -u]h(-u|\beta) \end{aligned} \quad (\text{A.1})$$

$$\begin{aligned} \partial g(u|\beta) / \partial \beta &= E[X f'(u + X[\beta - \beta_0]) | X\beta > -u][1 - H(-u|\beta)] \\ &\quad + E[X f(-X\beta_0) | X\beta = -u]h(-u|\beta) \end{aligned} \quad (\text{A.2})$$

$$\begin{aligned} \partial^2 g(u|\beta) / \partial u^2 &= E[f''(u + X[\beta - \beta_0]) | X\beta > -u][1 - H(-u|\beta)] \\ &\quad + h(-u|\beta)E[f'(-X\beta_0) | X\beta = -u] \\ &\quad + (\partial / \partial u)\{h(-u|\beta)E[f(-X\beta_0) | X\beta = -u]\} \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned}
\partial^2 g(u|\beta)/\partial\beta\partial\beta' &= E[XX' f''(u + X[\beta - \beta_0]) | X\beta > -u] [1 - H(-u|\beta)] \\
&\quad + h(-u|\beta) E[XX' f'(-X\beta_0) | X\beta = -u] \\
&\quad + (\partial/\partial u) \{ h(-u|\beta) E[XX' f(-X\beta_0) | X\beta = -u] \}
\end{aligned} \tag{A.4}$$

and similarly for the derivatives of  $G(u|\beta)$ , Assumption 5 could in principle be expressed in terms of a (rather lengthy) set of conditions on  $f$ ,  $h$ , and the conditional expected values of  $x$ .

*Assumption 6.* Let  $m_1$  and  $m_2$  be defined by (3.6)–(3.7), and the total derivatives by (2.18). Then (i)  $|m_1(u(\beta), x, \beta)|$  and  $|dm_1(u(\beta), x, \beta)/d\beta|$  are bounded by a function  $\phi_1(\varepsilon, x)$  such that  $E[\phi_1(\varepsilon, x)1(y > 0)] < \infty$ , and (ii)  $|m_2(v, x, \beta)|$  and  $|dm_2(v, x, \beta)/d\beta|$  are bounded a function  $\phi_2(v, x)$ , such that  $\int dv \phi_2(v, x)1(v > u(\beta))$  is also bounded uniformly in  $\beta$  by a function with finite expected value.

In the case  $\beta_0 = 0$ , the functions  $m_{0,1}$ ,  $m_{0,2}$ ,  $\dot{m}_{0,1}$  and  $\dot{m}_{0,2}$  defined by (A.30)–(A.33) are substituted for  $m_1$ ,  $m_2$  and their derivatives.

While there does not appear to be a straightforward way of expressing Assumption 6 in terms of primitive conditions on the underlying density functions  $f$  and  $h$ , we note that the conditions are satisfied in the case where  $x$  and  $\varepsilon$  are normally distributed. In that case, (i)  $m_1(u, x, \beta)$ ,  $m_2(u, x, \beta)$  and their derivatives have bounds of the form  $p_1(x)p_2(u)$ , where  $p_1$  and  $p_2$  are polynomials, and (ii)  $m_2$  and its derivatives decrease rapidly for large positive  $u$ . With bounds of that form, and  $\beta$  restricted to a compact set, Assumption 6 can readily be verified.

*Assumption 7.* The asymptotic information bound  $I_*$  is finite.

The asymptotic information bound is equal to  $E[m_1(\varepsilon, x, \beta_0)m_1(\varepsilon, x, \beta_0)'1(x\beta_0 + \varepsilon > 0)]$ . Explicit expressions for  $I_*$  are given by (3.9), or in the case  $\beta_0 = 0$  by (A.35).

The following Assumptions 8–10 are used in determining the rates of convergence of the kernel estimators of  $g$ ,  $G$  and their derivatives. In the next assumption, let

$$\begin{aligned}
g_1(u|\beta) &= \int dx h(x) x f(u + x[\beta - \beta_0])1(x\beta > -u) \\
g_2(u|\beta) &= \int dx h(x) x x' f(u + x[\beta - \beta_0])1(x\beta > -u)
\end{aligned}$$

and similarly for  $G_1(u|\beta)$  and  $G_2(u|\beta)$ .

*Assumption 8.* (i)  $f(\varepsilon)$  is bounded; (ii) in the case  $\beta_0 \neq 0$ , the derivatives of  $g(u|\beta)$ ,  $g_1(u|\beta)$ ,  $g_2(u|\beta)$ ,  $G(u|\beta)$ ,  $G_1(u|\beta)$  and  $G_2(u|\beta)$  up to fourth order are bounded for  $\beta \in B$ ; (iii) in the case  $\beta_0 = 0$ , the derivatives of  $f(\varepsilon)$  up to fourth order are bounded.

It follows from (i) that  $g(u|\beta)$  is bounded;  $G(u|\beta)$  is of course bounded by construction. Assumption 8(ii) is the usual type of condition needed for bounding the bias terms in the kernel estimators. Up to fourth derivatives are needed in (ii) because we need to estimate the second derivatives  $d^2\tilde{g}/d\beta d\beta'$  and  $d^2\tilde{G}/d\beta d\beta'$ , and two further derivatives are then needed to evaluate the bias term. If required, (A.1)–(A.4) and similar equations for the higher-order derivatives could be used to express Assumption 8 in terms of  $f$ ,  $h$ , and the relevant conditional moments of  $x$ .

*Assumption 9.*  $E[|x|^p] < \infty$  for some  $p > 4$ .

This assumption, together with Assumption 8(i), bounds the variances of the kernel estimators.

*Assumption 10.*  $E[|\varepsilon|^q] < \infty$  for some  $q > 0$ .

This rules out, for example, distributions with logarithmically decreasing tail probabilities. It follows from Assumptions 9 and 10 that we can find a set  $W_n \subset R^{k+1}$ , with volume increasing no faster than a power of  $n$ , such that

$$\Pr\{(x, \varepsilon) \notin W_n\} = o(n^{-1}) \quad (\text{A.5})$$

and therefore, with probability approaching 1 as  $n \rightarrow \infty$ , all sample observations are in  $W_n$ . This allows for uniform convergence on an expanding set (see Lemma B.1 in Ai, 1997).

The following conditions refer to the construction of the estimator, specifically the properties of the kernel function and the rates of convergence of the window width and the trimming parameters.

*Condition 1.*  $K$  is a conventional kernel function, i.e., a bounded, differentiable, symmetric function that satisfies  $\int K(u)du = 1$ ,  $\int u^2 K(u)du < \infty$ ,  $\int [dK(u)/du]^2 du < \infty$ , and  $\int [d^2 K(u)/du^2]^2 du < \infty$ .

*Condition 2.* Let the convergence rates of the kernel bandwidth  $h_n$  and the trimming parameters  $b_n$  and  $u_n$  be  $h_n \sim n^{-\alpha}$ ,  $b_n \sim n^{-\beta}$ ,  $u_n \sim n^\gamma$  (with  $\alpha > 0$ ,  $\beta > 0$ , and  $\gamma > 0$ ). Then  $\beta + \gamma < \frac{1}{2} - \frac{5}{2}\alpha$ ,  $2\beta + \gamma < \frac{1}{2}\alpha$ ,  $3\beta + \gamma < 4\alpha - \frac{1}{2}$ , and  $\alpha < (p-4)/(p+4)$  (where  $p$  comes from Assumption 9).

The rate restrictions in Condition 2 imply  $\frac{1}{8} < \alpha < \frac{1}{5}$ , and also  $p > \frac{36}{7}$ . (This could be relaxed by using a kernel with higher-order bias reduction, but in any event we need  $p > 4$ .) Note that one implication of Condition 2 is that  $h_n^{1/2}$  decreases faster than  $b_n$ ,

which is useful in the following proofs when we have to determine which terms have the slowest rate of convergence.

To see where the restrictions in Condition 2 come from, note that in Propositions 2(i) and 2(ii), the critical term (the one with the slowest rate of convergence) is the kernel estimation error  $\Delta d^2 \tilde{g} / d\beta d\beta'$ . This term is  $O(h_n^{-5/2}) o_p(n^{-1/2+})$ , and so we must have  $\alpha < \frac{1}{5}$ . With trimming, convergence requires  $u_n b_n^{-1} h_n^{-5/2} n^{-1/2} \rightarrow 0$ , which gives the first restriction. Because  $h_n$  decreases slower than  $n^{-1/5}$ , the kernel estimation errors  $\Delta \tilde{g}$ ,  $\Delta \tilde{G}$  and  $\Delta d\tilde{G} / d\beta$  are bounded by the bias, which is  $O(h_n^2)$ , whereas  $\Delta d\tilde{g} / d\beta$  may be either  $O(h_n^{-3/2}) o_p(n^{-1/2+})$  or  $O(h_n^2)$  depending on the value of  $\alpha$ . In Proposition 2(iii), the critical terms are products of kernel estimation errors, which have to converge to zero at a rate faster than  $n^{-1/2}$ . Terms of the form  $\Delta \tilde{g} \cdot \Delta d\tilde{g} / d\beta \cdot \tau_1'$  and  $(\Delta \tilde{g})^2 \cdot \tau_1''$  have worst-case convergence rates of  $O(b_n^{-2} h_n^{1/2} n^{-1/2+})$  and  $O(b_n^{-3} h_n^4)$ , and those rates give the second and third restrictions. A key step in the proof of Proposition 2(iii) will then be to show that the terms linear in the kernel estimators converge at a faster rate and so impose no further restrictions. The fourth restriction in Condition 2 is a technical requirement needed for uniform convergence of the kernel estimators.

The remaining two conditions are not needed if the case  $\beta_0 = 0$  can be excluded *a priori*.

*Condition 3.*  $K'(u)$  has rapidly decreasing tails, i.e.,  $dK(u)/du = o(|u|^{-\nu})$  for all  $\nu \geq 0$ .

This can be achieved by a function either with exponential tails or with bounded support. It implies that the kernel function  $K(u)$  and the associated functions  $[\bar{K}(u) - 1(u \geq 0)]$  and  $\bar{K}_1(u)$  also have rapidly decreasing tails, where  $\bar{K}(u) = \int_{-\infty}^u dv K(v)$  and  $\bar{K}_1(u) = \int_{-\infty}^u dv v K(v)$ . A thin-tailed kernel allows relatively fast shrinkage of the trimming at  $u = 0$ , as follows:

*Condition 4.* The convergence rate of the trimming parameter  $\mu_n$  is  $\mu_n \sim n^{-\rho}$  where  $\rho < \alpha$ .

## A.2. Proofs of propositions

First consider the case where  $\beta_0 \neq 0$ , in which case the neighborhood  $B$  can be chosen so as not to contain the point  $\beta = 0$ . The modifications needed to allow for the case  $\beta_0 = 0$  are given in the next subsection.

*Proposition A.* Under Assumptions 1–7,  $\bar{\beta}$  is consistent and asymptotically normal, with asymptotic variance given by the inverse of (3.9).

With Assumption 6,  $n^{-1}S_n(\beta)$  and  $n^{-1}\partial S_n(\beta)/\partial\beta$  converge in probability to their expected values uniformly in  $\beta$ . By Assumption 7,  $E[s(X, Y, \beta_0)s(X, Y, \beta_0)^T] < \infty$ . By explicit calculation, we can then check that  $E[s(X, Y, \beta_0)] = 0$  and that both  $\text{var}[s(X, Y, \beta_0)]$  and  $E[-\partial s(X, Y, \beta_0)/\partial\beta]$  are equal to the asymptotic information matrix (3.9). Consistency and asymptotic normality then follow from standard results (see, for example, Hansen, 1982).

An explicit identification condition is not needed, because under the above assumptions (3.9) is positive definite and is equal to  $E[-\partial s(X, Y, \beta_0)/\partial\beta]$ , so that  $E[s(X, Y, \beta)]$  has an isolated zero at  $\beta = \beta_0$ . Since we assumed that there is an initial consistent estimator, this is enough to identify  $\beta$ .

*Proof of Proposition 1*

We have to show that  $E[s^*(X, Y, \beta_0) | \varepsilon] = 0$ , where

$$s^*(x, y, \beta_0) = m_1^*(\varepsilon, x, \beta) 1(y > 0) - \int_0^\infty dv m_2^*(v + u, x, \beta)$$

with  $u = y - x\beta_0$ . In fact, this will follow from two slightly more general results. First, let

$$\mu(u, x) = G(u | \beta_0) dg(u | \beta_0) / d\beta - g(u | \beta_0) dG(u | \beta_0) / d\beta$$

From (3.8), we see that  $\mu(u, x)$  depends on  $x$  only through the factor  $(x - E[X | X\beta_0 > -u])$ . Since  $y > 0$  is the same as  $x\beta_0 > -\varepsilon$ , we have

$$E[\mu(\varepsilon, X) \eta(\varepsilon) 1(Y > 0) | \varepsilon] = 0 \tag{A.6}$$

for any function  $\eta(\varepsilon)$ . Next, the expected value over  $x$  (conditional on  $\varepsilon$ ) of the integral

$$\int_0^\infty dv \mu(v + u, x) \eta(v) = \int dv 1(v > u) \mu(v, x) \eta(v)$$

has the form

$$\int dx h(x) \int dv \mu(v, x) \eta(v) [1(x\beta_0 > -\varepsilon) 1(v > \varepsilon) + 1(x\beta_0 < -\varepsilon) 1(v > -x\beta_0)]$$

Rearranging the indicator functions as  $1(v > \varepsilon) 1(x\beta_0 > -v)$  and reversing the order of integration (the trimmed functions will have finite support) gives

$$\int_\varepsilon^\infty dv \int dx h(x) \mu(v, x) \eta(v) 1(x\beta_0 > -v) \tag{A.7}$$

As before the integral over  $x$  gives zero, so that

$$E[\int dv 1(v > U) \mu(v, X) \eta(v) | \varepsilon] = 0 \quad (\text{A.8})$$

for any function  $\eta(\varepsilon)$ . The proposition then follows from (A.6) with  $\eta = G\tau$  and (A.8) with  $\eta = g\tau$ .

*Proof of Proposition 2(a)*

Under Assumptions 8–10 and Condition 1, we can apply Lemma B.1 of Ai (1997) to show that the following convergence bounds hold uniformly in  $\beta \in B$  and  $(x, \varepsilon) \in W_n$ , with  $u(\beta) = y - x\beta$ . If we write  $\tilde{g}(u | \beta) = \sum_j a_j(u, \beta)$ , then  $|a_j| < c h_n^{-1}$ ,  $|\partial a_j / \partial u| < c h_n^{-2}$ ,  $|\partial a_j / \partial \beta| < c h_n^{-2} |x_j|$ , which is integrable, and  $E[a_j^2] < c h_n^{-1}$ . It follows from the lemma that<sup>13</sup>

$$|\tilde{g}(u(\beta) | \beta) - E[\tilde{g}(u(\beta) | \beta)]| = O(h_n^{-1/2}) o_p(n^{-1/2+}), \quad (\text{A.9})$$

uniformly in  $\beta$  and  $u$ . Similarly,

$$\begin{aligned} |\tilde{G}(u(\beta) | \beta) - E[\tilde{G}(u(\beta) | \beta)]| &= o_p(n^{-1/2+}) \\ \left| \frac{d\tilde{g}(u(\beta) | \beta)}{d\beta} - E\left(\frac{d\tilde{g}(u(\beta) | \beta)}{d\beta}\right) \right| &= (c + |x|) O(h_n^{-3/2}) o_p(n^{-1/2+}) \\ \left| \frac{d\tilde{G}(u(\beta) | \beta)}{d\beta} - E\left(\frac{d\tilde{G}(u(\beta) | \beta)}{d\beta}\right) \right| &= (c + |x|) O(h_n^{-1/2}) o_p(n^{-1/2+}) \end{aligned}$$

The bias terms such as  $|E[\tilde{g}] - g|$  are  $O(h_n^2)$  for  $\tilde{g}$  and  $\tilde{G}$ , and  $(c + x) O(h_n^2)$  for the derivatives. Because of the rate of convergence of  $h_n$  given in Condition 2, the bias terms dominate except in the case of  $d\tilde{g}/d\beta$ . Thus we have

$$|\Delta\tilde{g}(u(\beta) | \beta)| \equiv |\tilde{g}(u(\beta) | \beta) - g(u(\beta) | \beta)| = O(h_n^2) \quad (\text{A.10})$$

and similarly

$$\begin{aligned} |\Delta\tilde{G}(u(\beta) | \beta)| &= O(h_n^2) \\ |\Delta d\tilde{g}(u(\beta) | \beta) / d\beta| &= O(h_n^{-3/2}) o_p(n^{-1/2+}) + O(h_n^2) \\ |d\Delta\tilde{G}(u(\beta) | \beta) / d\beta| &= O(h_n^2) \end{aligned}$$

Using the first-order expansion

$$\tau_1(\tilde{g} \tilde{G}^2) - \tau_1(g G^2) = \tau_1'(\tilde{\gamma}) (\tilde{g} \tilde{G}^2 - g G^2)$$

---

<sup>13</sup> The notation  $n^{p+}$  means any power of  $n$  greater than  $p$ .

where  $\tau'_1(\gamma) = d\tau_1(\gamma)/d\gamma$  and  $\tilde{\gamma}$  is between  $\tilde{g}G^2$  and  $gG^2$ , and making repeated application of identities such as

$$\tilde{G} \frac{d\tilde{g}}{d\beta} - G \frac{dg}{d\beta} = \Delta\tilde{G} \cdot \Delta \frac{d\tilde{g}}{d\beta} + G \Delta \frac{d\tilde{g}}{d\beta} + \Delta\tilde{G} \cdot \frac{dg}{d\beta},$$

we can express  $\Delta\tilde{m}_1^* \equiv \tilde{m}_1^* - m_1^*$  and  $\Delta\tilde{m}_2^*$  as sums of terms, each of which is a product of (i) approximation errors like (A.10) with known rates of convergence and (ii) factors like  $\tau_{1n}$  and  $dg/d\beta$  with known upper bounds. By inspection, the term with the slowest rate of convergence in  $\Delta\tilde{m}_1^*$  is  $G^2 \Delta(d\tilde{g}/d\beta)\tau$  (taking into account the assumption that  $b_n$  converges to zero at a slower rate than  $h_n$ ), and therefore

$$|\Delta\tilde{m}_1^*| = (c + |x|) O(b_n^{-1}) [O(h_n^{-3/2}) o_p(n^{-1/2+}) + O(h_n^2)] \quad (\text{A.11})$$

uniformly in  $u$  and  $\beta$ . The same bound also holds for  $|\Delta\tilde{m}_2^*|$ . Since  $\Delta\tilde{m}_2^*(u, \beta)$  has support in  $u$  on an interval of order  $u_n$ ,

$$|\Delta\tilde{s}^*| = (c + |x|) O(u_n b_n^{-1}) [O(h_n^{-3/2}) o_p(n^{-1/2+}) + O(h_n^2)]$$

and then because  $n^{-1} \sum_i |x_i| \xrightarrow{P} E[|x|] < \infty$ , we get

$$n^{-1} (\tilde{S}_n^*(\beta) - S_n^*(\beta)) = O(u_n b_n^{-1}) [O(h_n^{-3/2}) o_p(n^{-1/2+}) + O(h_n^2)] \quad (\text{A.12})$$

According to Condition 2, the right-hand side converges to zero.<sup>14</sup>

Next, using  $|m_1^*| \leq |m_1|$  and  $|m_2^*| \leq |m_2|$  and the assumption that  $s(\beta)$  is bounded uniformly in  $\beta$  by an integrable function, we have

$$n^{-1} (S_n^*(\beta) - S_n(\beta)) \xrightarrow{P} E[s^*(\beta) - s(\beta)]$$

uniformly in  $\beta$  (the rate no longer matters at this point). Since  $m_2^*(v, x, \beta) \rightarrow m_2(v, x, \beta)$  pointwise in  $v$ , and since both the integrand and its derivative with respect to  $\beta$  in the following expression are by assumption bounded uniformly in  $\beta$  by integrable functions (see Assumption 6), we can use the bounded convergence theorem to get

$$\int_0^\infty dv m_2^*(u + v, x, \beta) \rightarrow \int_0^\infty dv m_2(u + v, x, \beta) \quad (\text{A.13})$$

uniformly in  $\beta$ . Then, with  $m_1^*(\epsilon, x, \beta) \rightarrow m_1(\epsilon, x, \beta)$ , a further application of bounded convergence gives  $E[s^*(\beta) - s(\beta)] \rightarrow 0$ . Finally, (A.5) shows that the trimming

---

<sup>14</sup> The bound on the integral in  $\Delta\tilde{s}_n^*$  may not be the best that could be obtained, because in general the integral of a kernel estimator will converge faster (by a factor of  $h_n$ ) than the kernel estimator itself. However, to take advantage of this one would have to establish a suitable lower bound on the rest of the integrand.

restriction  $(x_i, \varepsilon_i) \in W_n, i = 1, \dots, n$  (which was implicitly imposed up to this point) holds with probability approaching 1, and so can be dropped when considering convergence in probability. Thus

$$n^{-1} \left( \tilde{S}_n^*(\beta) - S_n(\beta) \right) \xrightarrow{P} 0$$

uniformly in  $\beta$ , as required.

*Proof of Proposition 2(b)*

This follows along the same lines as the proof of Proposition 2(a). The derivatives  $d\tilde{m}_1^*/d\beta$  and  $d\tilde{m}_2^*/d\beta$  contain the kernel estimators  $d^2\tilde{g}/d\beta d\beta'$  and  $d^2\tilde{G}/d\beta d\beta'$ . As before, these new kernel estimators can be shown to satisfy uniform bounds,

$$\begin{aligned} |\Delta(d^2\tilde{g}/d\beta d\beta')| &= (c + |x|^2) O(b_n^{-1} h_n^{-5/2}) o_p(n^{-1/2+}) \\ |\Delta(d^2\tilde{G}/d\beta d\beta')| &= (c + |x|^2) O(b_n^{-1} h_n^{-3/2}) o_p(n^{-1/2+}) \end{aligned}$$

Expanding  $\Delta(d\tilde{m}_1^*/d\beta)$  and  $\Delta(d\tilde{m}_2^*/d\beta)$  as before, identifying the terms with the slowest rate of convergence as those involving  $\Delta(d^2\tilde{g}/d\beta d\beta')\tau$  (again taking into account the assumption that  $b_n$  decreases more slowly than  $h_n$ ), we get

$$|\Delta(d\tilde{s}^*/d\beta)| = (c + |x|^2) O(u_n b_n^{-1} h_n^{-5/2}) o_p(n^{-1/2+}) \quad (\text{A.14})$$

We have  $E[x^2] < \infty$  and  $u_n b_n^{-1} h_n^{-5/2} n^{-1/2+} \rightarrow 0$ . The rest of the proof is then the same as for Proposition 2(a), leading to

$$n^{-1} \left( d\tilde{S}_n^*(\beta)/d\beta - dS_n(\beta)/d\beta \right) \xrightarrow{P} 0$$

uniformly in  $\beta$ , as required.

*Proof of Proposition 2(c)*

We start by expanding  $\Delta\tilde{m}_1^*$  and  $\Delta\tilde{m}_2^*$  in the kernel approximation errors  $\Delta\tilde{g}$ ,  $\Delta\tilde{G}$ , etc., as in the proof of Proposition 2(a). Since we will need to isolate terms that are linear in the kernel estimates  $\tilde{g}$ ,  $\tilde{G}$ , etc., we have to carry the expansion of  $\tau_1$  to second order this time:

$$\tau_1(\tilde{g} \tilde{G}^2) - \tau_1(g G^2) = \tau_1'(g G^2) (\tilde{g} \tilde{G}^2 - g G^2) + \tau_1''(\tilde{\gamma}) (\tilde{g} \tilde{G}^2 - g G^2)^2$$

Let

$$\Delta\tilde{s}^* = \Delta\tilde{s}_L^* + \Delta\tilde{s}_R^*$$

where  $\Delta\tilde{S}_L^*$  contains terms linear in kernel approximation errors, while  $\Delta\tilde{S}_R^*$  contains terms that are second order or higher in kernel approximation errors. ( $L$  stands for “linear”, and  $R$  for “remainder”.)

First consider  $\Delta\tilde{S}_R^*$ . The uniform convergence results used in the proof of Proposition 2(a) for general  $\beta$  of course still hold at  $\beta_0$ . The slowest rates of convergence come from (i) terms of the form  $G^4\Delta(d\tilde{g}/d\beta)\Delta\tilde{g}\cdot\tau'_1$  in  $\Delta\tilde{m}_{1R}^*$  and  $gG^3\Delta(d\tilde{g}/d\beta)\Delta\tilde{g}\cdot\tau'_1$  in  $\Delta\tilde{m}_{2R}^*$ , bounded by  $(c+|x|)O(b_n^{-2})[O(h_n^{1/2})o_p(n^{-1/2+})+O(h_n^4)]$ , and (ii) terms of the form  $G^4(\Delta\tilde{g})^2\cdot\tau'_1$  in  $\Delta\tilde{m}_{1R}^*$  and  $\Delta\tilde{m}_{2R}^*$ , bounded by  $O(b_n^{-3}h_n^4)$ . Then, as in the proof of Proposition 2(a),

$$n^{-1/2}\left(\tilde{S}_{R,n}^*(\beta)-S_{R,n}^*(\beta)\right)=O(u_nb_n^{-2}h_n^{1/2})o_p(n^+)+O(u_nb_n^{-3}h_n^4n^{1/2}), \quad (\text{A.15})$$

and according to Condition 2, the right-hand side converges to zero.

Next, consider  $\Delta\tilde{S}_L^*$ . The linear terms coming from  $\Delta\tilde{m}_1^*(\varepsilon_i, x_i, \beta_0)$  are

$$\begin{aligned} \Delta\tilde{m}_{1L}^* = & \left\{ \left( G^2\Delta(d\tilde{g}/d\beta) + 2G\Delta\tilde{G}\cdot dg/d\beta - gG\Delta(d\tilde{G}/d\beta) \right. \right. \\ & \left. \left. - g\Delta\tilde{G}\cdot dG/d\beta - G\Delta\tilde{g}\cdot dG/d\beta \right) \tau'_1 \right. \\ & \left. + G^2(Gdg/d\beta - g dG/d\beta)(G\Delta\tilde{g} + 2g\Delta\tilde{G})\tau'_1 \right\} \tau_2 1(x_i\beta_0 > -\varepsilon_i) \end{aligned} \quad (\text{A.16})$$

while the linear terms from  $\Delta\tilde{m}_2^*(v+u_i, x_i, \beta_0)$  (where  $v$  is the integration variable) are

$$\begin{aligned} \Delta\tilde{m}_{2L}^* = & \left\{ \left( gG\Delta(d\tilde{g}/d\beta) + G\Delta\tilde{g}\cdot dg/d\beta + g\Delta\tilde{G}\cdot dg/d\beta \right. \right. \\ & \left. \left. - g^2\Delta(d\tilde{G}/d\beta) - 2g\Delta\tilde{g}\cdot dG/d\beta \right) \tau_1 \right. \\ & \left. + gG(Gdg/d\beta - g dG/d\beta)(G\Delta\tilde{g} + 2g\Delta\tilde{G})\tau'_1 \right\} \tau_2 \end{aligned} \quad (\text{A.17})$$

Substituting the expressions for the kernel estimators given by (2.12) and (2.19)-(2.20), we can write the kernel approximation error in the linear part of the trimmed score functions as a double sum of the form

$$n^{-1/2}\left(\tilde{S}_{L,n}^*(\beta_0)-S_{L,n}^*(\beta_0)\right)=n^{-3/2}\sum_i\sum_{j\neq i}\psi_n(\varepsilon_i, x_i, \varepsilon_j, x_j) \quad (\text{A.18})$$

where, for example, the contribution to  $\psi_n$  of the first term in (A.16) is

$$\begin{aligned} & \left\{ h_n^{-2}(x_j-x_i)K'\left((\varepsilon_i-\varepsilon_j)/h_n\right)1(x_j\beta_0 > -\varepsilon_j) - dg(e_i(\beta_0)|\beta_0)/d\beta \right\} \\ & \cdot G(\varepsilon_i|\beta_0)^2\tau_1[g(\varepsilon_i|\beta_0)G(\varepsilon_i|\beta_0)^2]\tau_2(\varepsilon_i) \end{aligned}$$

The summation on the right-hand side of (A.18) can be symmetrized in the indices  $i$  and  $j$  to make a  $U$ -statistic. We can then apply Lemma 3.1 of Powell et al. (1989), which generalizes the usual projection theorem for  $U$ -statistics to the case where individual terms in the sum can grow with  $n$ , provided that they are  $o(n)$  in mean square. In the

present case we have  $E[\Psi_n^2] = O(u_n^2 b_n^{-2})[O(h_n^{-3} n^{-1}) + O(h_n^4)]$ , and the required rate of convergence follows from Condition 2. The projection theorem then gives

$$\begin{aligned} & n^{-3/2} \sum_i \sum_{j \neq i} \Psi_n(\varepsilon_i, x_i, \varepsilon_j, x_j) \\ &= n^{1/2} E[\Psi_n] + 2n^{-1/2} \sum_{i=1}^n \{r_n(\varepsilon_i, x_i) - E[\Psi_n]\} + o_p(1) \end{aligned} \quad (\text{A.19})$$

where

$$r_n(\varepsilon, x) = \{E[\Psi_n(\varepsilon, x, \varepsilon_j, x_j) | \varepsilon, x] + E[\Psi_n(\varepsilon_i, x_i, \varepsilon, x) | \varepsilon, x]\} / 2 \quad (\text{A.20})$$

First, taking expectations over  $(\varepsilon_j, x_j)$  conditional on  $(\varepsilon_i, x_i)$ , we have the bias term  $E[\Delta \tilde{g}(\varepsilon_i, \beta) | \varepsilon_i, x_i] = O(h_n^2)$ , and similarly for the other kernel estimators. It follows that

$$E[\Psi_n(\varepsilon_i, x_i, \varepsilon_j, x_j) | \varepsilon_i, x_i] = O(u_n b_n^{-2} h_n^2). \quad (\text{A.21})$$

Next, we take expectations over  $(\varepsilon_i, x_i)$  conditional on  $(\varepsilon_j, x_j)$ . Following the steps leading to (A.7) for a generic function  $\mu(v, x)$ , we have

$$\begin{aligned} & E\left[\int_0^\infty dv \mu(v+U, X)\right] \\ &= \int d\varepsilon f(\varepsilon) \int dx h(x) \int dv \mu(v, x) [1(x\beta_0 > -\varepsilon) 1(v > \varepsilon) + 1(x\beta_0 < -\varepsilon) 1(v > -x\beta_0)] \\ &= \int dv F(v) \int dx h(x) \mu(v, x) 1(x\beta_0 > -v) \end{aligned}$$

This can be used to evaluate the expectations of the terms coming from the integral of  $\Delta \tilde{m}_{2L}$ . This is where the specific functional form of the efficient score comes into play, as well as the use of a common trimmed denominator for all its components. After some calculations, we arrive at the nice result

$$E[\Psi_n(\varepsilon_i, x_i, \varepsilon_j, x_j) | \varepsilon_j, x_j] = 0. \quad (\text{A.22})$$

and therefore also  $E[\Psi_n] = 0$ . The right hand side of (A.19) now reduces to  $n^{-1/2}$  times a sum of independent terms with mean zero and bounded by (A.21), so application of a central limit theorem gives

$$n^{-1/2} \left( \tilde{S}_{L,n}^*(\beta_0) - S_{L,n}^*(\beta_0) \right) = O(u_n b_n^{-2} h_n^2) \quad (\text{A.23})$$

which tends to zero by Condition 2.

Finally, we have to show that the trimmed and untrimmed score functions  $S_n^*(\beta_0)$  and  $S_n(\beta_0)$  are asymptotically equivalent. Since they are both sums of i.i.d. terms with mean zero, we can use mean square convergence. We have

$$\mathbb{E} \left[ n^{-1} \left| S_n^*(\beta_0) - S_n(\beta_0) \right|^2 \right] = \mathbb{E} \left[ \left| s^*(\beta_0) - s(\beta_0) \right|^2 \right] \quad (\text{A.24})$$

with

$$s^*(\beta_0) - s(\beta_0) \equiv \Delta s(\beta_0) = \Delta m_1(\varepsilon, x, \beta_0) 1(y > 0) - \int_u^\infty dv \Delta m_2(v, x, \beta_0) \quad (\text{A.25})$$

The differences  $\Delta m_1 = m_1^* - m_1$  and  $\Delta m_2 = m_2^* - m_2$  are related by  $\Delta m_2(v, x, \beta_0) = [f(v)/F(v)] \Delta m_1(v, x, \beta_0)$  because of the common trimmed denominator. Evaluating (A.24), and using integration by parts to eliminate the cross product between the two terms on the right-hand side of (A.25), then gives

$$\mathbb{E} [ |\Delta m_1(\varepsilon, x, \beta_0)|^2 1(y > 0) ] + [R(\varepsilon)]_{-\infty}^\infty \quad (\text{A.26})$$

where

$$R(\varepsilon) = F(\varepsilon) \int dx h(x) 1(x\beta_0 > -\varepsilon) \left| \int dv f(v) 1(v > \varepsilon) \Delta m_1(v, x, \beta_0) / F(v) \right|^2$$

The remainder term in (A.26) evidently vanishes provided that the integrals exist. To bound  $R(\varepsilon)$ , apply the Cauchy-Schwarz inequality,

$$\begin{aligned} & \left| \int dv f(v) 1(v > \varepsilon) \Delta m_1(v, x, \beta_0) / F(v) \right|^2 \\ & \leq [F(\varepsilon)^{-1} - 1] \int dv f(v) 1(v > \varepsilon) |\Delta m_1(v, x, \beta_0)|^2 \end{aligned}$$

and then use the inequalities  $|\Delta m_1| \leq |m_1|$  and  $1(x\beta_0 > -\varepsilon) 1(v > \varepsilon) \geq 1(x\beta_0 > -v)$  to get

$$R(\varepsilon) \leq \int dx h(x) \int dv f(v) 1(x\beta_0 > -v) |m_1(v, x, \beta_0)|^2 = \text{tr } I_* < \infty.$$

It follows, by bounded convergence as  $\varepsilon \rightarrow \pm\infty$ , that  $R(\infty) = R(-\infty) = 0$ . That leaves the first term in (A.26), where  $\Delta m_1 \rightarrow 0$  pointwise. The bounded convergence theorem again applies because  $|\Delta m_1| \leq |m_1|$  and  $\mathbb{E}[|m_1(\varepsilon, x, \beta_0)|^2 1(y > 0)] = \text{tr } I_* < \infty$ . Therefore,

$$\mathbb{E} \left[ \left| s^*(\beta_0) - s(\beta_0) \right|^2 \right] \rightarrow 0 \quad (\text{A.27})$$

From (A.24) and (A.27),

$$n^{-1/2} \left( S_n^*(\beta_0) - S_n(\beta_0) \right) \xrightarrow{P} 0$$

Together with (A.15) and (A.23), this gives the required result

$$n^{-1/2} \left( \tilde{S}_n^*(\beta_0) - S_n(\beta_0) \right) \xrightarrow{P} 0$$

### A.3. Estimation and Asymptotic Efficiency when $\beta_0 = 0$

The derivation given above has to be modified when  $\beta_0 = 0$  because the conditional density function  $h(u | \beta)$  is singular at  $(u, \beta) = (0, 0)$ , and therefore the derivatives of  $g(u | \beta)$  and  $G(u | \beta)$  are not well defined at that point. However, the other key points of the derivation are not affected when  $\beta_0 = 0$ : the kernel estimates converge uniformly to their expected values at the same rates as before, and the conditional expected value in (A.22) is still zero.

First, the estimator is modified by incorporating the additional trimming factor  $\tau_3(u)$ , defined by (3.14), in the terms  $\tilde{m}_1^*(u, x, \beta)$  and  $\tilde{m}_2^*(u, x, \beta)$  of the trimmed score function. This excludes  $u$  from a neighborhood  $U_n = \{u : |u| < \mu_n\}$ , where  $\mu_n \rightarrow 0$  with  $\mu_n / h_n \rightarrow \infty$ .

Secondly, in the derivation of uniform convergence,  $\beta$  is restricted to a neighborhood  $B_{0,n} = \{\beta : |\beta| < a_n\}$ , where  $a_n \rightarrow 0$  with  $a_n n^{1/2} \rightarrow \infty$ . (This of course relies on an initial  $\sqrt{n}$ -consistent estimator of  $\beta$ .) This allows  $\tilde{g}(u | \beta)$  to be considered as an estimator of  $g(u | 0)$  rather than  $g(u | \beta)$ .

Define the following functions to replace the derivatives  $dg/d\beta$  and  $d^2g/d\beta d\beta'$

$$\dot{g}(u | \beta) = E[(X - x) f'(u + X\beta) 1(u + X\beta > 0)] \quad (\text{A.28})$$

$$\ddot{g}(u | \beta) = E[(X - x)(X - x)' f''(u + X\beta) 1(u + X\beta > 0)] \quad (\text{A.29})$$

and similarly for  $\dot{G}(u | \beta)$  and  $\ddot{G}(u | \beta)$ . Note that when the limit is taken as  $n \rightarrow \infty$  for fixed  $u$  with  $u \neq 0$ ,

$$\dot{g}(u | 0) = \lim E[d\tilde{g}(u | 0)/d\beta]$$

and similarly for the other derivative terms. The modified score function  $S_{0,n}(\beta) = \sum_i s_0(x_i, y_i, \beta)$  is then defined as in (3.5)–(3.7) but with the moment functions

$$m_{0,1}(u, x, 0) = [\dot{g}(u | 0)/g(u | 0) - \dot{G}(u | 0)/G(u | 0)] 1(u > 0) \quad (\text{A.30})$$

$$m_{0,2}(u, x, 0) = [g(u | 0)/G(u | 0)] m_{0,1}(u, x). \quad (\text{A.31})$$

The modified hessian matrix  $\dot{S}_{0,n}(\beta)$ , corresponding to  $dS_n(\beta)/d\beta$ , is defined similarly with

$$\dot{m}_{0,1}(u, x, 0) = \left( \frac{\ddot{g}(u | 0)}{g(u | 0)} - \frac{\ddot{G}(u | 0)}{G(u | 0)} - \frac{\dot{g}(u | 0)^2}{g(u | 0)^2} + \frac{\dot{G}(u | 0)^2}{G(u | 0)^2} \right) 1(u > 0) \quad (\text{A.32})$$

$$\dot{m}_{0,2}(u, x, 0) = [g(u | 0)/G(u | 0)] [\dot{m}_{0,1}(u, x) + m_{0,1}(u, x)^2]. \quad (\text{A.33})$$

The trimmed versions  $S_{0,n}^*(\beta)$  and  $\dot{S}_{0,n}^*(\beta)$  incorporate the trimming factor

$$\tau(u | 0) = \tau_1[D(u | 0)] \tau_2(u) \tau_3(u).$$

Evaluating the modified score  $S_{0,n}(\beta)$  at  $\beta = 0$ , we find

$$s_0(x, y, 0) = (1(y > 0) f'(y) / f(y) + 1(y \leq 0) f(0) / F(0)) (E[X] - x) \quad (\text{A.34})$$

This is the same as the efficient semiparametric score for the case  $\beta_0 = 0$ . Its variance is

$$V_{0*}^{-1} = \text{var}[X] \int_0^\infty du f(u) \left[ \frac{d}{du} \log \left( \frac{f(u)}{F(u)} \right) \right]^2 = \text{var}[X] \left\{ \int_0^\infty du \frac{[f'(u)]^2}{f(u)} + \frac{[f(0)]^2}{F(0)} \right\} \quad (\text{A.35})$$

where  $V_{0*}$  denotes the semiparametric efficiency bound for this case. (The efficient semiparametric score (A.34) is in fact the limit of (3.8) as  $\beta_0 \rightarrow 0$ , but it is safer to start with  $\beta_0 = 0$  and derive it via the relevant orthogonality conditions for the ‘‘worst-case’’ direction of approach.) Finally, evaluating the modified hessian matrix at  $\beta = 0$ , we find

$$E[\dot{s}_0(X, Y, 0)] = -V_{0*}^{-1} \quad (\text{A.36})$$

The random variable  $\bar{\beta}_0 = -[\dot{S}_{0,n}(0)]^{-1} S_{0,n}(0)$  is  $O_p(n^{-1/2})$  (and therefore  $\bar{\beta}_0 \in B_n$  with probability approaching 1) and is asymptotically normal with asymptotic variance

$$E[\dot{s}_0(X, Y, 0)]^{-1} \text{var}[s_0(X, Y, 0)] E[\dot{s}_0(X, Y, 0)]^{-1} = V_{0*}$$

The next result shows that  $\hat{\beta}$  and  $\bar{\beta}_0$  are asymptotically equivalent.

*Proposition 3.* Suppose that Assumptions 1–4, 6–10 and Conditions 1–4 hold, as given in Appendix A.1, and let the trimming factor be  $\tau(u | \beta) = \tau_1[D(u | \beta)] \tau_2(u) \tau_3(u)$ , as defined by (3.12)–(3.14). Then if  $\beta_0 = 0$ , the results of Proposition 2 hold with  $B$ ,  $S_n(\beta)$  and  $dS_n(\beta) / d\beta$  replaced by  $B_{0,n}$ ,  $S_{0,n}(\beta)$  and  $\dot{S}_{0,n}(\beta)$ .

It follows from the usual first-order series expansion of  $\tilde{S}_n^*(\beta)$  in  $(\beta - \beta_0)$  that (i) there is an  $O_p(n^{-1/2})$  consistent root (as previously defined), and (ii) if we identify  $\hat{\beta}$  with the consistent root, then  $\hat{\beta}$  has the same asymptotic distribution as  $\bar{\beta}_0$ , i.e.,  $\hat{\beta}$  achieves the semiparametric efficiency bound for  $\beta_0 = 0$ .

The following result is used in the proof of Proposition 3.

*Proposition B.* Suppose that (i)  $h_n = o(u_n)$  and  $\beta_n = o(u_n)$ , (ii)  $E[|x|^p] < \infty$ , and (iii)  $\kappa(\cdot)$  is a bounded function such that  $\kappa(z) = O(|z|^{-\nu})$  for all  $\nu > 0$  as  $z \rightarrow \pm\infty$ . Then for  $r < p$ ,

$$\int dx h(x) |x|^r \kappa([u_n + x\beta_n] / h_n) = O(|\beta_n / u_n|^{p-r}). \quad (\text{A.37})$$

*Proof.* First consider the integration region where  $|x\beta| \geq \frac{1}{2}|u|$ :

$$\begin{aligned} & \int dx h(x) |x|^r \cdot |\kappa([u_n + x\beta_n]/h_n)| \cdot 1(|x\beta_n| \geq \frac{1}{2}|u_n|) \\ & \leq c \int dx h(x) |x|^r 1(|x| \geq \frac{1}{2}|u|/|\beta_n|) \leq c |\beta_n/u|^{p-r} E[|x|^p] \leq c |\beta_n/u_n|^{p-r} \end{aligned} \quad (\text{A.38})$$

(where  $c$  represents a generic constant). In the remaining integration region,  $|x\beta| < \frac{1}{2}|u|$  implies  $|u + x\beta| > \frac{1}{2}|u|$ , and therefore

$$\begin{aligned} & \int dx h(x) |x|^r \cdot |\kappa([u_n + x\beta_n]/h_n)| \cdot 1(|x\beta_n| < \frac{1}{2}|u_n|) \\ & \leq E[|x|^r] O(|h_n/u_n|^v) = O(|h_n/u_n|^v) \end{aligned} \quad (\text{A.39})$$

For large enough  $v$  (depending on the rates of convergence of  $\beta_n$  and  $u_n$ ) (A.39) becomes negligible in comparison with (A.38), and the proposition follows.

### *Proof of Proposition 3*

We have to show that  $\tilde{S}_n^*(\beta)$  and  $d\tilde{S}_n^*(\beta)/d\beta'$  converge uniformly to  $S_{0,n}^*(\beta)$  and  $\dot{S}_{n,0}^*(\beta)$  at the same rates as in the proof of Proposition 2 (Appendix A.2). Since the uniform convergence of the kernel estimates to their expected values is not affected by the value of  $\beta_0$ , the essential step is to bound the ‘‘bias’’ terms, such as

$$E[d\tilde{g}(u(\beta)|\beta)/d\beta - \dot{g}(u(\beta)|0)]. \quad (\text{A.40})$$

by  $O(h_n^2)$ . However, the second derivative terms occur only in the hessian, and therefore need only be bounded by  $o(h_n^{-5/2} n^{-1/2+})$  according to (A.14).

Consider (A.40) as a typical bias term, with

$$E[d\tilde{g}_n(u|\beta)/d\beta] = h_n^{-1} E[(X-x) \int dw 1(wh_n < u + X\beta) K'(w) f(u + X\beta - wh_n)].$$

First integrate by parts with respect to  $w$  to isolate the term in  $h_n^{-1}$ , and then in the remaining term expand  $f'(u + X\beta - wh_n)$  to second order in powers of  $h_n$ . Then (A.40) can be written as

$$\begin{aligned} & \dot{g}(u|\beta) - \dot{g}(u|0) + h_n^{-1} f(0) E[(X-x) K([u + X\beta]/h_n)] \\ & + E[(X-x) f'(u + X\beta) \{ \bar{K}([u + X\beta]/h_n) - 1(u + X\beta > 0) \}] \\ & + E[(X-x) f''(u + X\beta) \bar{K}_1([u + X\beta]/h_n)] + O(h_n^2) \end{aligned} \quad (\text{A.41})$$

with

$$\begin{aligned} & |\dot{g}(u|\beta) - \dot{g}(u|0)| \leq E[|(X-x) \{ f'(u + X\beta) 1(u + X\beta > 0) - f'(u) 1(u > 0) \}|] \\ & \leq (c + |x|) \{ O(|\beta|) + \Pr(|X\beta| > |u|) \} \leq (c + |x|) \{ O(|\beta|) + O(|\beta/u|^{p-1}) \} \end{aligned}$$

The last inequality is derived in the same way as (A.38), assuming that  $E[|x|^p] < \infty$ . The remaining terms in (A.41) are bounded using (A.37), and the overall bound on the bias term (A.40) (with  $p = 4$ ) has the form

$$(c + |x|) \{O(a_n) + O(h_n^{-1} a_n^3 \mu_n^{-3}) + O(h_n^2)\}$$

According to Condition 2,  $h_n \gg n^{-1/5}$ , so we can choose  $a_n = O(h_n^2)$  and still have  $a_n \gg n^{-1/2}$ . Then, with  $h_n = o(\mu_n)$  from Condition 4, the overall bound on (A.40) becomes  $(c + |x|) O(h_n^2)$  as required.

The remaining bias terms in  $\tilde{S}_n^*(\beta)$  are bounded similarly, while the bias associated with the second derivative terms in  $d\tilde{S}_n^*(\beta)/d\beta'$  has the required bound

$$(c + |x|^2) \{O(a_n) + O(h_n^{-2} a_n^2 \mu_n^{-2}) + O(h_n^2)\} = (c + |x|^2) o(h_n^{-5/2} n^{-1/2+})$$

under the same assumptions, except that this time we need  $a_n = O(h_n^{3/4} n^{-1/4})$ .

The rest of the proof is the same as for Proposition 2, except the limit functions are now  $S_{0,n}(\beta)$  and  $\dot{S}_{0,n}(\beta)$  instead of  $S_n(\beta)$  and  $dS_n(\beta)/d\beta$ .

#### A.4. Random censoring

Linear regression with random censoring is closely related to the tobit model. In this case we have  $y_i = \max(c_i, x_i\beta + \varepsilon_i)$ , where the censoring points  $c_i$  are i.i.d. with density function  $p_c(c)$  and (in the case considered here) are independent of  $x$  and  $\varepsilon$ .<sup>15</sup>

Following the same procedure as in Section 2, the kernel estimators  $\tilde{g}(u|\beta)$  and  $\tilde{G}(u|\beta)$ , the estimated likelihood function  $\tilde{L}_n(\beta)$ , and the score function  $\tilde{S}_n(\beta)$  are the same as in (2.12)–(2.15), except that the indicator function is changed to  $1(y_i > c_i)$ . The underlying functions  $g(u|\beta)$  and  $G(u|\beta)$  are now defined by

$$g(u|\beta) = \int dx h(x) f(u + x[\beta - \beta_0]) P_c(x\beta + u) \quad (\text{A.42})$$

$$G(u|\beta) = \int dx h(x) \{1 - F(u + x[\beta - \beta_0])\} P_c(x\beta + u) \quad (\text{A.43})$$

where  $P_c$  is the distribution function of  $c$ . These versions of the functions are easier to work with because the integrands are differentiable.

As before, (A.42) and (A.43) can be used to define an artificial likelihood function  $\bar{L}_n(\beta)$  and score function  $S_n(\beta)$  as in (3.3) and (3.5). Evaluating the score function at  $\beta = \beta_0$  gives

---

<sup>15</sup> Since the estimator in Section 3 was derived for the tobit model, we continue here with left censoring. This can easily be translated into the more usual case of random right-censoring.

$$\begin{aligned}
s(x, y, \beta_0) = & -\frac{d}{du} \left( \log \frac{f(u)}{F(u)} \right) (x - E[X | X\beta_0 + u > C]) 1(y > c) \\
& + \int_u^\infty dv \frac{d}{dv} \left( \frac{f(v)}{F(v)} \right) (x - E[X | X\beta_0 + v > C])
\end{aligned} \tag{A.44}$$

where  $u = y - x\beta_0$ . The asymptotic information matrix for  $\bar{L}_n(\beta)$  is the variance of (A.44), i.e.,

$$\int du f(u) \left[ \frac{d}{du} \log \left( \frac{f(u)}{F(u)} \right) \right]^2 \Pr\{X\beta_0 + u > C\} \text{var}[X | X\beta_0 + u > C] \tag{A.45}$$

which is equal to the inverse of the semiparametric efficiency bound for the model with random censoring. The proof that  $\hat{\beta}$  achieves this bound is the same as before. No special treatment is needed in the case  $\beta_0 = 0$ .

#### A.5. Efficient semiparametric estimation of truncated regression

The same technique can be used to construct an efficient semiparametric likelihood-based estimator of the truncated regression model. As before,  $y = x\beta_0 + \varepsilon$ , but now the joint density of  $(\varepsilon, x)$  is

$$f(\varepsilon) h(x) 1(x\beta_0 + \varepsilon > 0) / \{ \iint d\varepsilon' dx' f(\varepsilon') h(x') 1(x'\beta_0 + \varepsilon' > 0) \}$$

Without loss this can be written in term of  $h_*(x)$ , the marginal density of  $x$  in the observable population (after truncation), as<sup>16</sup>

$$\{ f(\varepsilon) 1(\varepsilon > -x\beta_0) / [1 - F(-x\beta_0)] \} h_*(x)$$

The nonparametric maximum likelihood estimator of  $F(\varepsilon | \beta)$ , the distribution function of the residuals, is given by the following equation corresponding to (2.3) for the censored model,

$$\sum_j 1(-x_j\beta \leq \varepsilon < e_j) = [1 - \hat{F}(\varepsilon | \beta)] \sum_j 1(\varepsilon \geq -x_j\beta) [1 - \hat{F}(-x_j\beta | \beta)]^{-1} \tag{A.46}$$

where  $e_j = y_j - x_j\beta$ . The solution of this equation is the Lynden-Bell product-limit estimator for truncated sampling, analogous to the Kaplan-Meier estimator for censored sampling (see, for example, Tsui, Jewell and Wu, 1988). The smoothed equation corresponding to (2.5) is

---

<sup>16</sup> The constraint that  $h_*(x) / [1 - F(-x\beta_0)]$  is integrable has no effect in finite samples and so is ignored.

$$\begin{aligned} & \sum_j \left\{ \bar{K} \left( \frac{\varepsilon + x_j \beta}{h_n} \right) - \bar{K} \left( \frac{\varepsilon - (y_j - x_j \beta)}{h_n} \right) \right\} \\ & = [1 - \tilde{F}(\varepsilon | \beta)] \frac{1}{h_n} \int_{-\infty}^{\varepsilon} dv \frac{1}{1 - \tilde{F}(v | \beta)} \sum_j K \left( \frac{v + x_j \beta}{h_n} \right) \end{aligned} \quad (\text{A.47})$$

with  $K$  and  $\bar{K}$  as before. This is a linear integral equation in  $[1 - \tilde{F}(\varepsilon | \beta)]^{-1}$ , and the solution is

$$\tilde{F}(\varepsilon | \beta) = 1 - \exp \left( - \int_{-\infty}^{\varepsilon} dv \tilde{g}(v | \beta) / \tilde{G}(v | \beta) \right) \quad (\text{A.48})$$

with new expressions for  $\tilde{g}$  and  $\tilde{G}$ ,

$$\tilde{g}(u | \beta) = \frac{1}{nh_n} \sum_j K \left( \frac{u - (y_j - x_j \beta)}{h_n} \right) \quad (\text{A.49})$$

$$\tilde{G}(u | \beta) = \frac{1}{n} \sum_j \left\{ \bar{K} \left( \frac{u + x_j \beta}{h_n} \right) - \bar{K} \left( \frac{u - (y_j - x_j \beta)}{h_n} \right) \right\} \quad (\text{A.50})$$

These are kernel estimators of the functions

$$g(u | \beta) = \int dx h_*(x) f(u + x[\beta - \beta_0]) [1 - F(-x\beta_0)]^{-1} 1(x\beta > -u) \quad (\text{A.51})$$

$$G(u | \beta) = \int dx h_*(x) \{1 - F(u + x[\beta - \beta_0])\} [1 - F(-x\beta_0)]^{-1} 1(x\beta > -u) \quad (\text{A.52})$$

The score  $S_n(\beta)$  and its kernel estimator  $\tilde{S}_n(\beta)$  are then calculated as before. Then  $s(\beta_0)$  is equal to the efficient score for semiparametric estimation of the truncated regression model,

$$\begin{aligned} s(x, y, \beta_0) &= - \frac{d}{du} \left( \log \frac{f(u)}{1 - F(u)} \right) (x - E[X | X\beta_0 > -u]) \\ &+ \int_{-x\beta_0}^u dv \frac{d}{dv} \left( \frac{f(v)}{1 - F(v)} \right) (x - E[X | X\beta_0 > -v]) \end{aligned} \quad (\text{A.53})$$

with  $u = y - x\beta_0$ . Its variance is equal to the corresponding semiparametric information bound (under the assumption that the bound is finite)

$$\frac{1}{Q} \int du f(u) \left[ \frac{d}{du} \log \left( \frac{f(u)}{1 - F(u)} \right) \right]^2 [1 - H(u | \beta_0)] \text{var}[X | X\beta_0 > -u] \quad (\text{A.54})$$

where  $Q$  is the probability of selection into the truncated sample.<sup>17</sup>

## Appendix B

### *Simulation results*

Simulations were carried out to evaluate the performance of semiparametric maximum likelihood estimator for samples of moderate size ( $n = 100, 200, 400$ ) generated by the tobit model, and to compare it with the adaptive  $M$ -estimator of Kim and Lai (2000).

*Data generation.* The model is

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

with a scalar regressor  $x \sim N(0,1)$ , intercept  $\alpha_0 = 0$ , and slope parameter  $\beta_0 = 1$  or  $\beta_0 = 0$ . With a symmetric error distribution, this gives 50% censoring. (If the error variance can be normalized to 1, then the underlying regression equation with  $\beta_0 = 1$  has  $R^2 = 0.5$  in the absence of censoring.) Both  $x$  and  $u$  are randomly generated in each replication.

Results are given for the following error distributions: normal, logistic, Laplace,  $t(3)$ ,  $t(2)$ , Cauchy, and two normal mixtures with distribution functions of the form

$$p\Phi([z - \mu_1]/\sigma_1) + (1-p)\Phi([z - \mu_2]/\sigma_2).$$

The first mixture has parameters  $p = 0.9$ ,  $\mu_1 = \mu_2 = 0$ ,  $\sigma_1 = 1/3$ ,  $\sigma_2 = 3$  (representing “contamination” by an error component with large variance), and the other has  $p = 0.5$ ,  $\mu_1 = -\mu_2 = 1/\sqrt{2}$ ,  $\sigma_1 = \sigma_2 = 1/\sqrt{2}$  (a bimodal distribution). Conventional scaling is used for the  $t(2)$  and Cauchy distributions, while the other distributions are normalized to zero mean and unit variance. Because the estimator does not restrict  $F$  to be symmetric or to have zero median, the intercept  $\alpha$  is not identified as a regression parameter; instead, it can be estimated as the median of  $\hat{F}$ .

*Implementation of the estimator.* The kernel  $K$  is standard normal. For the bandwidth  $h_n$ , we adopt the simple rule given by Silverman (1986, p. 47), setting  $h_n = c R n^{-a}$  where  $R$  is a preliminary estimate of the interquartile range of the error distribution and  $c$  is a suitably chosen constant.<sup>18</sup> (The interquartile range is used here because even when the variance exists, estimates of the error variance from censored data tend to be unreliable.) For estimation of  $\beta$ , the scale factor  $c = 1$  gives good results (in terms of

<sup>17</sup> The conditional variance and the distribution function  $H$  in (A.54) refer to the underlying (untruncated) population, for which the marginal density of  $x$  is  $h(x) = Q h_*(x) / [1 - F(-x\beta_0)]$ .

<sup>18</sup> The convergence rate  $a = 0.19$  was used, so as to be formally compatible with Condition 2.

mean square error) across the variety of different error distributions considered here; the results with  $c = 0.67$  and  $c = 1.5$  (not reported here) are similar. For estimating the quantiles of  $F$ , smaller bandwidths are better, and the quantile estimates reported here are for  $c = 0.67$  (the value suggested by Silverman for density estimation). Likelihood cross-validation and least-squares cross-validation methods were generally not effective in selecting a good bandwidth. Locally adaptive bandwidths are not used in this study because of their computational cost.

As found in other studies, the simulation results are generally insensitive to the choice of trimming parameters. Simulations were run for  $b_n = 0$  (no trimming),  $b_n = 0.0001$  and  $b_n = 0.001$ . An order-of-magnitude upper bound for  $b_n$  in small samples can be found by evaluating the denominator  $g(u|\beta)G(u|\beta)^2$  at some suitable trimming point in the left tail of the distribution of residuals. For the sample designs used here, values at the 5th (10th) percentile vary from 0.00015 (0.0010) for the bimodal normal mixture to 0.0005 (0.0026) for the Cauchy distribution. The results given here are for  $b_n = 0.001$ .

The initial values  $\tilde{\beta}$  are the least absolute deviations (LAD) tobit estimates (Powell, 1984). Because that estimator can have wide dispersion for small samples with 50% censoring (Paarsch, 1984), the initial estimator is restricted to  $|\tilde{\beta}| \leq 5$  in the simulations with  $\beta_0 = 1$ , and  $|\tilde{\beta}| \leq 2$  when  $\beta_0 = 0$ .

Modifications in the practical implementation of the estimator are: (i) to reduce the computational cost of the numerical quadrature, terms with  $i = j$  are retained in the kernel estimates, because then the integrand is the same function for all observations; (ii) to reduce search costs, the untrimmed estimator ( $b_n = 0$ ) is computed by maximizing the estimated log likelihood and (iii) the trimmed estimator ( $b_n > 0$ ) uses a two-step method with the untrimmed estimator as the starting point for minimization of the square of the trimmed score function; (iv) the trimming factors  $\tau_2(u)$  and  $\tau_3(u)$  are dropped, as is the convergence rate (slower than  $n^{-1/10}$ ) for the trimming parameter  $b_n$ . In the small number of simulations where no solution was found for the trimmed score equation,  $\hat{\beta}$  was computed by minimizing the square of the trimmed score function.<sup>19</sup>

*Implementing the Kim and Lai estimator.* For comparison, we also give corresponding results from the adaptive  $M$ -estimator proposed by Kim and Lai (2000), where efficient weights are estimated by a spline-function method, using a split-sample technique. The specific implementation is the one given in Section 3 of Kim and Lai (2000), except that (i) the starting values  $\tilde{\beta}$  are the same as in the previous simulation (i.e., LAD estimates),

---

<sup>19</sup> See footnote 10. The rate of occurrence (averaged over the various error distributions) was about 1 in 10,000 simulations for  $n = 100$  and about 1 in 20,000 for  $n = 400$ .

(ii)  $x$  is expressed as a deviation from the sample mean  $\bar{x}$ , and (iii) the estimated moment equation is solved for  $\hat{\beta}$  by a direct search method, rather than the iterative procedure used by Kim and Lai.<sup>20</sup> There are many alternative choices available for the spline basis function, for the cross-validation criterion used to determine the number of knots, and for the trimming rule, and it should be noted that some of these might yield an improved version of the Kim and Lai estimator.

Kim and Lai make use of asymptotic results derived in Lai and Ying (1994). Their proof extends to the case of a fixed censoring point only if the parameter space excludes the point  $\beta_0 = 0$  a priori, as can be seen from assumption (4.14) of Lai and Ying (1994). This does not necessarily invalidate the estimator, of course, and some results are presented here for  $\beta_0 = 0$  as well as for  $\beta_0 = 1$ .<sup>21</sup>

*Results.* Table 1 reports sampling statistics (biases and standard errors) of  $\hat{\beta}$  for the simulations with  $\beta_0 = 1$ . The column labeled RMSE/ASD gives the ratio of the root mean square error to the asymptotic efficiency bound for the standard deviation (ASD). The performance of the SPMLE  $\hat{\beta}$  is best for normal errors, where it is only slightly less efficient than the conventional parametric maximum likelihood tobit estimator. (For comparison, the corresponding normalized root mean squared errors of the tobit MLE are 1.021, 1.007, and 0.996 for  $n = 100, 200$  and  $400$ .) With thicker-tailed error distributions there is substantial bias at small sample sizes, and the approach to the efficiency bound is slower. The Kim and Lai estimator shows smaller bias but larger variance. In contrast with the SPMLE it performs relatively better for the thicker-tailed distributions, and for the  $t(2)$  and Cauchy distributions it begins to match the SPMLE (in terms of root mean square error) at  $n = 400$ .

For a subset of the sample designs, simulations were also run with  $\beta_0 = 0$ , with results reported in Table 2. Because of the symmetry of the estimator when  $\beta_0 = 0$ , the bias in  $\hat{\beta}$  is negligible and only the standard deviations are given. In this case the Kim and Lai estimator has larger standard errors, and does not appear to be “catching up” with the SPMLE over this range of sample sizes.

---

<sup>20</sup> Since the simulations reported here involve only a one-dimensional search, Brent’s method was used (see Press *et al.*, 1993, Chapter 9).

<sup>21</sup> Another assumption, (4.12b) of Lai and Ying (1994), requires an error distribution with at least exponentially decreasing tails, but the simulations given here do not show any markedly different behavior of the Kim and Lai estimator when the errors have a  $t$ -distribution.

**Table 1** *Estimates of  $\beta$  with  $\beta_0 = 1$*

<i>N</i>	SPML estimator			Adaptive M-estimator			ASD
	bias	SD	RMSE/ASD	bias	SD	RMSE/ASD	
Normal							
100	0.054	0.155	1.17	0.027	0.237	1.68	0.1414
200	0.029	0.106	1.10	0.007	0.149	1.49	0.1000
400	0.020	0.073	1.07	0.005	0.100	1.41	0.0707
Logistic							
100	0.058	0.157	1.19	0.023	0.218	1.56	0.1407
200	0.034	0.107	1.13	0.012	0.135	1.36	0.0995
400	0.022	0.074	1.10	0.006	0.090	1.29	0.0703
Laplace							
100	0.066	0.157	1.39	0.022	0.192	1.59	0.1217
200	0.040	0.100	1.24	0.013	0.120	1.41	0.0860
400	0.026	0.068	1.20	0.005	0.081	1.33	0.0608
<i>t</i> (3)							
100	0.058	0.130	1.21	0.015	0.169	1.46	0.1168
200	0.029	0.090	1.15	0.008	0.106	1.28	0.0826
400	0.022	0.061	1.11	0.004	0.070	1.21	0.0584
<i>t</i> (2)							
100	0.112	0.223	1.38	0.033	0.272	1.48	0.1853
200	0.060	0.151	1.24	0.017	0.169	1.30	0.1310
400	0.039	0.100	1.15	0.008	0.111	1.20	0.0927
Cauchy							
100	0.215	0.293	1.72	0.051	0.349	1.67	0.2115
200	0.097	0.185	1.40	0.022	0.212	1.42	0.1495
400	0.066	0.122	1.30	0.009	0.137	1.30	0.1057
Mixture 1							
100	0.043	0.078	1.26	0.008	0.107	1.52	0.0706
200	0.025	0.053	1.16	0.003	0.065	1.31	0.0499
400	0.011	0.037	1.10	0.002	0.042	1.20	0.0353
Mixture 2							
100	0.060	0.151	1.22	0.024	0.261	1.96	0.1335
200	0.025	0.100	1.09	0.007	0.156	1.66	0.0944
400	0.018	0.071	1.09	0.005	0.110	1.65	0.0668

Mixture 1 is  $0.9 N(0, 1/9) + 0.1 N(0, 9)$

Mixture 2 is  $0.5 N(1/\sqrt{2}, 1/2) + 0.5 N(-1/\sqrt{2}, 1/2)$

ASD is the standard deviation corresponding to the asymptotic semiparametric efficiency bound

**Table 2** *Estimates of  $\beta$  with  $\beta_0 = 0$*

$n$	SPML estimator	Adaptive M-estimator	ASD
	SD	SD	
Normal			
100	0.128	0.231	0.1105
400	0.062	0.100	0.0553
Laplace			
100	0.115	0.202	0.0707
400	0.050	0.084	0.0353
$t(2)$			
100	0.189	0.279	0.1348
400	0.080	0.136	0.0674
Mixture 1			
100	0.053	0.124	0.0406
400	0.024	0.044	0.0203

**Table 3** *Simulation results for quantiles of the SPML estimator of  $F$*

$n$	median		IQR		$n$	median		IQR	
	mean (SD)		mean (SD)			mean (SD)		mean (SD)	
Normal			1.349		$t(2)$			1.633	
100	-0.079	(0.171)	1.527	(0.220)	100	-0.133	(0.221)	1.970	(0.370)
200	-0.048	(0.110)	1.469	(0.151)	200	-0.080	(0.142)	1.863	(0.245)
400	-0.039	(0.080)	1.440	(0.107)	400	-0.055	(0.098)	1.792	(0.167)
Logistic			1.211		Cauchy			2.000	
100	-0.084	(0.163)	1.387	(0.215)	100	-0.208	(0.277)	2.536	(0.542)
200	-0.056	(0.108)	1.341	(0.150)	200	-0.130	(0.171)	2.402	(0.364)
400	-0.035	(0.074)	1.300	(0.105)	400	-0.081	(0.112)	2.257	(0.264)
Laplace			0.980		Mixture 1			0.502	
100	-0.082	(0.136)	1.175	(0.200)	100	-0.028	(0.092)	0.563	(0.162)
200	-0.051	(0.090)	1.115	(0.137)	200	-0.018	(0.053)	0.544	(0.058)
400	-0.029	(0.063)	1.059	(0.096)	400	-0.010	(0.038)	0.526	(0.039)
$t(3)$			0.883		Mixture 2			1.486	
100	-0.058	(0.134)	1.016	(0.164)	100	-0.069	(0.178)	1.637	(0.234)
200	-0.031	(0.084)	0.970	(0.108)	200	-0.044	(0.118)	1.596	(0.165)
400	-0.029	(0.059)	0.959	(0.073)	400	-0.038	(0.087)	1.571	(0.109)

For each error distribution, the first figure in the IQR column is the true interquartile range.

See Table 1 for the definitions of the error distributions denoted “Mixture 1” and “Mixture 2”.

Table 3 presents results for the median and interquartile range of the SPML estimator  $\hat{F}$ , in the case  $\beta_0 = 1$ . Negative bias in the median and overestimation of the interquartile range both arise from over-correction of censoring in the left tail. Although asymptotic properties of  $\hat{F}$  and its quantiles are not analyzed in this paper, we note that the simulation results appear consistent with an  $n^{-2a}$  rate of convergence.

## References

- Ai, C. (1997). A semiparametric maximum likelihood estimator, *Econometrica* 65, pp 933–963.
- Bickel, P. J. (1982). On adaptive estimation, *Annals of Statistics* 10, pp. 647–671.
- Bickel, P. J., C. A. J. Klaassen, Y. Ritov and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins Press.
- Buckley, J., and I. James (1979). Linear regression with censored data, *Biometrika* 66, pp 429–436.
- Cosslett, S. R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models, *Econometrica* 55, pp. 559–585.
- Duncan, G. M. (1986). A semiparametric censored regression estimator, *Journal of Econometrics* 32, pp. 5–34.
- Efron, B. (1967). The two sample problem with censored data, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. Berkeley: University of California Press.
- Fernandez, L. (1986). Nonparametric maximum likelihood estimation of censored regression models, *Journal of Econometrics* 32, pp. 35–57.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica* 50, pp. 1029–1054.
- Horowitz, J. L. (1986). A distribution-free least squares estimator for censored linear regression models, *Journal of Econometrics* 32, pp. 59–84.
- Horowitz, J. L. (1988). Semiparametric M-estimation of censored linear regression models, in *Advances in Econometrics* 7, ed. G. F. Rhodes and T. B. Fomby, pp 45–83.
- Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics* 58, pp. 71–120.
- Ichimura, H., and L-F. Lee (1991). Semiparametric least squares estimation of multiple index models: single equation estimation, in *Nonparametric and Semiparametric*

- Methods in Econometrics and Statistics*, ed. W. A. Barnett, J. Powell, and G. E. Tauchen, Cambridge University Press, pp. 3–49.
- Kaplan, E. L., and P. Meier (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* 53, pp. 457–481.
- Kim, C-K., and T. L. Lai, (2000). Efficient score estimation and adaptive M-estimators in censored and truncated regression models, *Statistica Sinica* 10, pp. 731–749.
- Klein, R. W., and R. H. Spady (1993). An efficient semiparametric estimator of binary response models, *Econometrica* 61, pp. 387–421.
- Koul, H., V. Susarla and J. Van Ryzin (1981). Regression analysis with randomly right censored data, *Annals of Statistics* 9, pp. 1276–1288.
- Lai, T. L., and Z. Ying (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, *Annals of Statistics* 19, pp. 1370–1402.
- Lai, T. L., and Z. Ying (1992). Asymptotically efficient estimation in censored and truncated regression models, *Statistica Sinica* 2, pp. 17–46.
- Lai, T. L., and Z. Ying (1994). A missing information principle and M-estimators in regression analysis with censored and truncated data, *Annals of Statistics* 22, 1222–1255.
- Newey, W. K. (2001). Conditional moment restrictions in censored and truncated regression models, *Econometric Theory* 17, pp 863–888.
- Paarsch, H. J. (1984). A Monte Carlo comparison of estimators for censored regression models, *Journal of Econometrics* 21, pp 197–213.
- Pagan, A., and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge: Cambridge University Press.
- Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model, *Journal of Econometrics* 25, pp. 303–325.
- Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling (1993). *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press.
- Ritov, Y. (1986). Efficient estimation in a linear regression with censored data, working paper (unpublished).
- Ritov, Y. (1990). Estimation in a linear regression model with censored data, *Annals of Statistics* 18, pp. 303–328.
- Severini, T. A., and W. H. Wong (1992). Profile likelihood and conditionally parametric models, *Annals of Statistics* 20, pp. 1768–1802.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

- Tsai, W-Y. and J. Crowley (1985). A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *Annals of Statistics* 13, pp. 1317–1334.
- Tsui, K-L., N. P. Jewell and C. F. Wu (1988). A nonparametric approach to the truncated regression problem, *Journal of the American Statistical Association* 83, pp. 785–792.
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B* 38, pp.290–295.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.