

NONPARAMETRIC CENSORED REGRESSION

Arthur Lewbel*
Brandeis University

Oliver Linton†
Yale University

July 28, 1998

Abstract

The nonparametric censored regression model is $y = \max[c, m(x) + e]$, where both the regression function $m(x)$ and the distribution of the error e are unknown, but the fixed censoring point c is known. This paper provides a simple consistent estimator of the derivative of $m(x)$ with respect to each element of x . The convergence rate of this estimator is the same as for the derivatives of an uncensored nonparametric regression. We then estimate the regression function itself by solving the associated partial differential equation system. We show that our estimator of $m(x)$ achieves the same rate of convergence as the usual estimators in uncensored nonparametric regression. We also provide root n estimates of weighted average derivatives of $m(x)$, which equal the coefficients in any linear or partly linear specification for $m(x)$.

JEL Codes: C14 C24 C13

Keywords: Semiparametric, Nonparametric, Censored Regression, Tobit, Latent Variable

*Department of Economics, Brandeis University, Waltham, MA 02254 USA. Tel: (781)–736-2258.

†Cowles Foundation for Research in Economics, Yale University, Yale Station Box 208281, New Haven, CT 06520-8281, USA. Phone: (203) 432-3699. Fax: (203) 432-6167. E-mail address: linton@econ.yale.edu; <http://www.econ.yale.edu/~linton>.

1 Introduction

Consider the censored regression model $Y_i = \max[c, m(X_i) - e_i]$, where X_i is an observed d vector of regressors X_{ki} for $k = 1, \dots, d$, and e_i is an unobserved error that is mean independent of X_i (writing the model as $m - e$ instead of the more usual $m + e$ simplifies later results). Here, the censoring point c is a known constant, which, without loss of generality we can take to be zero. Both the regression function $m(\cdot)$ and the distribution $F(\cdot)$ of the error e is unknown. The errors are not assumed to be symmetric. For each element x_k of x , let $m_k(x) = \partial m(x) / \partial x_k$. This paper provides a simple consistent estimator of the derivatives $m_k(x)$. These derivatives are directly interpretable as the marginal effect of a change in x on the underlying uncensored population. They can also be used to test or estimate parametric or semiparametric specifications of $m(x)$. For example, $m_k(x)$ is constant if $m(x)$ is linear in x_k , and $m_k(x)$ depends only on x_k if $m(x)$ is additive in a function of x_k . Also, we show that the regression function $m(x)$ itself can be estimated by integrating the derivative estimates, and the distribution function of the errors can be estimated given $m(x)$. The proposed estimator can be extended to deal with heteroscedasticity of the form $y = \max[m(x) - p(x)e, 0]$, where both $m(x)$ and $p(x)$ are unknown functions.

Parametric and semiparametric estimators of censored regression models include Amemiya (1973), Heckman (1976), Buckley and James (1979), Koul, Suslara, and Van Ryzin (1981), Powell (1984), (1986a), (1986b), Duncan (1986), Fernandez (1986), Horowitz (1986,1988), Moon (1989), Powell, Stock and Stoker (1989), Nawata (1990), Ichimura (1993), Honoré and Powell (1994), Lewbel (1998a, 1998b), and Buchinsky and Hahn (1998). Unlike the present paper, most of these models either assume $m(x) = \beta'x$, or they provide estimates of average derivatives only up to an unknown scale.

Concerning estimation of nonparametric censored regression models, it is well known that nonparametric quantile regression is unaffected by the presence of a little censoring. Except when the error distribution is symmetric, the mean and median are estimating different quantities, so our results are not directly comparable to nonparametric quantile regressions. One advantage of our method over median regression is that the conditional median estimator is only consistent when censoring is less than 50%, while our procedure for the mean works for any amount of censoring less than 100%. In the case of mean regression functions when c is a random censoring point independent of X (which is a model adopted in many medical applications), there are a number of suitable methods for estimating m . See, e.g., Fan and Gijbels (1996, section 5.2). However, to our knowledge no one has provided consistent estimates of mean regression functions in the fixed censoring case that we treat. The fully nonparametric model is important because of the sensitivity of the parametric and semiparametric estimators to misspecification of functional form.

Our estimation methods are based on local polynomials, whose advantages are discussed in Fan

and Gijbels (1996). We show that the uniform convergence rate of the derivative estimator is the same as for the derivatives of an uncensored regression. We also establish that weighted averages of these derivatives can be estimated that are root n consistent and asymptotically normal. Finally, we show that our estimator of $m(x)$ is asymptotically normal and converges at the same rate as the corresponding estimator in the uncensored case.

2 The Main Idea

Let $Y_i^* = m(X_i) - e_i$ be an unobserved latent variable. The function m is differentiable and unknown. The error e_i is independent of X_i and continuously distributed with unknown distribution function $F(e)$. The observed dependent variable Y_i equals the latent variable censored at zero, so $Y_i = I(Y_i^* \geq 0)Y_i^*$, where I is the indicator function that equals one if its argument is true and zero otherwise. We assume throughout that our observed data are independent, identically distributed observations (Y_i, X_i) for $i = 1, \dots, n$, although our main results, Theorems 1-4, under reasonable conditions hold as stated when $\{Y_i, X_i\}$ is a stationary mixing process with $\{e_i\}$ independent of $\{X_i\}$, as in Robinson (1982).

ASSUMPTION A1. *Assume $Y^* = m(X) - e$ and $Y = I(Y^* \geq 0)Y^*$. Let Ω be a compact subset of the support of the $d \times 1$ vector x . The function m is differentiable and has finite derivatives $m_k(x) = \partial m(x) / \partial x_k$ for $k = 1, \dots, d$, for all $x \in \Omega$. The error e is continuously distributed, independent of x , with probability distribution function $F(e)$ and probability density function $f(e)$. $\mathfrak{F}[m(x)]$ exists for all $x \in \Omega$, where the function \mathfrak{F} is $\mathfrak{F}(m) = \int_{-\infty}^m F(e)de$.*

Theorem 1 *If Assumption A.1 holds then $E(Y|X = x) = \mathfrak{F}[m(x)]$, $E[I(Y > 0)|X = x] = F[m(x)]$, and for all $x \in \Omega$ having $F[m(x)] \neq 0$,*

$$m_k(x) = \frac{\partial E(Y|X = x) / \partial x_k}{E[I(Y > 0)|X = x]}, \quad k = 1, \dots, d. \tag{1}$$

PROOF. We have

$$E(Y|X = x) = E\{[m(x) - e]I[m(x) - e \geq 0]\}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} [m(x) - e]I[e \leq m(x)]f(e)de \\
&= \int_{-\infty}^{m(x)} [m(x) - e]f(e)de = m(x) \int_{-\infty}^{m(x)} f(e)de - \int_{-\infty}^{m(x)} ef(e)de.
\end{aligned}$$

Now do an integration by parts

$$E(Y|X = x) = m(x)F[m(x)] - [eF(e)]|_{-\infty}^{m(x)} + \int_{-\infty}^{m(x)} F(e)de = \mathfrak{F}[m(x)].$$

Next we have the standard result that $E[I(Y > 0)|X = x] = E\{I[e \leq m(x)]\} = F[m(x)]$, and therefore

$$\frac{\partial E(Y|X = x)}{\partial x_k} = \frac{\partial m(x)}{\partial x_k} \frac{\partial E(Y|X = x)}{\partial m} = \frac{\partial m(x)}{\partial x_k} F[m(x)] = \frac{\partial m(x)}{\partial x_k} E[I(Y > 0)|X = x]. \quad (2)$$

■

For the special case of $m(x) = \beta'x$, the fact that $E(Y|X = x) = \mathfrak{F}[m(x)]$, and hence equation (1) holds, has long been known. See, e.g., Rosett and Nelson (1975), Heckman (1976), McDonald and Moffitt (1980), and Horowitz (1986). Theorem 1 shows that this expression holds for arbitrary m and F , and so can be exploited for nonparametric estimation of $m_k(x)$.

Define $r(x) = E(Y|X = x)$, $r_k(x) = \partial r(x)/\partial x_k$, and $s(x) = E[I(Y > 0)|X = x]$. Theorem 1 showed that $m_k(x) = r_k(x)/s(x)$. The latter expression can be readily estimated using any differentiable nonparametric regression (conditional expectation) estimators for r and s , e.g., kernel regressions.

As discussed in an Extension section later, results similar to Theorem 1 hold for $E(Y^\kappa|X = x)$ for arbitrary positive integers κ . These can be exploited to yield additional estimators of $m_k(x)$, including estimation in the presence of heteroscedasticity of the form $y = \max[m(x) - p(x)e, 0]$, where both $m(x)$ and $p(x)$ are unknown functions.

2.1 Censored Regression Function

Given $m_k(x)$ for $k = 1, \dots, d$, the censored regression function $m(x)$ can be recovered under certain conditions by solving the implied system of partial differential equations. This problem is similar to Hausman and Newey (1995), who, based on Shephard's lemma, recover estimates of consumer surplus by integrating nonparametrically estimated demand equations. See Horowitz (1998) for another application of this idea, in his case to estimating additive transformation models. For the partial derivative system $m_k(x) = g_k(x)$, $k = 1, \dots, d$, to have a solution it suffices that we have a boundary condition¹ $m(\underline{x}) = 0$ for some \underline{x} and that the cross partials are symmetric, i.e., $m_{k\ell}(x) = m_{\ell k}(x)$ for

¹The boundary condition $m(\underline{x}) = 0$ can be assumed to hold without loss of generality for any $\underline{x} \in \Omega$, because the errors are not assumed to have mean zero.

all ℓ, k , where

$$m_{k\ell}(x) = \frac{r_{k\ell}(x)}{s(x)} - \frac{r_k(x) \cdot s_\ell(x)}{s^2(x)}, \quad (3)$$

where $r(x) = E(Y|X = x)$, $r_k(x) = \partial r(x)/\partial x_k$, and $s(x) = E[I(Y > 0)|X = x]$. Therefore, it suffices that

$$s_k(x) \cdot r_\ell(x) = s_\ell(x) \cdot r_k(x)$$

for all ℓ, k . If this is true, then the p.d.e. system has a unique solution given by

$$m(x) = \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} m_k(z_k(t)) dt, \quad (4)$$

where for each k , $z_k(t) = (x_1, \dots, x_{k-1}, t, \underline{x}_{k+1}, \dots, \underline{x}_d)$, see Varian (1984, p332). We will use (1) and (4) below to generate estimates of $m_k(x)$ and $m(x)$.

2.2 Average Derivatives and Partly Linear Models

Given any weighting function $w(x)$, define the average regression function derivative $\delta_{wk} = E[w(X)m_k(X)]/E[w(X)]$. Since $m_k(x) = r_k(x)/s(x)$, this δ_{wk} can be estimated at rate root n by replacing the expectations with sample averages and substituting in nonparametric regression based estimates of $r_k(x)$ and $s(x)$.

Taking $w(x) = 1$ results in unweighted average derivatives. Taking $w(x)$ to equal $s(x)$ times the density of x yields a particularly simple form for δ_{wk} if kernel regressions are used to estimate $r_k(x)$ and $s(x)$, since then δ_{wk} will equal the Powell, Stock, and Stoker's (1989) weighted average derivative divided by the mean of a kernel regression numerator (see, e.g., Lewbel 1995).

If the latent regression function is linear or partly linear, that is, if for some $j \leq d$, $m(x) = \beta_1 x_1 + \dots + \beta_j x_j + \tilde{m}(x_{j+1}, \dots, x_k)$, then for $1 \leq k \leq j$, $\beta_k = \delta_{wk}$. Root n estimation of the coefficients in uncensored partly linear regression models is described in Robinson (1988), among others. In contrast, what is provided here is estimation of the same parameters when the partly linear model is censored. As an estimator of β_k , δ_{wk} has the advantage that if $m(x)$ turns out to not be linear or partly linear, δ_{wk} will still equal the usual interpretation of β_k as a measure of the average effect on the latent variable of a marginal change in x_k .

2.3 The Error Distribution

For any e^* , $E[I(Y > 0)|m(X) = e^*] = F(e^*)$, where F is the distribution function of the errors e . Therefore, given the estimated regression function $\hat{m}(x)$, the distribution function F can be estimated

as a nonparametric regression of $I(Y > 0)$ on $\widehat{m}(x)$. In addition, Lemma 1 in Lewbel (1997) can be used to directly estimate the mean, variance, and other moments of e .

It may alternatively be possible to use Theorem 1 to estimate F , or at least functions of F , more directly. For example, it follows from Theorem 1 that $h(x) = F\{\mathfrak{F}^{-1}[g(x)]\}$, so a one dimensional nonparametric regression of $I(Y > 0)$ on $\widehat{g}(x)$ should yield a consistent estimate of the function $F[\mathfrak{F}^{-1}(g)]$.

3 Estimation

For the remainder of the paper we will discuss estimation using local polynomials. We use local polynomials instead of ordinary kernel or sieve estimators because of their attractive properties with regard to boundary bias and design adaptiveness, see Fan and Gijbels (1996) for discussion and references.

We shall use the following notation. For functions g and vectors $\mathbf{k} = (k_1, \dots, k_d)$ and $x = (x_1, \dots, x_d)$, let

$$\mathbf{k}! = k_1! \times \dots \times k_d!, \quad |\mathbf{k}| = \sum_{i=1}^d k_i, \quad x^{\mathbf{k}} = x_1^{k_1} \times \dots \times x_d^{k_d}$$

$$\sum_{0 \leq |\mathbf{k}| \leq p} = \sum_{j=0}^p \sum_{k_1=0}^j \dots \sum_{k_d=0}^j, \quad (D^{\mathbf{k}}g)(y) = \frac{\partial^{|\mathbf{k}|} g(y)}{\partial y_1^{k_1} \dots \partial y_d^{k_d}}.$$

To be consistent with our earlier usage, we will also use the special notation $g_{\mathbf{k}}(x) = D^{e_{\mathbf{k}}}g(x)$, where $e_{\mathbf{k}}$ is the k^{th} elementary vector, and $g_{\mathbf{k}\ell}(x) = D^{(e_{\mathbf{k}}+e_{\ell})}g(x)$. We also stack the first derivatives into a vector so that $Dg(x) = (g_1(x), \dots, g_d(x))'$.

3.1 Nonparametric Regression Derivatives

Given observations $\{Y_i, X_i\}_{i=0}^n$ and function $\Psi(Y_i)$, we shall estimate the regression function $g(x) = E[\Psi(Y_i)|X_i = x]$ and its derivatives using the multivariate weighted least squares criterion

$$\sum_{i=1}^n \left[\Psi(Y_i) - \sum_{0 \leq |\mathbf{k}| \leq p} b_{\mathbf{k}}(x)(X_i - x)^{\mathbf{k}} \right]^2 \mathcal{K}((X_i - x)/h_n), \quad (5)$$

where $\mathcal{K}(u)$ is a nonnegative weight function on \mathbb{R}^d and h_n is a bandwidth parameter, while p is an integer with $p \geq 2$. Minimizing (5) with respect to each $b_{\mathbf{k}}$ gives an estimate $\widehat{b}_{\mathbf{k}}(x)$ such that $(D^{\mathbf{k}}g)^{\wedge}(x) = \mathbf{k}! \widehat{b}_{\mathbf{k}}(x)$ estimates $(D^{\mathbf{k}}g)(x)$. Let also $\widehat{g}_{\mathbf{k}}(x) = (D^{e_{\mathbf{k}}}g)^{\wedge}(x)$ and $\widehat{D}g(x) = (\widehat{g}_1(x), \dots, \widehat{g}_d(x))'$.

3.2 The Censored Regression Function Derivatives

Let $\widehat{r}_k(x)$ and $\widehat{s}(x)$ be nonparametric estimators of the functions $r_k(x)$ and $s(x)$ as defined above. Specifically, for $\widehat{r}_k(x)$ and $\widehat{s}(x)$ we take $\Psi(y) = y$ and $\Psi(y) = 1(y > 0)$ in (5) respectively. We then let

$$\widehat{m}_k(x) = \frac{\widehat{r}_k(x)}{\widehat{s}(x)}, \quad k = 1, \dots, d. \quad (6)$$

3.3 The Censored Regression Function

To estimate $m(x)$ we use the representation (4). Since the solution to the p.d.e. system is explicit, we can just substitute estimates of $m_k(\cdot)$ and do numerical integration. Specifically, let

$$\widehat{m}(x) = \sum_{k=1}^d \frac{1}{J} \sum_{i=1}^J \frac{\widehat{m}_k(z_k(X_{ki}^*))}{\varphi_k(X_{ki}^*)} \mathbf{1}(X_{ki}^* \in [x_k, x_k]), \quad (7)$$

where $J(n)$ is a large number such that $J(n) \rightarrow \infty$ at a faster rate than n , each X_{ki}^* are i.i.d. with density φ_k , while $\widehat{m}_k(x)$ are nonparametric estimates of $m_k(x)$ defined above.

3.4 Symmetry

It is desirable for efficiency reasons to impose the symmetry condition (3) on the partial derivative estimates, which can be done as follows. The symmetry condition requires of the first partial derivatives of $r(x)$ and $s(x)$ that

$$\frac{r_k(x)}{r_\ell(x)} = \frac{s_k(x)}{s_\ell(x)} \quad \text{for all } x, k, \ell. \quad (8)$$

Letting $\delta = (\delta_1, \dots, \delta_{d+1})' \in \mathbb{R}^{d+1}$ and

$$(\delta_1, \delta_3 \cdot \delta_1, \dots, \delta_d \cdot \delta_1 | \delta_2, \delta_3 \cdot \delta_2, \dots, \delta_d \cdot \delta_2)' := g(\delta) = (g'_1(\delta) | g'_2(\delta))',$$

the symmetry condition (8) can be expressed conveniently as $(Dr(x), Ds(x))' = g(\delta)$, where we have suppressed the dependence of δ on x . Now let $\widehat{Dr}(x) = (\widehat{r}_1(x), \dots, \widehat{r}_d(x))$ and $\widehat{Ds}(x) = (\widehat{s}_1(x), \dots, \widehat{s}_d(x))$ be estimates of $Dr(x)$ and $Ds(x)$, and let

$$\widehat{\delta}(x) = \arg \min_{\delta} \left\{ \left(\begin{array}{c} \widehat{Dr}(x) \\ \widehat{Ds}(x) \end{array} \right) - g(\delta) \right\}' V_n^{-1} \left\{ \left(\begin{array}{c} \widehat{Dr}(x) \\ \widehat{Ds}(x) \end{array} \right) - g(\delta) \right\},$$

where V_n is weighting sequence with $V_n \rightarrow^p V > 0$ [we discuss choice of V_n below]. We then let $\widetilde{Dr}(x) = g_1(\widehat{\delta}(x))$ and $\widetilde{Ds}(x) = g_2(\widehat{\delta}(x))$ be our symmetry restricted estimates of $Dr(x)$ and $Ds(x)$, which can then be used to define a new estimate of $m_k(x)$, $Dm(x)$ which we denote by $\widetilde{m}_k(x)$, $\widetilde{Dm}(x)$. This can also be used in (7).

4 Asymptotic Properties

4.1 Nonparametric Regression

We first give some general definitions for our local polynomial kernel nonparametric regression estimators. Let

$$N_\ell = \binom{\ell + d - 1}{d - 1}$$

be the number of distinct d -tuples j with $|j| = \ell$. Arrange these N_ℓ d -tuples as a sequence in a lexicographical order (with highest priority to last position so that $(0, \dots, 0, \ell)$ is the first element in the sequence and $(\ell, 0, \dots, 0)$ the last element) and let ϕ_ℓ^{-1} denote this one-to-one map. Arrange the distinct values of $(D^{\mathbf{k}})^{\wedge}(g)$, $0 \leq |\mathbf{k}| \leq p$, as a column vector of dimension $N \times 1$, where $N = \sum_{\ell=0}^p N_\ell \times 1$, where the i^{th} element of that vector is obtained by the following relation

$$i = \phi_{|j|}^{-1}(j) + \sum_{k=0}^{|j|-1} N_k. \quad (9)$$

Similarly, arrange the vector $(D^{\mathbf{k}})(g)$. For each j with $0 \leq |j| \leq 2p$, let

$$\mu_j(\mathcal{K}) = \int_{\mathbb{R}^d} u^j \mathcal{K}(u) du, \quad \nu_j(\mathcal{K}) = \int_{\mathbb{R}^d} u^j \mathcal{K}^2(u) du,$$

and define the $N \times N$ dimensional matrices M and Γ and $N \times 1$ vector B by

$$M = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p} \\ \vdots & & & \vdots \\ M_{p,0} & M_{p,1} & \cdots & M_{p,p} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,p} \\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,p} \\ \vdots & & & \vdots \\ \Gamma_{p,0} & \Gamma_{p,1} & \cdots & \Gamma_{p,p} \end{bmatrix}, \quad B = \begin{bmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{bmatrix}, \quad (10)$$

where $M_{i,j}$ and $\Gamma_{i,j}$ are $N_i \times N_j$ dimensional matrices whose (ℓ, m) element are, respectively, $\mu_{\phi_i(\ell) + \phi_j(m)}$ and $\nu_{\phi_i(\ell) + \phi_j(m)}$. Note that the elements of the matrices M and Γ are simply multivariate moments of the kernel \mathcal{K} and \mathcal{K}^2 , respectively. Finally, arrange the N_{p+1} elements of the derivatives $(1/j!)(D^j g)(x)$ for $|j| = p + 1$ as a column vector $\mathcal{D}_{p+1}(x; g)$ using the lexicographical order introduced earlier.

For each j with $0 \leq |j| \leq 2p + 1$ define the function

$$H_j(u) = u^j \mathcal{K}(u).$$

We make the following assumptions on the kernel \mathcal{K} .

ASSUMPTION A2

- (a) The kernel \mathcal{K} is bounded with compact support ($\mathcal{K}(u) = 0$ for $\|u\| > A_0$).
- (b) For all j with $0 \leq |j| \leq 2p + 1$, there exists finite C_4 such that

$$|H_j(u) - H_j(v)| \leq C_4 \|u - v\|.$$

ASSUMPTION A3.

- (a) The regression functions r and s are $p + 1$ -times continuously differentiable.
- (b) The conditional distribution $G(y|u)$ of Y given $X = u$ is continuous at the point $u = x$.

REMARK. By dominated convergence, Assumption A3(b) implies that for each $L > 0$, the functions $E[Y1(|Y| < L)|X = u]$, $E[Y^21(|Y| < L)|X = u]$, are continuous at the point x . Hence for each $L > 0$, $\tilde{\sigma}_L^2(u) = \text{var}[Y \cdot 1(|Y| > L)|X = u]$ is continuous at the point x provided $m(\cdot)$ and $\sigma(\cdot)$ are continuous at the point x . This is needed in the proof of Theorem 2 where a truncation argument is employed and the continuity of $\tilde{\sigma}_L^2(u)$ at $u = x$ is required.

ASSUMPTION B

- (a) For any k with $|k| = p + 1$, there exists finite C_6 such that

$$|(D^k r)(u) - (D^k r)(v)|, |(D^k s)(u) - (D^k s)(v)| \leq C_6 \|u - v\|.$$

- (b) $E[|Y_1|^t] < \infty$ for some $t > 2$.
- (c) The Lebesgue density f of X and the regression function s satisfy

$$\inf_{x \in \mathcal{X}} f(x) > 0 \quad ; \quad \inf_{x \in \mathcal{X}} s(x) > 0$$

on some compact subset \mathcal{X} of \mathbb{R}^d .

We are now ready to give the asymptotic properties of our estimate $\widehat{Dm}(x)$ of $(Dm)(x)$ computed using our estimates $\widehat{Dr}(x)$ and $\widehat{s}(x)$. Define $\sigma_r^2(x) = \text{var}(Y|X = x)$ and $\sigma_s^2(x) = \text{var}[1(Y > 0)|X = x]$.

Theorem 2 Suppose that Assumptions A1-A3 hold and that $h_n = O(n^{-1/(d+2p+2)})$. Then, we have

$$\sqrt{nh_n^{d+2}} \left[\left\{ \widehat{Dm}(x) - Dm(x) \right\} - h_n^p \frac{(M^{-1}BD_{p+1}(x;r))_1}{s(x)} \right] \implies N \left[0, \frac{\sigma_r^2(x)}{f(x)s^2(x)} (M^{-1}\Gamma M^{-1})_{1,1} \right]$$

at continuity points x of $\{\sigma_r^2, \sigma_s^2, f, s\}$ whenever $f(x), s(x) > 0$. Here, $(M^{-1}\Gamma M^{-1})_{1,1}$ and $(M^{-1}BD_{p+1}(x;r))_1$ are the corresponding [as in (10)] submatrix of $M^{-1}\Gamma M^{-1}$ and subvector of $M^{-1}BD_{p+1}(x;r)$, respectively. Suppose in addition that Assumption B holds, and that the bandwidth $h_n \rightarrow 0$ slowly enough such that the right hand side of (11) below is $o(1)$. Then, we have with probability one

$$\sup_{x \in \mathcal{X}} |\widehat{Dm}(x) - (Dm)(x)| = O \left\{ \left(\frac{\ln n}{nh_n^{d+2}} \right)^{1/2} \right\} + O(h_n^p). \quad (11)$$

REMARKS A.

1. The optimal bandwidth for estimating the j^{th} derivative $(D^{e_j}m)(x)$ can be defined as the one which minimizes the sum of the squared bias and “variance” above. With $h_n^{\text{opt}} = O(n^{-1/(d+2p+2)})$ and using the above expressions for the bias and “variance” for the estimate of $m_j(x)$ it is seen that the rate of “mean-square convergence” is $O(n^{-2p/(d+2p+2)})$ which matches the optimal rate given by Stone (1980,1982) in the i.i.d. regression setting.
2. The quantity $s(x)$ measures the amount of censoring: when $s(x) = 1$ there is no censoring, while when $s(x) = 1/2$ there is 50% censoring. Both variance and bias deteriorate as $s(x)$ decreases, but the estimate is still consistent when $s(x) < 1/2$ in contrast to the usual nonparametric median estimator.

4.2 Symmetry Restricted Estimation

By a straightforward extension of Masry (1996b, Theorem 5), we can establish that

$$n^{p/(d+2p+2)} \begin{pmatrix} \widehat{Dr}(x) - Dr(x) \\ \widehat{Ds}(x) - Ds(x) \end{pmatrix} \implies N(b(x), \Omega(x)), \quad (12)$$

for some matrix functions $b(\cdot)$ and $\Omega(\cdot)$. Masry actually gives the marginal convergences but the joint convergence follows directly under our conditions provided the covariance $\sigma_{rs}(x) = \text{cov}(Y, 1(Y \geq 0)|X = x)$ is continuous at x . In fact,

$$b_{2d \times 1}(x) = \begin{pmatrix} b_r(x) \\ b_s(x) \end{pmatrix}, \quad \Omega_{2d \times 2d}(x) = \begin{pmatrix} \Omega_{rr}(x) & \Omega_{rs}(x) \\ \Omega_{sr}(x) & \Omega_{ss}(x) \end{pmatrix},$$

where

$$b_j(x) = \left\{ \lim_{n \rightarrow \infty} n^{p/(d+2p+2)} h_n^p \right\} \times (M^{-1} B \mathcal{D}_{p+1}(x; g_j))_1, \quad j = r, s$$

$$\Omega(x) = \left\{ \lim_{n \rightarrow \infty} n^{2p/(d+2p+2)} (n h_n^{d+2})^{-1} \right\} \times \frac{1}{f(x)} \begin{pmatrix} \sigma_r^2(x) & \sigma_{rs}(x) \\ \sigma_{rs}(x) & \sigma_s^2(x) \end{pmatrix} \otimes (M^{-1} \Gamma M^{-1})_{1,1}.$$

We have the following result.

Theorem 3 *Suppose that Assumptions A1-A3 hold and $h_n = O(n^{-1/(d+2p+2)})$. Then,*

$$n^{p/(d+2p+2)} \left\{ \widetilde{Dm}(x) - Dm(x) \right\} \implies N[\Phi b, \Phi \Omega \Phi'],$$

where $\Phi = G'_1(GV^{-1}G')^{-1}GV^{-1}/s$ and $G = (G_1, G_2)$ is the $(d+1) \times 2d$ vector of partial derivatives $\partial g_\ell / \partial \delta_k$ evaluated at δ , and we have suppressed dependence on x .

This theorem says that the pointwise asymptotic mean squared error matrix for $\widetilde{Dm}(x)$ is $\Phi[bb' + \Omega]\Phi'$. When the matrix $\Xi = bb' + \Omega$ is non-singular, the optimal weighting matrix according to mean squared error is $V = \Xi$, and the asymptotic mean squared error is $G'_1(G\Xi^{-1}G')^{-1}G_1$. In practice, we must replace Ξ by an estimate. Estimating the bias term b is quite difficult and likely to suffer from small sample effects. An alternative approach is to choose V to minimize the asymptotic variance, in which case $V^{-1} = \Omega^{-1}$ and the asymptotic variance is $G'_1(G\Omega^{-1}G')^{-1}G_1$ with bias $G'_1(G\Omega^{-1}G')^{-1}G\Omega^{-1}b$. The quantity Ω can be estimated using residuals. Finally, note that

$$\Omega - G'(G\Omega^{-1}G')^{-1}G = \Omega^{1/2} [I - \Omega^{-1/2}G'(G\Omega^{-1}G')^{-1}G\Omega^{-1/2}] \Omega^{1/2} \geq 0,$$

so that the symmetry restricted estimates are more efficient. Intuitively, the decrease in variance is about one half, reflecting the fact that symmetry effectively doubles the number of observations used in each estimation.

4.3 Estimation of The Regression Function

We now suppose that $p = 1$, i.e., we use a local linear estimator. Note that in the local linear case, the bias of \widetilde{Dm} is actually of order h_n^2 , provided r has three continuous derivatives, rather than the implied h_n suggested by Theorem 2.

We take $\mathcal{K}(u) = \prod_{\ell=1}^d K(u_\ell)$ and let $L(u) = \int_{-\infty}^u K(v)vdv$, which is symmetric about zero and has the same support as K , i.e., it can be interpreted as a sort of kernel [although it doesn't necessarily integrate to one]. We have the following theorem

Theorem 4 *Suppose that Assumptions A1-A3 and B hold except that $r(x)$ has three continuous derivatives, and that $\limsup_{n \rightarrow \infty} nh_n^{d+4} < \infty$. Suppose also that $x_j \neq \underline{x}_j$, $j = 1, \dots, d$, and that $\times_{k=1}^d [\underline{x}_k, x_k] \subseteq \mathcal{X}$. Then,*

$$\frac{\widehat{m}(x) - m(x) - \frac{h_n^2}{6}\beta(x)}{\sqrt{\frac{1}{nh_n^d}v(x)}} \implies N(0, 1),$$

where

$$\begin{aligned} \beta(x) &= \sum_{j=1}^d \sum_{l=1}^d \sum_{m=1}^d \frac{\int \mathcal{K}(u)u_k u_j u_l u_m du}{\mu_2(K)} \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \frac{3r_{jl}(z_k(t))f_m(z_k(t)) + r_{jlm}(z_k(t))f(z_k(t))}{s(z_k(t))f(z_k(t))} dt \\ &\quad - 3\mu_2(K) \sum_{j=1}^d \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \left\{ \frac{f_k(z_k(t))}{f(z_k(t))s(z_k(t))} r_{jj}(z_k(t)) + \frac{r_k(z_k(t))}{s^2(z_k(t))} s_{jj}(z_k(t)) \right\} dt, \end{aligned} \quad (13)$$

$$\begin{aligned} v(x) &= \frac{\|K\|_2^{2(d-1)} \|L\|^2}{\mu_2^2(K)} \sum_{k=1}^d \left\{ \frac{\sigma_r^2(z_k(x_k))}{(s^2 \cdot f)(z_k(x_k))} + \frac{\sigma_r^2(z_k(\underline{x}_k))}{(s^2 \cdot f)(z_k(\underline{x}_k))} \right\} \\ &\quad - \frac{2\|K\|^{2(d-2)} \langle K, L \rangle^2}{\mu_2^2(K)} \sum_{k=1}^{d-1} \frac{\sigma_r^2(z_k(x_k))}{(s^2 \cdot f)(z_k(x_k))}, \end{aligned} \quad (14)$$

where $\langle K, L \rangle = \int K(u)L(u)du$. Furthermore,

$$nh_n^d \text{cov}(\widehat{m}(x), \widehat{m}(z) | \mathcal{X}^n) \rightarrow \frac{\|K\|_2^{2(d-1)} \|L\|^2}{\mu_2^2(K)} \sum_{k=1}^d \frac{\sigma_r^2(\underline{x})}{(s^2 \cdot f)(\underline{x})}, \quad (15)$$

for any $x \neq z$ and $x_j, z_j \neq \underline{x}_j$, $j = 1, \dots, d$.

REMARKS B.

1. When K is Gaussian, $L(u) = -K(u)$. In this case, the stochastic part of $\widehat{m}(x) - m(x)$ is like the stochastic part of a (weighted) kernel estimator evaluated at x minus the stochastic part of the same estimator evaluated at \underline{x} , and

$$v(x) = \frac{1}{2^d \pi^{d/2}} \left\{ \frac{\sigma_r^2(x)}{(s^2 \cdot f)(x)} + \frac{\sigma_r^2(\underline{x})}{(s^2 \cdot f)(\underline{x})} \right\}.$$

This explains why there is a covariance between the estimates at distinct points x and z .

2. Standard errors can be calculated in the obvious way; all that is required are consistent estimates of $\sigma_r^2(\cdot)$, $s(\cdot)$, and $f(\cdot)$, see Härdle and Linton (1994) for discussion of this.
3. Using the symmetrized derivative estimators affects both bias and variance, reducing the latter by approximately one half.
4. The local quadratic estimate has the same variance and the same bias magnitude $[O(h_n^2)]$, although the exact form of the bias depends only on the fourth derivatives of r_k and is design adaptive.
5. Estimation of the distribution function can be analyzed using the theory of nonparametric regression for generated regressors treated in Ahn (1995) and Rilestone (1996).

4.4 An Average Censored Regression Derivative Estimator

For $w(\cdot)$ a given weight function and let $\delta_w = E[w(X_i)m_k(X_i)]$, and define

$$u_i = w(X_i) \left\{ m_k(X_i) - \widehat{\delta}_w \right\} + \left\{ \frac{w_k(X_i)}{s(X_i)} - \frac{w(X_i)}{s(X_i)} \frac{s_k(X_i)}{s(X_i)} \right\} \varepsilon_i - \frac{w(X_i)r_k(X_i)}{s^2(X_i)} \eta_i$$

$$\widehat{u}_i = w(X_i) \left\{ \widehat{m}_k(X_i) - \widehat{\delta}_w \right\} + \left\{ \frac{w_k(X_i)}{\widehat{s}(X_i)} - \frac{w(X_i)}{\widehat{s}(X_i)} \frac{\widehat{s}_k(X_i)}{\widehat{s}(X_i)} \right\} \widehat{\varepsilon}_i - \frac{w(X_i)\widehat{r}_k(X_i)}{\widehat{s}^2(X_i)} \widehat{\eta}_i,$$

where $\varepsilon_i = Y_i - r(X_i)$ and $\eta_i = 1(Y_i > 0) - s(X_i)$, while $\widehat{\varepsilon}_i = Y_i - \widehat{r}(X_i)$ and $\widehat{\eta}_i = 1(Y_i > 0) - \widehat{s}(X_i)$.

Theorem 5 *Suppose that conditions A1-A3 and B hold and that p, h_n are such that $nh_n^{2p} \rightarrow 0$ and $nh_n^d / \log n \rightarrow \infty$. Suppose also that $w(\cdot)$ is continuously differentiable. Then,*

$$\sqrt{n}(\widehat{\delta}_w - \delta_w) \implies N(0, E(u_i^2)) \quad ; \quad \frac{\sqrt{n}(\widehat{\delta}_w - \delta_w)}{\sqrt{\frac{1}{n} \sum_{i=1}^n \widehat{u}_i^2}} \implies N(0, 1). \quad (16)$$

5 Extensions and Conclusions

We have provided an estimator for the nonparametric censored regression model with fixed censoring, assuming the errors are mean independent of x . The estimator is based on the conditional means $r(x)$ and $s(x)$. Higher moments of Y can also be employed. In particular, for any integer $\kappa \geq 2$,

$$m_k(x) = \frac{\partial E(Y^\kappa | X = x) / \partial x_k}{\kappa E[I(Y^{\kappa-1}) | X = x]}, \quad k = 1, \dots, d. \quad (17)$$

The proof works in exactly the same way as Theorem 1. These higher moment based estimates could either be combined with the estimator based on Theorem 1 to improved efficiency, or compared to that estimator as a test of the (nonparametric) model specification.

The estimator can also be extended to handle some heteroscedasticity. Consider the model $y = \max[m(x) - p(x)e, 0]$, where now both $m(x)$ and $p(x)$ are unknown, differentiable functions. Assume $p(x) > 0$. Let $M(x) = m(x)/p(x)$ and $P(x) = \ln p(x)$. If $y = \max[m(x) - p(x)e, 0]$, it follows from Theorem 1 that now $s(x) = E[I(Y > 0)|X = x] = F[M(x)]$ and $r(x) = E(Y|X = x) = p(x)\mathfrak{F}[M(x)]$, where $\partial\mathfrak{F}(M)/\partial M = F(M)$. Similarly, by the above higher moment extension of Theorem 1 for $\kappa = 2$, $t(x) = E(Y^2|X = x) = p(x)^2\mathfrak{F}_2[M(x)]$, where $\partial\mathfrak{F}_2(M)/\partial M = 2\mathfrak{F}(M)$. Letting the subscript k denote taking the derivative with respect to x_k we have

$$r_k(x) = p(x)F[M(x)]M_k(x) + \mathfrak{F}[M(x)]p_k(x) = s(x)m_k(x) + [r(x) - s(x)]P_k(x)$$

$$t_k(x)/2 = p(x)^2\mathfrak{F}[M(x)]M_k(x) + \mathfrak{F}_2[M(x)]p(x)p_k(x) = r(x)m_k(x) + [t(x) - r(x)]P_k(x).$$

Solving this pair of equations for the regression function derivatives yields

$$m_k(x) = \frac{[t(x) - r(x)]r_k(x) - [r(x) - s(x)][t_k(x)/2]}{t(x)s(x) - r(x)^2} \quad (18)$$

The right side of this expression is a function of conditional expectations and their derivatives, and so can be estimated using kernel functions or local polynomials, and can be integrated to yield $m(x)$ or averaged to get average derivatives and coefficients in a linear or partly linear specification for $m(x)$. The above pair of equations can also be solved for the variance function derivatives, yielding

$$P_k(x) = \frac{[s(x)t_k(x)/2] - r(x)r_k(x)}{t(x)s(x) - r(x)^2}. \quad (19)$$

A Appendix

We first give some facts about the generic local linear estimator $\widehat{g}_k(x)$ of a partial derivative $g_k(x)$, which will be needed in the proof of Theorem 4. This can be decomposed as

$$\widehat{g}_k(x) - g_k(x) = e'_k M_n^{-1}(x)U_n(x) + e'_k M_n^{-1}(x)B_n(x),$$

where $e_k = (0, 0, \dots, 0, 1, 0, \dots, 0)'$ is the $d + 1$ vector with the one in the $k + 1$ position, while the $(d + 1) \times (d + 1)$ symmetric matrix $M_n(x)$ is

$$\begin{bmatrix} \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) & \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_1-X_{1i}}{h_n}\right) & \cdots & \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_d-X_{di}}{h_n}\right) \\ & \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_1-X_{1i}}{h_n}\right)^2 & \cdots & \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_1-X_{1i}}{h_n}\right) \left(\frac{x_d-X_{di}}{h_n}\right) \\ & & \ddots & \vdots \\ & & & \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_d-X_{di}}{h_n}\right)^2 \end{bmatrix},$$

the stochastic term

$$U_n(x) = \begin{bmatrix} \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \epsilon_i \\ \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_1-X_{1i}}{h_n}\right) \epsilon_i \\ \vdots \\ \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_d-X_{di}}{h_n}\right) \epsilon_i \end{bmatrix} = \begin{bmatrix} U_{n0}(x) \\ U_{n1}(x) \\ \vdots \\ U_{nd}(x) \end{bmatrix},$$

where ϵ_i is the mean zero error term, and finally the bias term

$$B_n(x) = \begin{bmatrix} \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \Delta_i(x) \\ \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_1-X_{1i}}{h_n}\right) \Delta_i(x) \\ \vdots \\ \frac{1}{nh_n^d} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h_n}\right) \left(\frac{x_d-X_{di}}{h_n}\right) \Delta_i(x) \end{bmatrix} = \begin{bmatrix} B_{n0}(x) \\ B_{n1}(x) \\ \vdots \\ B_{nd}(x) \end{bmatrix},$$

where $\Delta_i(x) = g(X_i) - g(x) - \sum_{k=1}^d g_k(x)(X_{ki} - x_k)$, with

$$\begin{aligned} \Delta_i(x) &= \frac{h_n^2}{2} \sum_{j=1}^d \sum_{l=1}^d \frac{\partial^2 g}{\partial x_j \partial x_l}(x) \left(\frac{X_{ji} - x_j}{h_n}\right) \left(\frac{X_{li} - x_l}{h_n}\right) \\ &\quad + \frac{h_n^3}{6} \sum_{j=1}^d \sum_{l=1}^d \sum_{m=1}^d \frac{\partial^3 g}{\partial x_j \partial x_l \partial x_m}(x) \left(\frac{X_{ji} - x_j}{h_n}\right) \left(\frac{X_{li} - x_l}{h_n}\right) \left(\frac{X_{mi} - x_m}{h_n}\right) + o(h_n^3), \end{aligned}$$

where the remainder is of the stated order on the set $\{X_i : \|X_i - x\| < h_n\}$. Using symmetry, we have

$$B_{n0}(x) = \frac{h_n^2}{2} \mu_2(K) \sum_{j=1}^d g_{jj}(x) \{1 + o_p(1)\} \quad (20)$$

$$B_{nk}(x) = \frac{h_n^3}{6} \sum_{j=1}^d \sum_{l=1}^d \sum_{m=1}^d \int \mathcal{K}(u) u_k u_j u_l u_m du \{3g_{jl}(x) f_m(x) + g_{jlm}(x) f(x)\} \{1 + o_p(1)\}. \quad (21)$$

We shall approximate $M_n(x)$ by

$$M_n(x) = f(x)M + h_n M^*(x) + O(h_n^2) + O_p \left(\sqrt{\frac{\log n}{nh_n^d}} \right)$$

[the errors are uniform in x], where

$$M = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \mu_2(K) & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \mu_2(K) \end{bmatrix} ; \quad M^*(x) = \mu_2(K) \begin{bmatrix} 0 & f_1(x) & \dots & f_d(x) \\ f_1(x) & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ f_d(x) & & & 0 \end{bmatrix}.$$

Therefore,

$$\begin{aligned} M_n^{-1}(x) &= f^{-1}(x)M^{-1} - h_n f^{-2}(x)M^{-1}M^*(x)M^{-1} + O(h_n^2) + O_p \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \\ &= \frac{1}{f(x)} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \frac{1}{\mu_2(K)} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \frac{1}{\mu_2(K)} \end{bmatrix} - \frac{h_n}{f^2(x)} \begin{bmatrix} 0 & f_1(x) & \dots & f_d(x) \\ f_1(x) & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ f_d(x) & & & 0 \end{bmatrix} + O(h_n^2) + O_p \left(\sqrt{\frac{\log n}{nh_n^d}} \right). \end{aligned}$$

■

We next give the proofs of Theorems 1-4.

PROOF OF THEOREM 1 AND 2. By a geometric expansion and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} & \sup_{x \in \mathcal{X}} \left| \frac{\widehat{r}_k(x)}{\widehat{s}(x)} - \frac{r_k(x)}{s(x)} - \frac{\widehat{r}_k(x) - r_k(x)}{s(x)} + \frac{r_k(x)}{s(x)} \frac{\widehat{s}(x) - s(x)}{s(x)} \right| \\ &= \sup_{x \in \mathcal{X}} \left| \frac{\widehat{r}_k(x) - r_k(x)}{s(x)} \frac{\widehat{s}(x) - s(x)}{s(x)} + \frac{\widehat{r}_k(x)}{\widehat{s}(x)} \left\{ \frac{\widehat{s}(x) - s(x)}{s(x)} \right\}^2 \right| \\ &\leq \frac{\sup_{x \in \mathcal{X}} |\widehat{r}_k(x) - r_k(x)|}{\inf_{x \in \mathcal{X}} |s(x)|} \sup_{x \in \mathcal{X}} |\widehat{s}(x) - s(x)| + \frac{\sup_{x \in \mathcal{X}} |\widehat{r}_k(x)|}{\inf_{x \in \mathcal{X}} |\widehat{s}(x)s(x)|} \left\{ \sup_{x \in \mathcal{X}} |\widehat{s}(x) - s(x)| \right\}^2 \\ &= O_p \left(\frac{\log n}{nh_n^{d+1}} \right) + O_p(h_n^4), \end{aligned}$$

since $\sup_{x \in \mathcal{X}} |\widehat{r}_k(x)| = O_p(1)$ and

$$\begin{aligned} \inf_{x \in \mathcal{X}} |\widehat{s}(x)s(x)| &\geq \inf_{x \in \mathcal{X}} |s^3(x)| - \sup_{x \in \mathcal{X}} s^2(x) |\widehat{s}(x) - s(x)| \\ &\geq \left\{ \inf_{x \in \mathcal{X}} |s(x)| \right\}^3 + o_p(1). \end{aligned}$$

Therefore, we can restrict attention to the linearization

$$L_n(x) = \frac{\widehat{r}_k(x) - r_k(x)}{s(x)} - \frac{r_k(x)}{s(x)} \frac{\widehat{s}(x) - s(x)}{s(x)},$$

whose asymptotic distribution and rate of uniform convergence follows from Masry (1996a,b). Specifically, $\widehat{s}(x) - s(x) = O_p(n^{-1/2}h_n^{-d/2}) + O(h_n^{p+1})$, which is of smaller order than the first term, while $\{\widehat{r}_k(x) - r_k(x)\}/s(x) = O_p(n^{-1/2}h_n^{-(d+2)/2}) + O(h_n^p)$. ■

PROOF OF THEOREM 3. Let

$$Q_n(\delta) = \left\{ \begin{pmatrix} \widehat{D}r \\ \widehat{D}s \end{pmatrix} - g(\delta) \right\}' V_n^{-1} \left\{ \begin{pmatrix} \widehat{D}r \\ \widehat{D}s \end{pmatrix} - g(\delta) \right\}, \quad Q(\delta) = \left\{ \begin{pmatrix} Dr \\ Ds \end{pmatrix} - g(\delta) \right\}' V^{-1} \left\{ \begin{pmatrix} Dr \\ Ds \end{pmatrix} - g(\delta) \right\}.$$

Then, for any compact subset Δ of \mathbb{R}^{d+1} , we have

$$\sup_{\delta \in \Delta} |Q_n(\delta) - Q(\delta)| = o(1)$$

with probability one. The pointwise convergence follows by Theorem 5 of Masry (1996b). Uniform convergence follows from the quadratic form of Q_n and the smoothness of g . Furthermore, when the restrictions are true, $Q(\delta)$ is uniquely minimized by δ_0 . Therefore, $\widehat{\delta}$ is strongly consistent.

Asymptotic normality of $\widehat{\delta}$ follows from the following two results:

$$n^{2p/(d+2p+2)} \frac{\partial Q_n}{\partial \delta}(\delta_0) \implies N(2GV^{-1}b, 4GV^{-1}\Omega V^{-1}G') \quad (22)$$

and

$$\frac{\partial^2 Q_n}{\partial \delta \partial \delta'}(\delta_n) \longrightarrow_p 2G'V^{-1}G \quad (23)$$

for any sequence $\delta_n \rightarrow \delta_0$, see Amemiya (1985, Chapter 4). We have

$$\frac{\partial Q_n}{\partial \delta}(\delta_0) = 2G'V_n^{-1} \left\{ \begin{pmatrix} \widehat{D}r \\ \widehat{D}s \end{pmatrix} - g(\delta_0) \right\} = 2G'V^{-1} \left\{ \begin{pmatrix} \widehat{D}r \\ \widehat{D}s \end{pmatrix} - g(\delta_0) \right\} \{1 + o_p(1)\},$$

and (22) follows from an extension of Theorem 5 of Masry (1996b) to cover the joint asymptotic behaviour. The matrix G consists of ones, zeros, and δ_j , while the second derivatives of g are either

one or zero. Therefore, standard arguments can be applied to establish condition (23). To obtain the distribution of $g_1(\hat{\delta})$, we use the delta method

$$g_1(\hat{\delta}) - g_1(\delta) = G'_1(\hat{\delta} - \delta_0) + O_p(\|\hat{\delta} - \delta_0\|^2).$$

■

PROOF OF THEOREM 4. The first approximation we make is that

$$\begin{aligned} \hat{m}(x) &= \sum_{k=1}^d \frac{1}{J} \sum_{i=1}^J \frac{\hat{m}_k(z_k(X_{ki}^*))}{\varphi_k(X_{ki}^*)} \mathbf{1}(X_{ki}^* \in [\underline{x}_k, x_k]) \\ &= \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \hat{m}_k(z_k(t)) dt + O(J^{-1}) \end{aligned} \quad (24)$$

with probability one. Then, we say that

$$\begin{aligned} \hat{m}(x) - m(x) &= \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \hat{m}_k(z_k(t)) dt \\ &= \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \left\{ \frac{\hat{r}_k(z_k(t)) - r_k(z_k(t))}{s(z_k(t))} - \frac{r_k(z_k(t)) \hat{s}(z_k(t)) - s(z_k(t))}{s(z_k(t))} \right\} dt + O_p\left(\frac{\log n}{nh_n^{d+1}}\right) + O_p(h_n^4) \end{aligned}$$

by the uniform convergence result (11). The next step is to linearize $\hat{r}_k(z_k(t)) - r_k(z_k(t))$ and $\hat{s}(z_k(t)) - s(z_k(t))$ and integrate term by term.

We first turn to the stochastic part of our estimator $\hat{m}(x)$, which is

$$\frac{1}{h_n} \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \frac{e'_k M_n^{-1}(z_k(t)) U_n(z_k(t); r)}{s(z_k(t))} dt - \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \frac{r_k(z_k(t)) e'_k M_n^{-1}(z_k(t)) U_n(z_k(t); s)}{s(z_k(t))} dt, \quad (25)$$

where we use the notation $U_n(z_k(t); r)$ to show which regression function is involved. The second term is of smaller order.

Consider the term

$$\begin{aligned} \frac{1}{h_n} \int_{\underline{x}_k}^{x_k} \frac{e'_k M_n^{-1}(z_k(t)) U_n(z_k(t); r)}{s(z_k(t))} dt &= \frac{1}{h_n} \int_{\underline{x}_k}^{x_k} \frac{e'_k M^{-1} U_n(z_k(t); r)}{s(z_k(t)) f(z_k(t))} dt \\ &= \frac{1}{\mu_2(K)} \frac{1}{h_n} \int_{\underline{x}_k}^{x_k} \frac{U_{nk}(z_k(t); r)}{s(z_k(t)) f(z_k(t))} dt \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\mu_2(K)} \frac{1}{nh_n^{d+1}} \sum_{i=1}^n \epsilon_i \prod_{\ell < k} K\left(\frac{x_\ell - X_{\ell i}}{h_n}\right) \prod_{\ell > k} K\left(\frac{\underline{x}_\ell - X_{\ell i}}{h_n}\right) \\
&\quad \times \int_{\underline{x}_k}^{x_k} K\left(\frac{t - X_{ki}}{h_n}\right) \left(\frac{t - X_{ki}}{h_n}\right) \frac{dt}{(sf)(z_k(t))} \\
&= \frac{1}{nh_n^d} \sum_{i=1}^n \epsilon_i \prod_{\ell < k} K\left(\frac{x_\ell - X_{\ell i}}{h_n}\right) \prod_{\ell > k} K\left(\frac{\underline{x}_\ell - X_{\ell i}}{h_n}\right) \\
&\quad \times \left\{ \frac{L\left(\frac{x_k - X_{ki}}{h_n}\right)}{(sf)(z_k(x_k))} - \frac{L\left(\frac{\underline{x}_k - X_{ki}}{h_n}\right)}{(sf)(z_k(\underline{x}_k))} \right\} \{1 + o_p(1)\},
\end{aligned}$$

where the approximation error is small by dominated convergence arguments. Furthermore,

$$\begin{aligned}
\int_{\underline{x}_k}^{x_k} e'_k M^{-1} M^*(z_k(t)) M^{-1} f^{-2}(z_k(t)) U_n(z_k(t); r) dt &= \int_{\underline{x}_k}^{x_k} f_k(z_k(t)) f^{-2}(z_k(t)) U_{n0}(z_k(t); r) dt \\
&= O_p(n^{-1/2})
\end{aligned}$$

and is of smaller order. In conclusion, we have written the stochastic part of $\widehat{m}(x) - m(x)$ as

$$T_n = \frac{1}{\mu_2(K)} \sum_{k=1}^d \{T_{nk}(z_k(x_k)) - T_{nk}(z_k(\underline{x}_k))\} + o_p(n^{-1/2} h_n^{-d/2}), \quad (26)$$

where

$$T_{nk}(z_k(x_k)) = \frac{1}{nh_n^d} \sum_{i=1}^n \epsilon_i \prod_{\ell < k} K\left(\frac{x_\ell - X_{\ell i}}{h_n}\right) \prod_{\ell > k} K\left(\frac{\underline{x}_\ell - X_{\ell i}}{h_n}\right) L\left(\frac{x_k - X_{ki}}{h_n}\right) \frac{1}{(sf)(z_k(x_k))}.$$

Thus T_n has variance (14) because $z_k(x_k) = z_{k+1}(\underline{x}_{k+1})$, $k = 1, 2, \dots, d-1$, and

$$E [T_{nk}(z_k(x_k)) T_{n,k+1}(z_{k+1}(\underline{x}_{k+1}))] = \frac{1}{nh_n^d} \|K\|^{2(d-2)} \langle K, L \rangle^2 \frac{\sigma_r^2(z_k(x_k))}{(s^2 \cdot f)(z_k(x_k))},$$

while $E [T_{nk}(z_k(x_k)) T_{n,k+j}(z_{k+j}(\underline{x}_{k+j}))] = 0$ for all k and $j > 1$. The central limit theorem follows immediately from the representation (26).

We next examine the bias term. In this case, we must include terms from the estimation of s , that is, the bias is

$$\sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \left\{ \frac{\frac{1}{h_n} e'_k M_n^{-1}(z_k(t)) B_n(z_k(t); r)}{s(z_k(t))} - \frac{r_k(z_k(t)) e'_0 M_n^{-1}(z_k(t)) B_n(z_k(t); s)}{s(z_k(t))} \right\} dt. \quad (27)$$

We have

$$\begin{aligned} \frac{1}{h_n} \int_{\underline{x}_k}^{x_k} \frac{e'_k M^{-1}(z_k(t)) B_n(z_k(t); r)}{f(z_k(t)) s(z_k(t))} dt &= \frac{1}{h_n} \int_{\underline{x}_k}^{x_k} \frac{B_{nk}(z_k(t); r)}{f(z_k(t)) s(z_k(t)) \mu_2(K)} dt = O(h_n^2) \\ \int_{\underline{x}_k}^{x_k} \frac{e'_k M^{-1} M^*(z_k(t)) M^{-1} B_n(z_k(t); r)}{f^2(z_k(t)) s(z_k(t))} dt &= \int_{\underline{x}_k}^{x_k} \frac{f_k(z_k(t)) B_{n0}(z_k(t); r)}{f^2(z_k(t)) s(z_k(t))} dt = O(h_n^2). \end{aligned}$$

We then also subtract the bias term contributed by \widehat{s} , which is

$$- \sum_{k=1}^d \int_{\underline{x}_k}^{x_k} \left\{ \frac{r_k(z_k(t))}{s(z_k(t))} \frac{B_{n0}(z_k(t); s)}{f(z_k(t)) s(z_k(t))} \right\} dt = O(h_n^2).$$

The result follows by substituting in the expressions (20) and (21). ■

PROOF OF THEOREM 5. The asymptotic distribution of the average derivative estimator follows from the expansion

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n w(X_i) \{ \widehat{m}_k(X_i) - \delta_w \} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n w(X_i) \{ m_k(X_i) - \delta_w \} + \frac{1}{\sqrt{n}} \sum_{i=1}^n w(X_i) \frac{\widehat{r}_k(X_i) - r_k(X_i)}{s(X_i)} \\ &\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{w(X_i) r_k(X_i) \widehat{s}(X_i) - s(X_i)}{s(X_i)} + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n w(X_i) \{ m_k(X_i) - \delta_w \} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{w_k(X_i)}{s(X_i)} - \frac{w(X_i) s_k(X_i)}{s(X_i)} \right\} \varepsilon_i \\ &\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{w(X_i) r_k(X_i)}{s^2(X_i)} \eta_i + o_p(1). \end{aligned}$$

The first approximation follows from our uniform convergence results, while the second line is standard [almost] - it is established using integration by parts and bias reduction [the implementation of \widehat{m}_k has to use higher order polynomials as implied by the conditions]. ■

ACKNOWLEDGEMENTS

We would like to thank Joel Horowitz for sharing his research with us and Andrew Chesher for helpful comments. We would both like to thank the National Science Foundation for financial support.

References

- [1] AHN, H. (1995): “Nonparametric two-stage estimation of conditional choice probabilities in a binary choice model under uncertainty,” *Journal of Econometrics* **67**, 337-378.
- [2] AMEMIYA, T. (1973), “Regression Analysis When the Dependent Variable is Truncated Normal,” *Econometrica*, 41, 997–1016.
- [3] AMEMIYA, T. (1985) *Advanced Econometrics*. Harvard University Press.
- [4] ANDREWS, D. W. K., (1995), “Nonparametric Kernel Estimation for Semiparametric Models,” *Econometric Theory*, 11, 560–596.
- [5] BUCHINSKY, M. AND J. HAHN, “An Alternative Estimator for the Censored Quantile Regression Model,” *Econometrica*, 66, 653-671.
- [6] BUCKLEY, J. AND I. JAMES, (1979), “Linear Regression With Censored Data,” *Biometrika*, 66, 429–436.
- [7] COLLOMB, G. AND W. HÄRDLE (1986), “Strong Uniform Convergence Rates in Robust Nonparametric Time Series Analysis and Prediction: Kernel Regression Estimation From Dependent Observations,” *Stochastic Processes and Their Applications*, 23, 77–89.
- [8] DUNCAN, G. M., (1986), “A Semi-parametric Censored Regression Estimator,” *Journal of Econometrics*, 32, 5–24.
- [9] FAN, J., AND I. GIJBELS (1996), *Local Polynomial Modelling and Its Applications* Chapman and Hall.
- [10] FERNANDEZ, L., (1986), “Non-parametric Maximum Likelihood Estimation of Censored Regression Models,” *Journal of Econometrics*, 32, 35–57.
- [11] HÄRDLE, W., AND O.B. LINTON (1994): “Applied nonparametric methods,” *The Handbook of Econometrics*, vol. IV, eds. D.F. McFadden and R.F. Engle III. North Holland.
- [12] HÄRDLE, W. AND T. M. STOKER (1989), “Investigating Smooth Multiple Regression by the Method of Average Derivatives,” *Journal of the American Statistical Association*, 84, 986–995.
- [13] HAUSMAN, J. A. AND W. K. NEWEY (1995), “Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss,” *Econometrica*, 63, 1445–1476.

- [14] HECKMAN, J. J. (1976), “The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models,” *Annals of Economic and Social Measurement*, 15, 475–492.
- [15] HONORÉ, B. E. AND J. L. POWELL, (1994), “Pairwise Difference Estimators of Censored and Truncated Regression Models,” *Journal of Econometrics*, 64, 241–278.
- [16] HOROWITZ, J. L., (1986), “A Distribution Free Least Squares Estimator for Censored Linear Regression Models,” *Journal of Econometrics*, 32, 59-84.
- [17] HOROWITZ, J. L., (1988), “Semiparametric M-Estimation of Censored Linear Regression Models,” *Advances in Econometrics*, 7, 45-83.
- [18] HOROWITZ, J. L., (1998), “Nonparametric estimation of a generalized additive model with an unknown link function,” Iowa City Manuscript.
- [19] ICHIMURA, H. (1993), “Semiparametric Least Squares (SLS) and Weighted SLS estimation of Single-index Models,” *Journal of Econometrics*, 58, 71–120.
- [20] KOUL, H., V. SUSLARA, AND J. VAN RYZIN (1981), “Regression Analysis With Randomly Right Censored Data,” *Annals of Statistics*, 42, 1276–1288.
- [21] LEWBEL, A. (1995), “Consistent Nonparametric Tests With An Application to Slutsky Symmetry,” *Journal of Econometrics*, 67, 379–401.
- [22] LEWBEL, A. (1997), “Semiparametric Estimation of Location and Other Discrete Choice Moments,” *Econometric Theory*, 13, 32-51.
- [23] LEWBEL, A. (1998a), “Semiparametric Latent Variable Model Estimation With Endogenous or Mismeasured Regressors,” *Econometrica*, 66, 105–121.
- [24] LEWBEL, A. (1998b), “Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables,” Unpublished Manuscript.
- [25] MCDONALD, J. AND R. MOFFITT (1980), “The Uses of Tobit Analysis,” *Review of Economics*, 62, 318–321.
- [26] MADDALA, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Econometric Society Monograph No. 3, Cambridge: Cambridge University Press.

- [27] MASRY, E. (1996a), “Multivariate local polynomial regression for time series: Uniform strong consistency and rates,” *J. Time Ser. Anal.* 17, 571-599.
- [28] MASRY, E., (1996b), “Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and their Applications* 65, 81-101.
- [29] MOON, C.-G., (1989), “A Monte Carlo Comparison of Semiparametric Tobit Estimators. *Journal of Applied Econometrics*, 4, 361-382.
- [30] NAWATA, K. (1990), “Robust Estimation Based on Group-Adjusted Data in Censored Regression Models,” *Journal of Econometrics*, 43, 337–362.
- [31] NEWEY, W. K. (1994), “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–1382.
- [32] POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989), “Semiparametric Estimation of Index Coefficients,” *Econometrica* 57, 1403–1430.
- [33] POWELL, J. L. (1984), “Least Absolute Deviations Estimation for the Censored Regression Model,” *Journal of Econometrics*, 25, 303–325.
- [34] POWELL, J. L. (1986a), “Symmetrically Trimmed Least Squares Estimation For Tobit Models,” *Econometrica*, 54, 1435–1460.
- [35] POWELL, J. L. (1986b), “Censored Regression Quantiles,” *Journal of Econometrics*, 32, 143–155.
- [36] RILESTONE, P. (1996): “Nonparametric estimation of models with generated regressors,” *International Economic Review* 37, 299-313.
- [37] ROBINSON, PETER M. (1982), “On the Asymptotic Properties of Estimators of Models Containing Limited Dependent Variables,” *Econometrica*, 50, 27-41.
- [38] ROBINSON, PETER M. (1988), “Root- N -Consistent Semiparametric Regression,” *Econometrica*, 56, 931–954.
- [39] ROSETT, R. AND F. NELSON (1975), “Estimation of the Two-Limit Probit Regression Model,” *Econometrica*, 43, 141–146.
- [40] SILVERMAN, B. W. (1978) “Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and its Derivatives,” *Annals of Statistics*, 6, 177–184.

- [41] STOKER, THOMAS M. (1991), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, W. A. Barnett, J. Powell, and G. Tauchen, Eds., Cambridge University Press.
- [42] STONE, C.J. (1980), "Optimal rates of convergence for nonparametric estimators," *Annals of Statistics* **8**, 1348-1360.
- [43] STONE, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* **8**, 1040-1053.
- [44] VARIAN, H.R. (1984): *Microeconomic Analysis*. W.H. Norton & Company: New York.

Figures overleaf show results of estimation in the model $y = y^*1(y^* \geq 0)$, where $y^* = x^3 + e$, where x is uniformly distributed and e is standard normal. Sample size is $n = 1000$ and $h_n = 2n^{-1/7}$.