A Simple Ordered Data Estimator for Inverse Density Weighted Expectations

Arthur Lewbel* and Susanne M. Schennach
Boston College and University of Chicago

Abstract

We consider estimation of means of functions that are scaled by an unknown density, or equivalently, integrals of conditional expectations. The "ordered data" estimator we provide is root n consistent, asymptotically normal, and is numerically extremely simple, involving little more than ordering the data and summing the results. No sample size dependent smoothing is required. A similarly simple estimator is provided for the limiting variance. The proofs include new limiting distribution results for functions of nearest neighbor spacings. Potential applications include endogeneous binary choice, willingness to pay, selection, and treatment models.

JEL codes: C14, C21, C25, J24

Keywords: Semiparametric, Conditional Expectation, Density Estimation, Binary Choice, Binomial Response.

*Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. phone: 617-552-3678, fax: 617-552-2308, lewbel@bc.edu, http://www2.bc.edu/~lewbel/

1 Introduction

Given an iid dataset $(x_1, y_1), \ldots, (x_n, y_n)$, on a bivariate random variable (x, y), we propose to estimate an expectation of the inverse density weighted form

$$\theta = E\left[\frac{y}{f(x)}\right] \tag{1}$$

where $f(x)$ is the unknown density of the continuously distributed x. The "ordered data" estimator we provide possesses the rather surprising property of achieving root n consistency and asymptotic normality without requiring sample-size-dependent smoothing. It also offers the advantage of being numerically extremely simple, requiring little more than ordering the data and summing the results, thereby avoiding issues regarding the selection of smoothers such as bandwidths, kernels, etc. A similarly simple estimator is provided for the limiting variance.

Inverse density weighted estimation applies generically to the estimation of definite integrals of conditional expectations. Suppose

$$\theta = \int_{x \in \mathcal{X}} E(w|x)dx \tag{2}$$

for some random variable w, where $\mathcal{X} \subset \text{supp}(x)$. Then θ can be rewritten as $\theta = E[y/f(x)]$ with $y = wI(x \in \mathcal{X})$, and where I is the indicator function that equals one if its argument is true and zero otherwise.

A number of existing semiparametric estimators make use of inverse density weighted expectations either directly or indirectly, via their relationship with integrated conditional expectations. Examples include density weighted least squares (Newey and Ruud, 1984), average derivative estimation (Härdle and Stoker, 1993), estimators of willingness-to-pay models and general estimators of moments from binomial data (Lewbel, 1997, McFadden, 1999, Lewbel, Linton and McFadden, 2002), semiparametric estimators of consumer surplus (Hausman and Newey, 1995, Newey, 1997), some discrete choice, sample selection, and other latent variable model estimators (Lewbel, 1998, 2000, 2002), entropy measures of dependence (Hong and White, 2000) and semiparametric functional tests (Hall and Yatchew, 2005).

Let $\left(y_{[i]}, x_{[i]}\right)$ denote the i-th observation when the data are sorted in increasing order of x, so $x_{[i]}$ is the i-th order statistic and $y_{[i]}$ is the concomitant statistic to $x_{[i]}$. Let (y_i, x_i) denote the i-th observation when the data are left unsorted. Let $F(x)$ denote the unknown distribution function of x. We show that the

numerically trivial "ordered data" estimator

$$\hat{\theta} = \sum_{i=1}^{n-1} \left(y_{[i+1]} + y_{[i]}\right)\left(x_{[i+1]} - x_{[i]}\right)/2 \tag{3}$$

is root n consistent and asymptotically normal. Specifically, $n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, 3\sigma^2/2)$ where

$$\sigma^2 = E\left[\frac{\mathrm{Var}(y|x)}{f^2(x)}\right] \tag{4}$$

and a consistent estimator of σ^2 is the simple expression

$$\hat{\sigma}^2 = \frac{n}{4}\sum_{i=1}^{n-1}\left(y_{[i+1]} - y_{[i]}\right)^2\left(x_{[i+1]} - x_{[i]}\right)^2. \tag{5}$$

For some intuition for this estimator, let x be a point that lies between $x_{[i]}$ and $x_{[i+1]}$ for some i, and let y be a corresponding point that lies between $y_{[i]}$ and $y_{[i+1]}$. Then $y \approx \left(y_{[i+1]} + y_{[i]}\right)/2$ and

$$\frac{1}{f(x)} = \left(\frac{dF(x)}{dx}\right)^{-1} \approx \frac{x_{[i+1]} - x_{[i]}}{F(x_{[i+1]}) - F(x_{[i]})} \approx \frac{x_{[i+1]} - x_{[i]}}{1/n} \tag{6}$$

where the last step results from replacing F with the corresponding empirical distribution function. The estimator $\hat{\theta}$ is then just an average of $y/f(x)$, using these approximations for y and $f(x)$. This differencing of the empirical distribution function for x does not yield a consistent estimator for $1/f(x)$, but averaging over x speeds the rate of convergence to yield root n consistency. This result is rather exceptional among semiparametric estimators of nonlinear functionals of the data generating process in that it attains root n consistency without the use of explicit smoothing techniques.

The ability to bypass kernel and bandwidth selection procedures constitutes a substantial benefit of the proposed estimator. The well-known bandwidth selection rules used in nonparametric estimation, such as cross-validation, are not generally applicable to semiparametric settings because the optimal bandwidth in nonparametric settings typically fails to undersmooth at the rate needed to achieve a $o\left(n^{-1/2}\right)$ bias in the semiparametric functional. In practice, a bandwidth "slightly" smaller than the one given by cross-validation is often used instead, in an attempt to undersmooth. As the required amount of undersmoothing is a sample-size-dependent quantity, this informal method does not provide much guidance and may lead to a nonnegligible bias. The derivation and application of formal data-driven bandwidth selection rules in semiparametric settings is rarely done, as it involves technical higher-order asymptotic analyses (see,

for instance, Härdle *et al.*, 1992, Hall and Horowitz, 1990) that must be rederived for each semiparametric estimator of interest, and may depend strongly on the unknown precise degree of smoothness of the estimand.

The method for approximating the reciprocal of a density using longer spacings to achieve consistency is known (see, e.g., Bloch and Gastwirth, 1968). Our contribution consists mainly of providing limiting distribution theory for an average of such estimators, for fixed or increasing spacings. The complication that arises in doing so is accounting for the fact that the spacings $x_{[i+1]} - x_{[i]}$ are not independent. In fact, the statistical dependence among them is of a form that is not covered by standard central limit theorems for dependent processes. One spacing depends equally strongly on arbitrarily distant spacings and not only on its neighbors. To handle this difficulty, we substantially extend a technique of proof proposed by Weiss (1958) in the case of homogenous functions of spacings of a sample drawn from a uniform distribution to cover the more general functional (1).

While the use of nearest neighbors in differencing schemes in semiparametric[1] settings is not new (e.g., Yatchew, 1997), our use of a statistic based on nearest neighbor *spacings* is innovative. The asymptotic analysis of differencing techniques only relies on the fact that spacings converge to zero as sample size increases, while our results require an analysis of their asymptotic distribution.

In addition to providing the limiting distribution of an estimator of (1), we provide an extension of the limiting distribution theory to the case where x itself is estimated and consider a generalizations to multivariate y. We also show how the ordered data estimator could be used in some semiparametric models, we provide a small Monte Carlo analysis of the estimator, and we apply the estimator in a small empirical study. Proofs are in the Appendix.

2 Asymptotics

2.1 Main results

The derivation of the asymptotic properties of our "ordered data" estimator relies on a few standard assumptions.

Condition 1 *(y_i, x_i) is a sequence of i.i.d. random variables.*

[1] The use of differencing using k-nearest neighbors, where $k \to \infty$ as $n \to \infty$ is also well-established (see, e.g. Robinson, 1987) but our asymptotic analysis under fixed k is fundamentally different.

Condition 2 *The support of $f(x)$, denoted $\mathcal{F}$, is a finite interval and $\inf_{x\in\mathcal{F}} f(x) > 0$.*

Condition 3 *$f(x)$ is uniformly Hölder continuous of exponent $h_f > 1/2$ over the interior of its support (i.e. $|f(x) - f(\xi)| \leq H_f |x-\xi|^{h_f}$ for $x, \xi \in \mathcal{F}$).*

Condition 4 *$g(x) = E[y|x]$ is uniformly Hölder continuous of exponent $h_g > 1/2$ over the interior of the support of $f(x)$ (i.e. $|g(x) - g(\xi)| \leq H_g |x-\xi|^{h_g}$ for $x, \xi \in \mathcal{F}$).*

Condition 5 *$E\left[\frac{\mathrm{Var}[y|x]}{f^2(x)}\right] < \infty$.*

Assumption 1 is very common in cross-sectional analysis. Assumptions 3 and 4 impose smoothness constraints on $f(x)$ and $E[y|x]$ that are slightly stronger than continuity and weaker than assuming that $f(x)$ and $E[y|x]$ are Lipschitz (which would corresponds to $h_f = h_g = 1$). Assumption 5 is crucial in order to obtain root n consistency but could probably be relaxed using sample size-dependent trimming if only consistency is desired. Assumption 2 is frequently made in the literature focusing on inverse density weighted estimators. It can probably be relaxed at the expense of substantial complications in the proofs, however, we do not pursue such extensions in the present work, because Assumption 5 is rarely satisfied when $f(x)$ is not bounded away from zero.[2]

Theorem 1 *Under Assumptions 1 through 5, $n^{1/2}\left(\tilde{\theta} - \theta\right) \xrightarrow{d} N\left(0, 2\sigma^2\right)$ where*

$$\tilde{\theta} = \sum_{i=1}^{n-1} y_{[i]}\left(x_{[i+1]} - x_{[i]}\right) \tag{7}$$

$$\theta = E\left[\frac{y}{f(x)}\right] = \int E[y|x]\, dx \tag{8}$$

$$\sigma^2 = E\left[\frac{\mathrm{Var}[y|x]}{f^2(x)}\right]. \tag{9}$$

[2] Another paper that employs this support restriction, for similar reasons, is Abadie and Imbens (2002). This support restriction can likely be relaxed using trimming or boundary methods similar to those used for kernel estimators, though doing so may require the introduction of data dependent parameters (such as asymptotic trimming terms). One of the main advantages of the estimator is that it does not require selection of data dependent parameters like bandwidths. Alternatively, it is always possible to map a variable with unbounded support onto one with a bounded support by using a fixed nonlinear mapping. Our estimator will be root n consistent for this transformed variable, provided that the Jacobian of this known transformation is included in our inverse-density weighted expectation. This would just amount to redefining the y variable and verifying that the resulting $E[y|x]$ is sufficiently smooth and bounded.

Corollary 2 *Under Assumptions 1 through 5,* $n^{1/2}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N\left(0,(3/2)\,\sigma^2\right)$ *where*

$$\hat{\theta}=\sum_{i=1}^{n-1}\left(y_{[i]}+y_{[i+1]}\right)\left(x_{[i+1]}-x_{[i]}\right)/2. \tag{10}$$

and where θ *and* σ^2 *are defined in Theorem 1.*

The fundamental reason why root n consistency is possible without smoothing is because spacings directly estimate the reciprocal of the density and θ is a linear functional of the reciprocal of the density, thus allowing substantial undersmoothing without introducing nonnegligible nonlinear remainder terms. Kernel-based estimation would first proceed by estimating the density and then take its reciprocal, a nonlinear operation that precludes the use of such substantial undersmoothing. The need for smoothing in nonlinear functional estimation, but not in linear functional estimation, parallels Newey's linearization condition (Newey, 1994, Assumption 5.1), which places different constraints on the rate of convergence of plugged-in nonparametric estimates, depending on whether the functional of interest is linear or not.

The factor 2 in the variance of $\tilde{\theta}$ was also noted by Mack and Müller (1988) in the related problem of estimating a nonparametric conditional expectation using the Priestley and Chao (1972) estimator. See also Wand and Jones (1995, p. 131). Corollary 2 reduces this factor to 3/2. This factor can be further reduced to arbitrarily close to one by considering longer spacings. For instance, it can be shown along the same lines as Theorem 1 that

$$\check{\theta}_k=\sum_{i=1}^{n-k} y_{[i]}\left(x_{[i+k]}-x_{[i]}\right)/k. \tag{11}$$

has an asymptotic variance equal to $\sigma^2\left(1+\frac{1}{k}\right)$. Note that for $k=2$, Equation (11) is asymptotically equivalent to Equation (10).[3] Centered differences can also be used, e.g.

$$\hat{\theta}_c=\sum_{i=2}^{n-1} y_{[i]}\left(x_{[i+1]}-x_{[i-1]}\right)/2. \tag{12}$$

is asymptotically equivalent to both $\hat{\theta}$ and $\tilde{\theta}_2$.

The following theorem shows that when the length of the spacings grows with sample size, the asymptotic variance reaches the value σ^2, which can be shown to be semiparametrically efficient using the variance expressions in Newey (1994).[4]

[3] This can be shown along the following lines $\sum_{i=1}^{n-1} y_{[i]}\left(x_{[i+1]}-x_{[i]}\right)/2 + \sum_{i=1}^{n-1} y_{[i+1]}\left(x_{[i+1]}-x_{[i]}\right)/2 = \sum_{i=1}^{n-1} y_{[i]}\left(x_{[i+1]}-x_{[i]}\right)/2+\sum_{i=2}^{n} y_{[i]}\left(x_{[i]}-x_{[i-1]}\right)/2 \approx \sum_{i=2}^{n-1} y_{[i]}\left(x_{[i+1]}-x_{[i-1]}\right) \approx \sum_{i=1}^{n-2} y_{[i]}\left(x_{[i+2]}-x_{[i]}\right)$.

[4] Let the quantities defined in Newey (1994) be denoted by a N subscript. Our estimator can be written in terms of a nonparametric estimate $h_N(x)$ of the conditional expectation $E[y|x]$ as $\hat{\theta}=\int h_N(x)\,dx=\int \frac{h_N(x)}{f(x)} f(x)\,dx = E\left[\frac{h_N(x)}{f(x)}\right] =$

Theorem 3 *Let k_n be a deterministic sequence of integers such that $k_n = o(\ln n)$ and $k_n \to \infty$. Under Assumptions 1 through 5, $n^{1/2}\left(\tilde{\theta}_{k_n} - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right)$ where $\tilde{\theta}_k$ is given in Equation (11) and where θ and σ^2 are defined in Theorem 1.*

While Theorem 3 requires an upper bound on the rate at which k_n can grow, it imposes *no lower bound* (i.e. k_n can diverge arbitrarily slowly). This is unusual among semiparametric estimators, which typically require both an upper bound and a lower bound, in order to control both the bias and the variance. Here, the variance is already finite even when no smoothing is performed and hence, the mild requirement that $k_n \to \infty$ is sufficient to reduce the asymptotic variance term $2\sigma^2$ to its efficient value of σ^2.

The variance term σ^2 can be consistently estimated by[5]

$$\hat{\sigma}^2 = \frac{n}{4}\sum_{i=1}^{n-1}\left(y_{[i+1]} - y_{[i]}\right)^2\left(x_{[i+1]} - x_{[i]}\right)^2, \tag{13}$$

as established in the following Theorem.

Theorem 4 *Under Assumptions 1 through 5, $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.*

One important point needs to be emphasized regarding the asymptotic variance σ^2. If f where known, the most obvious estimator of θ would be $\hat{\theta}_0 \equiv n^{-1}\sum_{i=1}^{n} y_i/f(x_i)$. An interesting feature of this model is that $\hat{\theta}_0$ is inefficient, that is, it is more efficient to plug an appropriate estimator for $f(x_i)$ into this sample average than to use the true f. This can be immediately verified by

$$\begin{aligned}\operatorname{Var}[y/f(x)] &= E[\operatorname{Var}[y|x]/f^2(x)] + \operatorname{Var}[E(y|x)/f(x)] \\ &= \sigma^2 + \operatorname{Var}[E(y|x)/f(x)] \geq \sigma^2.\end{aligned} \tag{14}$$

Similar efficiency gains from using estimated rather than true functions has been observed in a model that is scaled by a conditional density function (see Magnac and Maurin, 2003) and in a model scaled by a propensity score function (see Hirano, Imbens and Ridder, 2003).

$E[m_N(z_N, h_N)]$ where $z_N = (x, y)$ and $m_N(z_N, h_N) = \frac{h_N(x)}{f(x)}$. Since $m_N(z_N, h_N)$ is linear in $h_N(x)$, Equation 4.4 in Newey (1994) holds for $\delta_N(x) = (f(x))^{-1}$ and Proposition 4 yields the influence function $\alpha_N(z) = (f(x))^{-1}(y - E[y|x])$, the variance of which is $E\left[\operatorname{Var}[y|x]/(f(x))^2\right] \equiv \sigma^2$.

[5] The idea of employing $\left(y_{[i+1]} - y_{[i]}\right)$ to "differentiate out" the contribution of $E[y|x]$ has been used for instance by Yatchew (1997) to estimate the partially linear model and by Abadie and Imbens (2002) to estimate the variance of a matching estimator of treatment effects models. However, our variance estimator combines differencing and spacings, making it impossible to use these existing results.

2.2 Extensions

2.2.1 Multivariate case

The result presented so far can be extended in various directions. First, the scalar y can be replaced with a vector, as follows.

Corollary 5 *Let v_i for $i = 1, \ldots, n$ denote observations from random vector v taking value in $\mathbb{R}^k$ and let $v_{[i]}$ denote the concomitant statistic to $x_{[i]}$. If Assumptions 1 through 5 hold for $y_i = v_{i1}, \ldots, v_{ik}$, then $n^{1/2}\left(\hat{\psi} - \psi\right) \xrightarrow{d} N\left(0, (3/2)\, V\right)$ where $\hat{\psi} = \sum_{i=1}^{n-1}\left(v_{[i]} + v_{[i+1]}\right)\left(x_{[i+1]} - x_{[i]}\right)/2$,$\psi = E\left[v/f\left(x\right)\right] = \int E\left[v|x\right] dx$ and $V = E\left[\left(E\left[vv'|x\right] - E\left[v|x\right] E\left[v'|x\right]\right) f^{-2}\left(x\right)\right]$. Moreover, $\hat{V} \xrightarrow{p} V$, where*

$$\hat{V} = \frac{n}{4}\sum_{i=1}^{n-1}\left(v_{[i+1]} - v_{[i]}\right)\left(v_{[i+1]} - v_{[i]}\right)'\left(x_{[i+1]} - x_{[i]}\right)^2. \tag{15}$$

2.2.2 Estimated x

In some applications, it is useful to be able to use an estimated value of x. While the ordered data estimator requires no modification *per se* to handle this generated regressor, its asymptotic variance needs to account for the error in the estimation of x in the first step. For this purpose, let x be a random scalar and w be a random vector taking a value in $\mathbb{R}^{N_w}$ that are related through $x = X\left(w, \gamma\right)$ for some function $X : \mathbb{R}^{N_w} \times \mathbb{R}^{N_\gamma} \mapsto \mathbb{R}$ and for some parameter vector γ whose true value is γ^*. Letting $\theta\left(\gamma\right) = E\left[y/f_{x|\gamma}\left(x|\gamma\right)\right]$, we can then establish the following result.

Condition 6 *The support of $f_{x|\gamma}\left(x|\gamma\right)$, denoted $\mathcal{F}_\gamma$, is a finite interval and $\inf_{\gamma\in\Gamma}\inf_{x\in\mathcal{F}} f_{x|\gamma}\left(x|\gamma\right) > 0$, where Γ is some neighborhood of γ^*.*

Condition 7 $\left|f_{x|\gamma}\left(x|\gamma\right) - f_{x|\gamma}\left(\xi|\gamma\right)\right| \leq H_f\left|x - \xi\right|^{h_f}$ *for $x, \xi \in \mathcal{F}_\gamma$ and any $\gamma \in \Gamma$.*

Condition 8 $\left|g\left(x\right) - g\left(\xi\right)\right| \leq H_g\left|x - \xi\right|^{h_g}$ *for $x, \xi \in \cup_{\gamma\in\Gamma}\mathcal{F}_\gamma$.*

Condition 9 $E\left[\sup_{\gamma\in\Gamma}\mathrm{Var}\left[y|x\right]/f^2_{x|\gamma}\left(x|\gamma\right)\right] < \infty$.

Condition 10 $E\left[\sup_{\gamma\in\Gamma}\left|\partial X\left(w,\gamma\right)/\partial\gamma\right|\right] < \infty$.

Condition 11 $\partial E\left[y/f_{x|\gamma}\left(X\left(w,\gamma\right)|\gamma\right)\right]/\partial\gamma = E\left[\partial\left(y/f_{x|\gamma}\left(X\left(w,\gamma\right)|\gamma\right)\right)/\partial\gamma\right]$ *for all $\gamma \in \Gamma$ and $\partial E\left[y/f_{x|\gamma}\left(X\left(w,\gamma\right)|\gamma\right)\right]/\partial\gamma$ is continuous in γ at γ^*.*

Theorem 6 *Let $\tilde{\theta}_{k_n}(\gamma)$ and k_n respectively denote the estimator and the sequence described in Theorem 3. Let $\hat{\gamma}$ be a consistent estimate of γ^* with influence function[6] Ψ_i. Under Assumptions 1 and 6 through 11, $n^{1/2}\left(\tilde{\theta}_{k_n}(\hat{\gamma}) - \theta(\gamma^*)\right) \to N(0, V)$ where*

$$V = E\left[\frac{(y - E[y|w,\gamma^*])^2}{f^2_{x|\gamma}(x|\gamma^*)}\right] + 2B'E\left[\frac{(y - E[y|w,\gamma^*])}{f_{x|\gamma}(x|\gamma^*)}\Psi_i\right] + B'E[\Psi_i\Psi_i']B \tag{16}$$

$$B = E\left[\frac{y}{f_{x|\gamma}(x|\gamma^*)}\frac{\partial}{\partial x}E\left[\frac{\partial X(w,\gamma^*)}{\partial\gamma}|x\right]\right] + \\ -E\left[\frac{y}{\left(f_{x|\gamma}(x|\gamma^*)\right)^2}\frac{\partial f_{x|\gamma}(x|\gamma^*)}{\partial x}\left(\frac{\partial X(w,\gamma^*)}{\partial\gamma} - E\left[\frac{\partial X(w,\gamma^*)}{\partial\gamma}|x\right]\right)\right]. \tag{17}$$

Moreover, if $E[\partial X(w,\gamma^*)/\partial\gamma|x,y] = E[\partial X(w,\gamma^*)/\partial\gamma]$, then $B = 0$.

Assumptions 6 through 9 replace Assumptions 2 through 5 to ensure that the regularity conditions hold uniformly for $\gamma \in \Gamma$, since $\tilde{\theta}_{k_n}(\gamma)$ is a random function instead of a random scalar.

Theorem 6 assumes $k_n \to \infty$, omitting the case where the spacing width k is fixed. With fixed spacings the function $\tilde{\theta}_k(\gamma)$ exhibits a peculiar mode of convergence to $\theta(\gamma)$, in that the derivative $d\tilde{\theta}_k(\gamma)/d\gamma$ diverges as $n \to \infty$ almost everywhere and changes discontinuously every time the order of the $x_{[i]}$ changes. As a result, we do not know if $n^{1/2}\left(\tilde{\theta}_k(\hat{\gamma}) - \theta(\hat{\gamma})\right) - n^{1/2}\left(\tilde{\theta}_k(\gamma^*) - \theta(\gamma^*)\right) \xrightarrow{p} 0$ with fixed k. The function $\tilde{\theta}_{k_n}(\gamma)$ with $k_n \to \infty$ does not exhibit these problems.

3 Monte Carlo

Here we provide a Monte Carlo analysis to assess the small sample behavior of the estimator. We draw x_i and e_i as independent standard normals and let $y_i = 2x_i(1 + e_i)I(0 < x_i < 1)$. Table 1 reports results from estimating $\theta = E[y/f(x)] = 1$ using this simulated data. The sample size is $n = 100$, the number of replications is $10,000$, and the reported summary statistics are, respectively, the mean, standard deviation, quartiles (lower, median, upper), root mean squared error, mean absolute error, and median absolute error.

We report results for seven estimators. The first is $\widehat{\theta}_0 = n^{-1}\sum_{i=1}^n y_i/f(x_i)$, an estimator that uses the true, normal, density function $f(x)$, which in a typical application would be unknown. The next is $\widehat{\theta}_1 = \sum_{i=1}^{n-1}\left(y_{[i+1]} + y_{[i]}\right)\left(x_{[i+1]} - x_{[i]}\right)/2$, the ordered data estimator of Corollary 2. We also report $\widehat{\theta}_k = \sum_{i=1}^{n-k}\left(y_{[i+k]} + y_{[i]}\right)\left(x_{[i+k]} - x_{[i]}\right)/(2k)$ for $k = 2$ and $k = 3$, which are ordered data estimators with wider

[6] That is, $n^{1/2}(\hat{\gamma} - \gamma^*) = n^{1/2}\sum_{i=1}^n \Psi_i + o_p(1)$ where Ψ_i is i.i.d with mean zero finite variance.

spacings. The next estimator is $\widehat{\theta}_4 = n^{-1} \sum_{i=1}^{n} y_i / \widehat{f}(x_i, b)$ where $\widehat{f}(x, b)$ is a kernel density estimator, using a quartic kernel and bandwidth b given by Silverman's rule of thumb. Finally, $\widehat{\theta}_5$ and $\widehat{\theta}_6$ are the same kernel density based estimators, except using bandwidths $b/2$ and $2b$, respectively.

In terms of mean squared, mean absolute, or median absolute error, the estimator having the best fit is $\widehat{\theta}_5$, though this kernel estimator also has the most mean and median bias. This estimator is undersmoothed relative to (Silverman's approximation of) pointwise optimality. Efficiency and root n convergence of plug-in kernel estimators require undersmoothing relative to pointwise optimality of the nonparametric density estimator (see, e.g., Newey, 1994).

The ordered data estimator $\widehat{\theta}_1$ had the smallest mean and median bias of all the estimators, but somewhat larger mean squared errors. These errors decrease as expected as the spacings increase. The estimator using the true density, $\widehat{\theta}_0$, has largest errors, at least in part reflecting the inefficiency of that estimator as discussed earlier.

4 Examples

4.1 Latent Moments From Binomial Data

Consider a model of the form $d_i = I(w_i > x_i)$, where d is an observed dummy variable, x is an observed continuously distributed random variable, and w is an unobserved latent random variable that is drawn from a distribution that is independent of x. The goal is estimation of moments of w.

Problems like this arise in survey research, where w is an attribute of an individual such as wealth or willingness to pay for a public good. The individual is asked if w exceeds some randomly chosen value x, and d is the response. This form of survey design is used, because it is likely to produce less biased responses than directly asking for w (see, e.g., McFadden, 1999 and reference therein).

This model may also be applied in destructive testing, e.g., w could be the speed at which a car safety device fails, x would be the speed at which the car was tested, and d an indicator of outcome failure, such as whether a test dummy was injured. Similarly, in bioassay w might be the time required for an animal to suffer an abnormality, x is the time at which the animal is sacrificed to test for the abnormality, d is indicator of the test result for abnormality at time x.

Let $\theta = E(w^\lambda) - c^\lambda$ where λ is a moment chosen for estimation and c is any chosen element of the support

of x (e.g., c could be the median of x). A special case of results in Lewbel, Linton, and McFadden (2002) is

$$\theta = E\left(\frac{\lambda x^{\lambda-1}[d - I(x < c)]}{f(x)}\right), \tag{18}$$

assuming that $\text{supp}\,(w) \subset \text{supp}\,(x)$.[7] Therefore, letting $y_i = \lambda x_i^{\lambda-1}[d_i - I(x_i < c)]$, a simple estimator of $E(w^\lambda)$ is $\hat{\theta} + c^\lambda$ where $\hat{\theta} = \sum_{i=1}^{n-1}\left(y_{[i+1]} + y_{[i]}\right)\left(x_{[i+1]} - x_{[i]}\right)/2$, $n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, 3\sigma^2/2)$, and $\hat{\sigma}^2 = \sum_{i=1}^{n-1}\left(y_{[i+1]} - y_{[i]}\right)^2\left(x_{[i+1]} - x_{[i]}\right)^2 n/4$.

4.2 Selection and Treatment Effects

Consider the model $y_i = y_i^* d_i$, $d_i = I(0 \leq w_i + x_i \leq a)$, where a is a constant (which could equal infinity), y_i is an individual's observed outcome, d is an observed dummy variable that indicates if the individual is selected or treated, x is an observed continuously distributed random variable, and y^* and w are unobserved latent random variables that are drawn from a distribution that is independent of x. The goal is estimation of moments of the potential outcome y^*.

An example is a wage model, where y^* is an individual's wage if employed, y is the individual's observed wage, d is the indicator of whether an individual is employed, a is infinite, $-x$ is some form of nonwage income such as a government defined benefit, and w is a latent variable such that the individual chooses to work if $w + x$ is sufficiently large. More generally, in a treatment context y^* is an individual's outcome if treated (the potential outcome), d is a treatment indicator, and x is a variable that only affects the decision to treat but not the outcome if treated. An example with a finite a (two sided censoring) would be ordered treatment, where the latent $w_i + x_i$ determines the treatment, which if negative would indicate a lesser (or no) treatment and if greater than a would indicate a stronger treatment, where the possible treatments are, e.g., dosages of a drug or years of schooling.

Define ω by

$$\omega = \frac{E[y/f(x)]}{E[d/f(x)]} \tag{19}$$

Theorem 1 and Corollary 1 in Lewbel (2002) show that, if a is finite and x has a sufficiently large support, then $\omega = E(y^*)$, and that even without these assumptions, $\omega \approx E(y^*)$.

Both the numerator and denominator of ω are inverse density weighted means, so we propose the simple

[7] See also Lewbel (1997) and McFadden (1999).

estimator

$$\widehat{\omega} = \frac{\sum_{i=1}^{n-1} \left(y_{[i+1]} + y_{[i]}\right)\left(x_{[i+1]} - x_{[i]}\right)}{\sum_{i=1}^{n-1} \left(d_{[i+1]} + d_{[i]}\right)\left(x_{[i+1]} - x_{[i]}\right)} \tag{20}$$

The limiting distribution for $\widehat{\omega}$ is obtained by applying Corollary 5 with $v = (y, d)$ to obtain the joint distribution of the numerator and denominator of $\widehat{\omega}$, then applying the delta method. The result is $n^{1/2}\left(\widehat{\omega} - \omega\right) \xrightarrow{d} N\left(0, (3/2)s^2\right)$, where

$$s^2 = \left[E\left(\frac{d}{f(x)}\right)\right]^{-2} E\left[\frac{\mathrm{Var}\left[(y - \omega d)|x\right]}{f^2(x)}\right] \tag{21}$$

which can be consistently estimated by

$$\widehat{s}^2 = \frac{n\sum_{i=1}^{n-1}\left[\left(y_{[i+1]} - \widehat{\omega} d_{[i+1]}\right) - \left(y_{[i]} - \widehat{\omega} d_{[i]}\right)\right]^2 \left(x_{[i+1]} - x_{[i]}\right)^2}{\left[\sum_{i=1}^{n-1}\left(d_{[i]} + d_{[i+1]}\right)\left(x_{[i+1]} - x_{[i]}\right)\right]^2} \tag{22}$$

4.3 Endogeneous Binary Choice Models

Consider the binary choice or binomial response model

$$d_i = I(x_i + z_i'\beta + e_i \geq 0) \tag{23}$$

Where for each individual i, d_i is an observed zero or one outcome, z_i is a vector of possibly endogeneous observed regressors, x_i is an observed scalar regressor with coefficient normalized to equal one, β is a vector of coefficients to be estimated, and e_i is an unobserved error term. If e is independent of x, z then β can be estimated either parametrically by maximum likelihood if the distribution of e is known or semiparametrically using, e.g., Klein and Spady (1993).

Suppose that the joint distribution of e, z is unknown, but we observe a vector of instrumental variables r_i such that $E(re) = 0$. If x is independent of z, r, e then, given some regularity and support assumptions[8], Lewbel (2000) shows that

$$E(rz')\beta = E\left(r\frac{d - I(x > 0)}{f(x)}\right) \tag{24}$$

so β can be consistently estimated by a linear two stage least squares regression of $[d - I(x > 0)]/f(x)$ on z using instruments r. Note that r and z may contain a constant term, so the regression includes estimation of location in $z'\beta$ assuming e has unconditional mean zero.

[8] In particular, x should have a continuous distribution, be demeaned or otherwise located to contain zero in its support, and contain the support of $-(z'\beta + e)$ in its support. Magnac and Maurin (2003) discuss alternative restrictions.

Corollary 7 below provides a simple method for implementing this result, by applying linear two stage least squares and using Corollary 5 to deal with the density in the right side of equation (24).

Make the following definitions. Let $v_i = r_i[d_i - I(x_i > 0)]$, $\psi = E[v/f(x)]$, $\Sigma_{zr} = E(zr')$, $\Sigma_{rr} = E(rr')$, and $\Delta = (\Sigma_{zr}\Sigma_{rr}^{-1}\Sigma_{zr}')^{-1}\Sigma_{zr}\Sigma_{rr}^{-1}$. Assuming the inverses in the definition of Δ exists, it follows from equation (24) that $\beta = \Delta\psi$. The corresponding finite sample expressions are $\widehat{\Sigma}_{zr} = \sum_{i=1}^{n-1} z_{[i]}r_{[i]}'/n$, $\widehat{\Sigma}_{rr} = \sum_{i=1}^{n-1} r_{[i]}r_{[i]}'/n$, $\hat{\psi} = \sum_{i=1}^{n-1}\left(v_{[i]} + v_{[i+1]}\right)\left(x_{[i+1]} - x_{[i]}\right)/2$, $\widehat{\beta} = \widehat{\Delta}\hat{\psi}$.and

$$\widehat{\Delta} = (\widehat{\Sigma}_{zr}\widehat{\Sigma}_{rr}^{-1}\widehat{\Sigma}_{zr}')^{-1}\widehat{\Sigma}_{zr}\widehat{\Sigma}_{rr}^{-1}. \tag{25}$$

Corollary 7 *Given the above definitions, if Assumptions 1 through 5 hold for* $y_i = v_{i1}, \ldots, v_{ik}$, *then* $n^{1/2}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, \Omega)$ *and* $\widehat{\Omega} \xrightarrow{p} \Omega$ *where* $\Omega = \Delta E\left[(W+U)(W+U)'\right]\Delta'$, $W = (rz' - \Sigma_{zr}')\beta$, $U = y - E[y|x]$, $\widehat{\Omega} = n^{-1}\widehat{\Delta}\left(\sum_{i=1}^{n-1}\frac{3}{8}\hat{U}_{[i]}\hat{U}_{[i]}' + 2\hat{U}_{[i]}\widehat{W}_{[i]}' + \widehat{W}_{[i]}\widehat{W}_{[i]}'\right)\widehat{\Delta}'$, $\widehat{W}_{[i]} = \left(r_{[i]}z_{[i]}' - \widehat{\Sigma}_{zr}'\right)\widehat{\beta}$ *and* $\hat{U}_{[i]} = \left(y_{[i]} - y_{[i+1]}\right)\left(x_{[i+1]} - x_{[i]}\right)n$.

As noted by Lewbel (2000), the binary choice or binomial response model just described is identified under more general settings. In particular, the binary outcome can be generated from

$$d_i = I(\tilde{x}_i + z_i'\beta + e_i \geq 0) \tag{26}$$

where $\tilde{x}$ does not need to be independent from z and r. In this case, a consistent estimation method consists in constructing a new variable x from the residuals of the least-squares projection of $\tilde{x}$ on z and r. Provided that the residual x is independent from z and r, the normal equation 24 still holds and the methodology outlined in the beginning of this section applies. However, the asymptotic variance and this estimator is affected by the fact that the variable x is estimated in a first step. Theorem 6 can be used to handle this situation, by making the following identifications between the quantities used in the statement of the Theorem (on the left-hand side) and the quantities introduced in the present section (on the right-hand side):

$$w_i = (z_i', r_i') \tag{27}$$

$$\hat{\gamma} = \left(\sum_{i=1}^{n} w_i w_i'\right)^{-1}\sum_{i=1}^{n} w_i\tilde{x}_i \tag{28}$$

$$X(w, \gamma) = w_i'\gamma. \tag{29}$$

5 Empirical Application

We now apply the results of the previous section to estimation of a model by Cogneau and Maurin (2002) on the effects of parental income on school attendance in Madagascar. The data set consists of a representative sample of 1401 children aged six to eight, from a World Bank survey conducted in 1993-1994. The model is equation (23) where d_i equals one if child i is enrolled in school on time (that is, by age 6), and zero otherwise, x_i is the date of birth of the child in the relevant year, normalized to vary from -1/2 to 1/2, and the other regressors z_i are a constant, the child's sex, parents' income, and the mother's education level (Cogneau and Maurin report virtually no difference between using mother's or father's education level).

The latent error e_i may be correlated with parent's income and education level, because of family fixed effects (common unobserved determinants of parent's resources and decisions) and measurement errors, in that parent's observed income and education level are rough proxies for permanent income and other measures of total household resources. To control for this endogeneity, instruments r_i are defined as the child's sex, differences between parents and grandparents education levels, and the difference between fathers and grandfathers sector of labor activity (agriculture vs nonagriculture). For more details about the data, the model, and alternative estimators, see Cogneau and Maurin (2002).

Table 2 below reports results using two different estimators, both based on equation (24) and the implication that $\beta = \Delta\psi$. The first estimator is from Lewbel (2000), which uses the same estimator of Δ as equation (25), except that observation $[n]$ is not omitted, and estimates ψ as $n^{-1}\sum_{i=1}^{n} v_i/\widehat{f}(v_i)$, where $\widehat{f}(v_i)$ is a quartic kernel estimator. Bandwidth selection and kernel based standard error estimates for this estimator are constructed using the methods described in Lewbel (2000). The second estimator in Table 2 is the ordered data estimator in Corollary 7.

Both the kernel and ordered data estimators are equivalent to a linear two stage least squares regression of an estimate of $[d - I(x > 0)]/f(x)$ on regressors z using instruments. Each estimator is applied twice in Table 2. The first application of each uses z as instruments as well as regressors, and so is equivalent to an ordinary least squares instead of a two stage least squares, which fails to control for endogeneity. The second application of each estimator uses r as described above as instruments.[9]

[9] The first-step F statistics of the regression of each element of x on z are relatively large (greater than 50), indicating that our results do not suffer from a weak instrument problem (see Equations (3.5) and (3.7) in Staiger and Stock, 1997).

The kernel and ordered data based estimates are generally quite similar. The kernel based two stage least squares estimates are all within two standard errors of the ordered data based two stage least squares estimates, and vice versa. The main empirical finding is that controlling for endogeneity more than quadruples the estimated effect of parental income on the decision to start children's schooling on time. After controlling for income and endogeneity, the effects of mother's education level and the sex of the child are small and statistically insignificant

6 Conclusion

We provide the limiting root n distribution for a simple "ordered data" estimator of means of functions that are scaled by an unknown density, or equivalently, integrals of conditional expectations. We show that the ordered data estimator is a viable alternative to more complicated estimators that require smoothing parameters such as kernels and bandwidths.

Our asymptotic distribution theory is complicated by the fact that the dependence among sorted data spacings $x_{[i+1]}-x_{[i]}$ is of a form that is not covered by standard central limit theorems for dependent processes. Each spacing depends equally strongly on arbitrarily distant spacings and not only on its neighbors. We substantially extend Weiss (1958) to derive asymptotic distribution theory for these spacings.

Although our "ordered data" approach only covers the case where x is scalar, it should be noted that the majority of empirical applications involving nonparametric density estimation are likely to be univariate, e.g., the popular STATA econometrics package has built in commands for univariate, but not multivariate, nonparametric kernel density and kernel regression estimation.

It is possible to extend the approach presented here to obtain consistent estimators for K-dimensional x by replacing spacings by suitable functions of first-nearest neighbors distances (details are available from the authors upon request). However, achieving root n consistency involves handling some of the technical difficulties that also plague semiparametric multivariate kernel estimation. In conventional kernel estimators, boundary effects introduce a $O(h)$ bias, where h is the bandwidth (Cheng *et al.*, 1997), regardless of the order of the kernel. By analogy, since the effective bandwidth in a nearest-neighbor estimator is $O\left(n^{-1/K}\right)$, the bias is $O\left(n^{-1/K}\right)$, which is sufficiently large to prevent root n consistent estimation with no bias in the

asymptotic distribution[10] for $K > 1$. In the context of kernel estimation, this bias is dealt with using some form of asymptotic trimming or kernel refinement, such as a local polynomial kernel smoother (Stone, 1977 and Cleveland, 1979). An analogous local nearest-neighbor polynomial estimator would consist of regressing y_j on a polynomial in x_j for all x_j in some $O\left(n^{-1/K}\right)$ neighborhood of x_i and using the resulting predicted value $\hat{y}_i$ in a sample average weighted by nearest-neighbor distances to the power K. It would be interesting to see if such extensions can attain root n consistency with a fixed number of neighbors in a multivariate setting.

Appendix A

Lemma 8 *Let (i) g_i be an iid sequence drawn from a $e^{-g_i}1\left(g_i \geq 0\right)$ density, (ii) u_i be an iid sequence drawn from a uniform density on $[0,1]$, and (iii) z_i be a sequence such that the following quantities are defined: $\mu = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left[z_i\right]$, $\rho^2 = \lim_{n\to\infty} \mathrm{Var}\left[n^{-1/2}\sum_{i=1}^{n} z_i\right]$, and $\tau^2 = \lim_{n\to\infty} n^{-1}\left(\sum_{i=1}^{n} E\left[z_i^2\right]\right)$. If (i) z_i is independent from g_i and $\left(u_{[i+1]} - u_{[i]}\right)$ and (ii) if $n^{-1/2}\left(\sum_{i=1}^{n} z_i g_i - n\mu\right)$ is asymptotically normal, then*

$$n^{-1/2}\left(\sum_{i=1}^{n-1} z_{[i]} n\left(u_{[i+1]} - u_{[i]}\right) - n\mu\right) \xrightarrow{d} N\left(0, \rho^2 + \tau^2 - \mu^2\right). \tag{30}$$

Proof. This proof is analogous to Weiss' (1958) derivation of the asymptotic distribution of homogenous functions of spacings. Let $G_{[i]} = \sum_{j=0}^{i+1} g_{[j]}$ and $s_{[i]} = g_{[i]}/G_n$ for $i = 1, \ldots, n-1$. It can be shown (see Weiss, 1958) that the joint distribution of the $s_{[i]}$ is identical to the one of the $u_{[i+1]} - u_{[i]}$. The desired result can thus be established by relating the distribution of $n^{-1}\sum_{i=1}^{n-1} z_{[i]} g_{[i]}$ to the one of $n^{-1}\sum_{i=1}^{n-1} z_{[i]} s_{[i]}$.

Let us first calculate the mean and the variance of $n^{-1}\sum_{i=1}^{n-1} z_{[i]} g_{[i]}$. We have $\lim_{n\to\infty} E\left[n^{-1}\sum_{i=1}^{n-1} z_{[i]} g_{[i]}\right]$ $= \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n-1} E\left[z_{[i]} g_{[i]}\right] = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left[z_i g_i\right] = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left[z_i\right] E\left[g_i\right] = \lim_{n\to\infty} n^{-1}$ $\sum_{i=1}^{n} E\left[z_i\right]\cdot 1 = \mu$. Also, $\lim_{n\to\infty} \mathrm{Var}\left[n^{-1/2}\sum_{i=1}^{n-1} z_{[i]} g_{[i]}\right] = \lim_{n\to\infty} n^{-1} E\left[\left(\sum_{i=1}^{n} z_i g_i\right)^2\right] - \mu^2 = \lim_{n\to\infty} n^{-1}$ $\sum_{i=1}^{n} E\left[z_i^2\right] E\left[g_i^2\right] + \sum_{i=1}^{n}\sum_{j\neq i} E\left[z_i z_j\right] E\left[g_i\right] E\left[g_j\right] - \mu^2 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left[z_i^2\right]\cdot 2 + \sum_{i=1}^{n}\sum_{j\neq i} E\left[z_i z_j\right]\cdot$ $1\cdot 1 - \mu^2 = \lim_{n\to\infty} n^{-1}\left(\sum_{i=1}^{n} E\left[z_i^2\right]\right) + n^{-1}\left(\sum_{i=1}^{n} E\left[z_i^2\right] + \sum_{i=1}^{n}\sum_{j\neq i} E\left[z_i z_j\right]\right) - \mu^2 = \tau^2 + \lim_{n\to\infty}$

[10] The same issue of $O\left(n^{-1/K}\right)$ bias in an otherwise root n consistent estimator was noted in a nearest neighbor matching estimator investigated by Abadie and Imbens (2002), although their estimator does not make use of nearest neighbor distances to evaluate an inverse density.

$E\left[\left(n^{-1/2}\sum_{i=1}^{n} z_i\right)^2\right] - \mu^2 = \tau^2 + \rho^2$. Thus,

$$X_0 = \frac{\sum_{i=1}^{n-1} z_{[i]} g_{[i]} - n\mu}{n^{1/2}\gamma} \tag{31}$$

where $\gamma = \sqrt{\tau^2 + \rho^2}$, has a $N(0,1)$ asymptotic distribution. However, X_0 can also be written (see Weiss, 1958) as

$$\begin{aligned}
X_0 &= \frac{\sum_{i=1}^{n-1} z_{[i]} g_{[i]} - G_n\mu + G_n\mu - n\mu}{n^{1/2}\gamma} \\
&= \frac{\sum_{i=1}^{n-1} z_{[i]} g_{[i]} - G_n\mu}{\left(\frac{G_n}{n}\right) n^{1/2}\gamma} + \frac{G_n\mu - n\mu}{\left(\frac{G_n}{n}\right) n^{1/2}\gamma} + O_p\left(n^{-1}\right) \\
&= \frac{\sum_{i=1}^{n-1} z_{[i]} n\left(g_{[i]}/G_n\right) - n\mu}{n^{1/2}\gamma} + \frac{G_n\mu - n\mu}{\left(\frac{G_n}{n}\right) n^{1/2}\gamma} + O_p\left(n^{-1}\right) \\
&= \frac{\sum_{i=1}^{n-1} z_i n s_{[i]} - n\mu}{n^{1/2}\gamma} + \frac{G_n\mu - n\mu}{\left(\frac{G_n}{n}\right) n^{1/2}\gamma} + O_p\left(n^{-1}\right) \\
&\equiv X_1 + X_2 + O_p\left(n^{-1}\right)
\end{aligned} \tag{32}$$

where X_1 and X_2 are independent (because it can be shown (see Weiss, 1958) that d_i and G_n are independent). Moreover,

$$X_2 = \frac{G_n\mu - n\mu}{\left(\frac{G_n}{n}\right) n^{1/2}\gamma} = \left(\frac{G_n - n}{n^{1/2}}\right)\left(\frac{\mu}{\left(\frac{G_n}{n}\right)\gamma}\right) \xrightarrow{p} \left(\frac{G_n - n}{n^{1/2}}\right)\left(\frac{\mu}{\gamma}\right). \tag{33}$$

It follows that $X_2 \xrightarrow{d} N\left(0, (\mu/\gamma)^2\right)$. Since (i) $X_0 = X_1 + X_2$, (ii) $X_0 \xrightarrow{d} N(0,1)$, (iii) $X_2 \xrightarrow{d} N\left(0, (\mu/\gamma)^2\right)$ and (iv) X_1 and X_2 are independent, it follows that the asymptotic distribution of X_1 is the one of X_0, "deconvoluted" by the one of X_2. For independent normals, the deconvolution operation simply amounts to subtracting the variances. Thus, $X_1 \xrightarrow{d} N\left(0, 1 - (\mu/\gamma)^2\right)$. Since $X_1 = \left(\sum_{i=1}^{n-1} z_{[i]} n s_{[i]} - n\mu\right) / \left(n^{1/2}\gamma\right)$, then $n^{-1/2}\left(\sum_{i=1}^{n-1} z_{[i]} n s_{[i]} - n\mu\right) \xrightarrow{d} N\left(0, \gamma^2 - \mu^2\right)$, where $\gamma^2 - \mu^2 = \rho^2 + \tau^2 - \mu^2$. ■

Lemma 9 *Let g_i be an iid sequence drawn from a $e^{-g_i} 1(g_i \geq 0)$ density and let u_i be an iid sequence drawn from a uniform density on $[0,1]$. If (i) z_i is independent from g_i, and $\left(u_{[i+1]} - u_{[i]}\right)$ and (ii) if $n^{-1}\sum_{i=1}^{n} z_i g_i^{\alpha} \xrightarrow{p} \mu$, then $n^{-1}\sum_{i=1}^{n} z_{[i]} n^{\alpha}\left(u_{[i+1]} - u_{[i]}\right)^{\alpha} \xrightarrow{p} \mu$.*

Proof. Let $G_{[i]} = \sum_{j=0}^{i+1} g_{[j]}$. As noted by Weiss (1958), $n^{-1}\sum_{i=1}^{n-1} z_{[i]} n^{\alpha}\left(u_{[i+1]} - u_{[i]}\right)^{\alpha}$ has the same distribution as $n^{-1}\sum_{i=1}^{n} z_{[i]} n^{\alpha}\left(g_{[i]}/G_{[n]}\right)^{\alpha} = n^{-1}\sum_{i=1}^{n} z_i n^{\alpha}\left(g_i/G_{[n]}\right)^{\alpha} = \left(n/G_{[n]}\right)^{\alpha} n^{-1}\sum_{i=1}^{n-1} z_i g_i^{\alpha} = \left(1 + o_p(1)\right) n^{-1}\sum_{i=1}^{n} z_i g_i^{\alpha}$,where $n^{-1}\sum_{i=1}^{n} z_i g_i^{\alpha} \xrightarrow{p} \mu$, by assumption. ■

Lemma 10 *If x_i is an iid sequence drawn from a continuous density $f(x)$ satisfying assumption 2, then $\sum_{i=1}^{n-1}\left(x_{[i+1]}-x_{[i]}\right)^{\alpha}=O_p\left(n^{1-\alpha}\right)$ for $\alpha>0$.*

Proof. Let $\underline{f}=\inf_x f(x)$ and $u_i=F(x_i)$. By Assumption 2, $\underline{f}>0$ and the inverse function $F^{-1}(\cdot)$ is uniquely defined. We then have, by the mean value theorem and the continuity of $f(x)$,

$$\begin{aligned}
&\sum_{i=1}^{n-1}\left(x_{[i+1]}-x_{[i]}\right)^{\alpha}=\sum_{i=1}^{n-1}\left(F^{-1}\left(u_{[i+1]}\right)-F^{-1}\left(u_{[i]}\right)\right)^{\alpha}\\
=&\sum_{i=1}^{n-1}\left(f\left(\xi_{[i]}\right)\right)^{-\alpha}\left(u_{[i+1]}-u_{[i]}\right)^{\alpha}\ \text{for some } \xi_{[i]}\in\left[u_{[i]},u_{[i+1]}\right]\\
\leq&\ \underline{f}^{-\alpha}\sum_{i=1}^{n-1}\left(u_{[i+1]}-u_{[i]}\right)^{\alpha}=\underline{f}^{-\alpha}n^{1-\alpha}\left(n^{-1}\sum_{i=1}^{n-1}n^{\alpha}\left(u_{[i+1]}-u_{[i]}\right)^{\alpha}\right)\\
=&\ \underline{f}^{-\alpha}n^{1-\alpha}O_p(1)=O_p\left(n^{1-\alpha}\right) \qquad (34)
\end{aligned}$$

where the second to last equality follows from Lemma 9. ■

Lemma 11 *If $\mathrm{Var}[a_i]<\infty$, then $\sup_{i\in\{1,\ldots,n\}}a_i=O_p\left(n^{1/2}\right)$.*

Proof. Combining $P[A\cup B]\leq P[A]+P[B]$ with Tschebytchev's inequality, we have that $P[a_i\geq Cn^{1/2}\text{ for some }i\leq n]\leq\sum_{i=1}^{n}P\left[a_i\geq Cn^{1/2}\right]\leq\sum_{i=1}^{n}\mathrm{Var}[a_i]C^{-2}n^{-1}=\mathrm{Var}[a_i]C^{-2}$, which can be made arbitrarily small for all n by choosing a C sufficiently large. ■

Proof. (of Theorem 1) Let $g(x)=E[y|x]$, $\Delta y_{[i]}=y_{[i]}-g\left(x_{[i]}\right)$, $d_{[i]}=F\left(x_{[i+1]}\right)-F\left(x_{[i]}\right)$ and write $\tilde{\theta}-\theta=N_1+R_1+R_2$ where

$$\begin{aligned}
N_1&=n^{-1}\sum_{i=1}^{n-1}\frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)}nd_{[i]} \qquad (35)\\
R_1&=\sum_{i=1}^{n-1}g\left(x_{[i]}\right)\left(x_{[i+1]}-x_{[i]}\right)-\int g(x)\,dx \qquad (36)\\
R_2&=\sum_{i=1}^{n-1}\Delta y_{[i]}\left(x_{[i+1]}-x_{[i]}\right)-\sum_{i=1}^{n-1}\frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)}d_{[i]}. \qquad (37)
\end{aligned}$$

R_1 can be bounded in probability using the mean value theorem, the Hölder properties of $g(x)$, and Lemma 10: $|R_1|=\left|\sum_{i=1}^{n-1}g\left(x_{[i]}\right)\left(x_{[i+1]}-x_{[i]}\right)-\int g(x)\,dx\right|=\left|\sum_{i=1}^{n-1}\left(g\left(x_{[i]}\right)\left(x_{[i+1]}-x_{[i]}\right)-\int_{x_{[i]}}^{x_{[i+1]}}g(x)\,dx\right)\right|=\left|\sum_{i=1}^{n-1}\left(g\left(x_{[i]}\right)\left(x_{[i+1]}-x_{[i]}\right)-g\left(\xi_{[i]}\right)\left(x_{[i+1]}-x_{[i]}\right)\right)\right|$ for some $\xi_{[i]}\in\left[x_{[i]},x_{[i+1]}\right]$. We thus have $|R_1|\leq\sum_{i=1}^{n-1}\left|g\left(x_{[i]}\right)-g\left(\xi_{[i]}\right)\right|\left(x_{[i+1]}-x_{[i]}\right)\leq\sum_{i=1}^{n-1}H_g\left|\xi_{[i]}-x_{[i]}\right|^{h_g}\left(x_{[i+1]}-x_{[i]}\right)\leq H_g\sum_{i=1}^{n-1}\left(x_{[i+1]}-x_{[i]}\right)^{1+h_g}$ $=O_p\left(n^{-h_g}\right)=o_p\left(n^{-1/2}\right)$.

The second remainder term R_2 can be similarly bounded with the help of the Cauchy-Schwartz inequality:

$$\begin{aligned}
|R_2| &= \left|\sum_{i=1}^{n-1} \Delta y_{[i]} \left(x_{[i+1]} - x_{[i]}\right) - \sum_{i=1}^{n-1} \frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)} d_{[i]}\right| \\
&= \left|\sum_{i=1}^{n-1} \frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)} \left(\left(x_{[i+1]} - x_{[i]}\right) f\left(x_{[i]}\right) - \left(F\left(x_{[i+1]}\right) - F\left(x_{[i]}\right)\right)\right)\right| \\
&= \left|\sum_{i=1}^{n-1} \frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)} \left(x_{[i+1]} - x_{[i]}\right) \left(f\left(x_{[i]}\right) - f\left(\xi_{[i]}\right)\right)\right| \text{ for some } \xi_{[i]} \in \left[x_{[i]}, x_{[i+1]}\right] \\
&\leq \sum_{i=1}^{n-1} \frac{\left|\Delta y_{[i]}\right|}{f\left(x_{[i]}\right)} \left(x_{[i+1]} - x_{[i]}\right) \left|f\left(x_{[i]}\right) - f\left(\xi_{[i]}\right)\right| \leq H_f n^{-1} \sum_{i=1}^{n-1} \frac{\left|\Delta y_{[i]}\right|}{f\left(x_{[i]}\right)} n \left(x_{[i+1]} - x_{[i]}\right)^{1+h_f} \\
&\leq H_f \left(n^{-1} \sum_{i=1}^{n-1} \frac{\left(\Delta y_{[i]}\right)^2}{f^2\left(x_{[i]}\right)}\right)^{1/2} \left(n^{-1} \sum_{i=1}^{n-1} n^2 \left(x_{[i+1]} - x_{[i]}\right)^{2+2h_f}\right)^{1/2} \\
&= H_f \left(n^{-1} \sum_{i=1}^{n-1} \frac{\left(\Delta y_{[i]}\right)^2}{f^2\left(x_{[i]}\right)}\right)^{1/2} \left(n \sum_{i=1}^{n-1} \left(x_{[i+1]} - x_{[i]}\right)^{2+2h_f}\right)^{1/2} \\
&= O_p(1) O_p\left(n^{-h_f}\right) = o_p\left(n^{-1/2}\right). \qquad (38)
\end{aligned}$$

We now show that the N_1 term is asymptotically normal and root n consistent. One cannot simply use the Lindeberg-Levy CLT to determine the asymptotics of this sum, because the $d_{[i]}$ are dependent. However, by Lemma 8, one can still achieve asymptotic normality and root n consistency, if two requirements are met: (i) $\Delta y_{[i]}/f\left(x_{[i]}\right)$ is independent from $d_{[i]}$ and (ii) $n^{-1}\sum_{i=1}^{n} \Delta y_i g_i / f\left(x_i\right)$ is asymptotically normal and root n consistent, where g_i is an iid sequence independent from $\Delta y_i / f\left(x_i\right)$ and drawn from a $e^{-g_i} 1\left(g_i \geq 0\right)$ density.

To show that $\left(\Delta y_{[i]}, x_{[i]}\right)$ is asymptotically independent (a.i.) from $d_{[i]}$, we use the fact, shown by Barbe (1994), that $u_{[i]}$ and $d_{[i]}$ are a.i. First note that $x_{[i]}$ and $d_{[i]}$ are also a.i. since $x_{[i]} = F^{-1}\left(u_{[i]}\right)$. Then, to see that $\Delta y_{[i]}$ and $d_{[i]}$ are a.i., observe that, asymptotically, $P\left[\Delta y_{[i]} | x_{[i]}, d_{[i]}\right] = P\left[\Delta y_{[i]} | x_{[i]}, x_{[i+1]}\right] = P\left[\Delta y_{[i]} | x_{[i]}\right]$.

Now, $n^{-1}\sum_{i=1}^{n-1} \Delta y_{[i]} g_{[i]} / f\left(x_{[i]}\right) \xrightarrow{p} n^{-1}\sum_{i=1}^{n} \Delta y_{[i]} g_{[i]} / f\left(x_{[i]}\right) = n^{-1}\sum_{i=1}^{n} \Delta y_i g_i / f\left(x_i\right)$, which can be shown to be asymptotically normal by the Lindeberg-Levi CLT, since all the variables are iid and since $E\left[\left(\Delta y g / f\left(x\right)\right)^2\right] = E\left[\left(\Delta y / f\left(x\right)\right)^2\right] E\left[g^2\right] = E\left[\mathrm{Var}\left(y|x\right)/f^2\left(x\right)\right] \cdot 2 < \infty$ by assumption.

Since both conditions of Lemma 8 are met, we can conclude that the sum $n^{-1}\sum_{i=1}^{n-1} \left(\Delta y_{[i]} / f\left(x_{[i]}\right)\right) n d_{[i]}$ is asymptotically normal with mean $\mu = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left[\Delta y_i / f\left(x_i\right)\right] = 0$ and variance $\rho^2 + \tau^2 - \mu^2$ where $\rho^2 = \lim_{n\to\infty} \mathrm{Var}\left[n^{-1/2}\sum_{i=1}^{n} \Delta y_i / f\left(x_i\right)\right] = E\left[\left(\Delta y_i / f\left(x_i\right)\right)^2\right] = E\left[\mathrm{Var}\left[y_i|x_i\right]/f^2\left(x_i\right)\right] = \sigma^2$ and $\tau^2 = \lim_{n\to\infty} n^{-1}\left(\sum_{i=1}^{n} E\left[\left(\Delta y_i / f\left(x_i\right)\right)^2\right]\right) = \sigma^2$. ■

Proof. (of Corollary 2) First observe that

$$\hat{\theta} = \sum_{i=1}^{n-1} y_{[i]} \left(x_{[i+1]} - x_{[i]} \right) /2 + \sum_{i=1}^{n-1} y_{[i+1]} \left(x_{[i+1]} - x_{[i]} \right) /2 \equiv T_1 + T_2. \tag{39}$$

Theorem 1 directly implies that the T_1 term is asymptotically normal. After reversing the order of the data, Theorem 1 also implies that the T_2 term is asymptotically normal. We then need to compute the asymptotic variance, which can be done by applying Lemma 8 with $z_{[i]} = \Delta y_{[i]} / \left(2f \left(x_{[i]} \right) \right) + \Delta y_{[i+1]} / \left(2f \left(x_{[i+1]} \right) \right)$, since it allows for dependent sequences such as z_i. We simply have to derive the new values of ρ^2 and τ^2: $\rho^2 = \lim_{n\to\infty} \mathrm{Var} \left[n^{-1/2} \sum_{i=1}^{n-1} \left(\Delta y_{[i]} / \left(2f \left(x_{[i]} \right) \right) + \Delta y_{[i+1]} / \left(2f \left(x_{[i+1]} \right) \right) \right) \right] = \lim_{n\to\infty} n^{-1} \mathrm{Var} [\sum_{i=1}^{n} \Delta y_i / f \left(x_i \right)] = E \left[\mathrm{Var} \left[y|x \right] / f^2 \left(x \right) \right] = \sigma^2$ and $\tau^2 = \lim_{n\to\infty} n^{-1} \left(\sum_{i=1}^{n-1} \mathrm{Var} \left(y_{[i]} / \left(2f \left(x_{[i]} \right) \right) + y_{[i+1]} / \left(2f \left(x_{[i+1]} \right) \right) \right) \right) = \lim_{n\to\infty} (2n)^{-1} \sum_{i=1}^{n} \mathrm{Var} \left(y_{[i]} / f \left(x_{[i]} \right) \right) = \mathrm{Var} \left(\Delta y_i / f \left(x_i \right) \right) /2 = E \left[\mathrm{Var} \left[y|x \right] / f^2 \left(x \right) \right] /2 = \sigma^2/2$, implying that $n^{1/2} \left(\hat{\theta} - \theta \right) \xrightarrow{d} N \left(0, \rho^2 + \tau^2 \right)$, where $\rho^2 + \tau^2 = 3\sigma^2/2$. ■

Proof. (of Theorem 3) Let $g \left(x \right) = E \left[y|x \right]$, $\Delta y_{[i]} = y_{[i]} - g \left(x_{[i]} \right)$, $u_{[i]} = F \left(x_{[i]} \right)$, and $d_{k,[i]} = u_{[i+k]} - u_{[i]}$. Then, $\tilde{\theta}_{k_n} - \theta = \sum_{i=1}^{n-k_n} y_{[i]} \left(x_{[i+k_n]} - x_{[i]} \right) /k_n - \int g \left(x \right) dx = N_1 - R_0 + R_1 + R_2 + R_3$, where

$$N_1 = n^{-1} \sum_{i=1}^{n} \frac{\Delta y_{[i]}}{f \left(x_{[i]} \right)} \tag{40}$$

$$R_0 = n^{-1} \sum_{i=n-k_n+1}^{n} \frac{\Delta y_{[i]}}{f \left(x_{[i]} \right)} \tag{41}$$

$$R_1 = \sum_{i=1}^{n-k_n} g \left(x_{[i]} \right) \left(x_{[i+k_n]} - x_{[i]} \right) /k_n - \int g \left(x \right) dx \tag{42}$$

$$R_2 = \sum_{i=1}^{n-k_n} \Delta y_{[i]} \left(x_{[i+k_n]} - x_{[i]} \right) /k_n - \sum_{i=1}^{n-k_n} \frac{\Delta y_{[i]}}{f \left(x_{[i]} \right)} d_{k_n,[i]} /k_n \tag{43}$$

$$R_3 = n^{-1} \sum_{i=1}^{n-k_n} \frac{\Delta y_{[i]}}{f \left(x_{[i]} \right)} \left(n d_{k_n,[i]} /k_n - 1 \right). \tag{44}$$

First, we have $N_1 = n^{-1} \sum_{i=1}^{n} \Delta y_{[i]} / f \left(x_{[i]} \right) = n^{-1} \sum_{i=1}^{n} \Delta y_i / f \left(x_i \right)$, which is asymptotically normal by the i.i.d. assumption and the assumption that $E \left[\mathrm{Var} \left[\Delta y_i | x_i \right] / f^2 \left(x_i \right) \right] < \infty$. Next, we bound the remainder terms. $R_0 \equiv r_0 \left(x_{[n-k_n+1]} \right)$ where

$$r_0 \left(\xi \right) = n^{-1} \sum_{i=1}^{n} \frac{\Delta y_{[i]}}{f \left(x_{[i]} \right)} 1 \left(x_{[i]} \geq \xi \right) = n^{-1} \sum_{i=1}^{n} \frac{\Delta y_i}{f \left(x_i \right)} 1 \left(x_i \geq \xi \right) \tag{45}$$

By Tschebytchev's inequality, for any $\varepsilon > 0$, $P \left[\left| n^{1/2} r_0 \left(\xi \right) \right| \geq \varepsilon | \xi \right] \leq \varepsilon^{-2} E \left[1 \left(x_i \geq \xi \right) \left(\Delta y_i / f \left(x_i \right) \right)^2 \right]$. As $\mathrm{Var} \left(y_i | x_i \right) / f^2 \left(x_i \right)$ is positive and $E \left[\mathrm{Var} \left(y_i | x_i \right) / f^2 \left(x_i \right) \right]$ is finite, $E \left[1 \left(x_i \geq \xi \right) \left(\Delta y_i / f \left(x_i \right) \right)^2 \right] \to 0$ as $\xi \to$

$\bar{x} \equiv \sup_{x \in \mathcal{F}} x$. Let ξ_n be a deterministic sequence such that $\xi_n \to \bar{x}$ and such that $x_{[n-k_n+1]} \geq \xi_n$ with probability approaching one, which is possible since $\operatorname{plim} x_{[n-k_n+1]} = \operatorname{plim} x_{[n]} = \bar{x}$. Then, w.p.a. 1, $P\left[\left|n^{1/2} r_0\left(x_{[n-k_n+1]}\right)\right| \geq \varepsilon\right] \leq P\left[\left|n^{1/2} r_0\left(\xi_n\right)\right| \geq \varepsilon\right] \to 0$, implying that $P\left[\left|n^{1/2} r_0\left(x_{[n-k_n+1]}\right)\right| \geq \varepsilon\right] \to 0$ and that $r_0\left(x_{[n-k_n+1]}\right) = o_p\left(n^{-1/2}\right)$.

Some of the changes of variables in the summations below introduce "boundary" terms, denoted by B, which can be shown to be $o_p\left(n^{-1/2}\right)$ (The proof is available upon request). Next, we consider, $|R_1| = \left|\sum_{i=1}^{n-k_n} g\left(x_{[i]}\right)\left(x_{[i+k_n]} - x_{[i]}\right)/k_n - \sum_{i=1}^{n-1} \int_{x_{[i]}}^{x_{[i+1]}} g\left(x\right) dx\right| = |\sum_{i=1}^{n-k_n} g\left(x_{[i]}\right)\left(x_{[i+k_n]} - x_{[i]}\right)/k_n - k_n^{-1} \sum_{i=1}^{n-k_n} \int_{x_{[i]}}^{x_{[i+k_n]}} g\left(x\right) dx + B|$ for $\xi_{[i]} \in \left[x_{[i]}, x_{[i+k_n]}\right]$.

Next, $|R_1| \leq H_g |k_n^{-1} \sum_{i=1}^{n-k_n} \left(x_{[i+k_n]} - x_{[i]}\right)^{1+h_g}| + |B| \leq H_g \left|\sum_{i=1}^{n-k_n} \left(k_n^{-1} \sum_{j=1}^{k_n} \left(x_{[i+j]} - x_{[i+j-1]}\right)\right)^{1+h_g}\right| + |B| \leq H_g \left|\sum_{i=1}^{n-k_n} k_n^{-1} \sum_{j=1}^{k_n} \left(x_{[i+j]} - x_{[i+j-1]}\right)^{1+h_g}\right| + |B| \leq H_g \left|\sum_{i=1}^{n-k_n} \left(x_{[i+1]} - x_{[i]}\right)^{1+h_g}\right| + |B| = O_p\left(n^{-h_g}\right) = o_p\left(n^{-1/2}\right)$

The R_2 remainder term can be similarly shown to be $o_p\left(n^{-1/2}\right)$. Finally, $R_3 = n^{-1} \sum_{i=1}^{n-k_n} f^{-1}\left(x_{[i]}\right) \Delta y_{[i]} \left(n d_{k,[i]}/k_n - 1\right) = n^{-1} \sum_{i=1}^{n-k_n} f^{-1}\left(x_{[i]}\right) \Delta y_{[i]} \left(\left(k_n^{-1} \sum_{j=1}^{k_n} n d_{1,[i+j-1]}\right) - 1\right)$. Proceeding as in Lemma 9 and introducing $g_{[i]}$ and $G_{[n]}$ defined therein, R_3 has the same distribution as

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n-k_n} \frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)} \left(\left(\frac{n}{G_{[n]}}\right)\left(k_n^{-1} \sum_{j=1}^{k_n} g_{[i+j-1]}\right) - 1\right) \\ = & \left(\frac{n}{G_{[n]}}\right) n^{-1} \sum_{i=1}^{n-k_n} \frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)} \left(\left(k_n^{-1} \sum_{j=1}^{k_n} g_{[i+j-1]}\right) - 1\right) + \left(1 - \left(\frac{n}{G_{[n]}}\right)\right) n^{-1} \sum_{i=1}^{n-k_n} \frac{\Delta y_{[i]}}{f\left(x_{[i]}\right)} \\ = & \left(1 + o_p\left(1\right)\right) R_{31} + o_p\left(1\right) O_p\left(n^{-1/2}\right) \end{aligned} \tag{46}$$

where $R_{31} = n^{-1} \sum_{i=1}^{n-k_n} f^{-1}\left(x_{[i]}\right) \Delta y_{[i]} \left(k_n^{-1} \sum_{j=1}^{k_n} \left(g_{[i+j-1]} - 1\right)\right)$. Employing well-known techniques used for the study of U-statistics, it can be shown that $E\left[R_{31}^2\right] = O\left(k_n^{-1} n^{-1}\right)$, implying that $R_3 = O_p\left(k_n^{-1/2} n^{-1/2}\right) = o_p\left(n^{-1/2}\right)$. ■

Proof. (of Theorem 4) We first observe that $\hat{\sigma}^2 = \frac{n}{4} \sum_{i=1}^{n-1} \left(\Delta y_{[i+1]} - \Delta y_{[i]} + r_{[i]}\right)^2 \left(x_{[i+1]} - x_{[i]}\right)^2$ where $\left|r_{[i]}\right| = \left|\left(y_{[i+1]} - y_{[i]}\right) - \left(\Delta y_{[i+1]} - \Delta y_{[i]}\right)\right| = \left|\left(g\left(x_{[i+1]}\right) - g\left(x_{[i]}\right)\right)\right| \leq H_g \left|x_{[i+1]} - x_{[i]}\right|^{h_g}$, by the Hölder property of $g\left(x\right)$. The remainder of the proof (available upon request) is a straightforward but tedious extension of the techniques used in Theorem 1. ■

Proof. (of Corollary 5) This result can be shown along the same lines as Theorems 1, 4 and Corollary 2, with the Cramer-Wold device, letting $y_i = \sum_{j=1}^{k} \alpha_j v_{ij}$, where $(\alpha_1, \ldots, \alpha_k)$ is a vector of arbitrary constants, and noting that $E\left[Covar\left[v_{ij}, v_{ij'}|x_i\right]/f^2\left(x_i\right)\right] \leq \left(E\left[\text{Var}\left[v_{ij}|x_i\right]/f\left(x_i\right)\right] E\left[\text{Var}\left[v_{ij'}|x_i\right]/f\left(x_i\right)\right]\right)^{1/2}$ by Cauchy-Schwartz, thus implying that all elements of the matrix V are bounded. ■

Lemma 12 *Let x be a random scalar and w be a random vector taking value in $\mathbb{R}^{N_w}$ related through $x = X\left(w, \gamma\right)$ for some function $X : \mathbb{R}^{N_w} \times \mathbb{R}^{N_\gamma} \mapsto \mathbb{R}$ and for some parameter vector γ. If $\partial X\left(w, \gamma\right)/\partial\gamma$ exists and is such that $E\left[\left|\partial X\left(w, \gamma\right)/\partial\gamma\right|\right] < \infty$ then*

$$\frac{\partial f_{x|\gamma}\left(x|\gamma\right)}{\partial\gamma} = -\frac{\partial}{\partial x}\left(f_{x|\gamma}\left(x|\gamma\right) E\left[\frac{\partial X\left(w, \gamma\right)}{\partial\gamma}|x\right]\right). \tag{47}$$

Proof. Let us introduce the sequence

$$S_m\left(\xi\right) = \begin{cases} 1 \text{ if } \xi < -m^{-1} \\ \left(1 - m\xi\right)/2 \text{ if } \xi \in \left[-m^{-1}, m^{-1}\right] \\ 0 \text{ if } \xi > m^{-1} \end{cases}, \tag{48}$$

which converges pointwise to the indicator function $1\left(\xi \leq 0\right)$, except at $\xi = 0$. By the Dominated Convergence Theorem, we have $F_{x|\gamma}\left(x|\gamma\right) = \int 1\left[X\left(w, \gamma\right) \leq x\right] dF_w\left(w\right) = \int \lim_{m\to\infty} S_m\left(X\left(w, \gamma\right) - x\right) dF_w\left(w\right) = \lim_{m\to\infty} \int S_m\left(X\left(w, \gamma\right) - x\right) dF_w\left(w\right)$ since the integrand is dominated, for all m, by the absolutely integrable measure $dF_w\left(w\right)$. Differentiating yields $\partial F_{x|\gamma}\left(x|\gamma\right)/\partial\gamma = \lim_{m\to\infty} \int S'_m\left(X\left(w, \gamma\right) - x\right) \partial X\left(w, \gamma\right)/\partial\gamma dF_w\left(w\right)$, where $S'_m\left(\xi\right) = -2m1\left(\left|x\right| \leq m^{-1}\right)$ and where the operator $\partial/\partial\gamma$ commutes with the limit and integral because $S'_m\left(\xi\right)$ is absolutely integrable by construction and so is $\left(\partial X\left(w, \gamma\right)/\partial\gamma\right) dF_w\left(w\right)$, by assumption. Noting that $S'_m\left(\xi\right)$ forms a sequence of functions converging to minus the Dirac delta distribution $-\delta\left(\xi\right)$, we have

$$\begin{aligned} \frac{\partial}{\partial\gamma} F_{x|\gamma}\left(x|\gamma\right) &= -\int \delta\left(X\left(w, \gamma\right) - x\right) \frac{\partial X\left(w, \gamma\right)}{\partial\gamma} dF_w\left(w\right) \\ &= -f_{x|\gamma}\left(x|\gamma\right) \frac{\int \delta\left(X\left(w, \gamma\right) - x\right) \frac{\partial X(w,\gamma)}{\partial\gamma} dF_w\left(w\right)}{f_{x|\gamma}\left(x|\gamma\right)} \\ &= -f_{x|\gamma}\left(x|\gamma\right) E\left[\frac{\partial X\left(w, \gamma\right)}{\partial\gamma}|x\right], \end{aligned} \tag{49}$$

where the last equality uses the definition of $E\left[\partial X\left(w, \gamma\right)/\partial\gamma|x\right]$. Finally, differentiating Equation (49) with respect to x yields Equation (47). ■

Proof. (of Theorem 6) We can decompose the estimation error as $n^{1/2}\left(\tilde{\theta}_{k_n}(\hat{\gamma})-\theta(\gamma^*)\right)=N_1+R_1+N_2+R_2$ where

$$N_1 = n^{1/2}\left(\tilde{\theta}_{k_n}(\gamma^*)-\theta(\gamma^*)\right) \tag{50}$$

$$N_2 = n^{1/2}\frac{\partial\theta(\gamma^*)}{\partial\gamma'}(\hat{\gamma}-\gamma^*) \tag{51}$$

$$R_1 = n^{1/2}\left(\tilde{\theta}_{k_n}(\hat{\gamma})-\theta(\hat{\gamma})\right)-n^{1/2}\left(\tilde{\theta}_{k_n}(\gamma^*)-\theta(\gamma^*)\right) \tag{52}$$

$$R_2 = n^{1/2}\left(\frac{\partial\theta(\hat{\gamma})}{\partial\gamma'}-\frac{\partial\theta(\gamma^*)}{\partial\gamma'}\right)(\hat{\gamma}-\gamma^*) \tag{53}$$

where $\dot{\gamma}$ is a mean value located along the segment joining γ^* and $\hat{\gamma}$. We will evaluate, in turn, (i) R_1, (ii) R_2, (iii) N_1 (iv) N_2 and (v) the asymptotic variance of N_1+N_2.

(i) Theorem 3 shows that $n^{1/2}\left(\tilde{\theta}_{k_n}(\gamma)-\theta(\gamma)\right)$ for a given γ as asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}(y_i-E[y_i|x_i])/f_{x|\gamma}(x_i|\gamma)$. In order to handle the fact that all the quantities are function of γ, Theorem 3 needs to be adapted for the order of the remainder terms to hold jointly for γ^* and $\hat{\gamma}$. This is achieved by replacing Assumptions 2 through 5 by corresponding Assumptions 6 through 9 that hold uniformly for $\gamma\in\Gamma$. It follows that R_1 is such that

$$\begin{aligned} R_1 &= n^{-1/2}\sum_{i=1}^{n}\frac{y_i-E[y_i|x_i=X(w_i,\hat{\gamma})]}{f_{x|\gamma}(x_i|\hat{\gamma})}-n^{-1/2}\sum_{i=1}^{n}\frac{y_i-E[y_i|x_i=X(w_i,\gamma^*)]}{f_{x|\gamma}(x_i|\gamma^*)}+o_p(1) \\ &= R_{11}+R_{12}+o_p(1) \end{aligned} \tag{54}$$

where

$$R_{11} = n^{-1/2}\sum_{i=1}^{n}\frac{f_{x|\gamma}(x_i|\gamma^*)-f_{x|\gamma}(x_i|\hat{\gamma})}{f_{x|\gamma}(x_i|\hat{\gamma})f_{x|\gamma}(x_i|\gamma^*)}\Delta y_i \tag{55}$$

$$R_{12} = n^{-1/2}\sum_{i=1}^{n}\left(\frac{E[y_i|x_i=X(w_i,\hat{\gamma})]-E[y_i|x_i=X(w_i,\gamma^*)]}{f_{x|\gamma}(x_i|\hat{\gamma})}\right) \tag{56}$$

and where $\Delta y_i=y_i-E[y_i|x_i=X(w_i,\gamma^*)]$. Tschebychev's inequality lets us write, for any $\varepsilon>0$,

$$P[|R_{11}|\geq\varepsilon\mid\hat{\gamma}] \leq \varepsilon^{-2}E\left[\left(n^{-1/2}\sum_{i=1}^{n}\frac{\left(f_{x|\gamma}(x_i|\gamma^*)-f_{x|\gamma}(x_i|\hat{\gamma})\right)}{f_{x|\gamma}(x_i|\hat{\gamma})f_{x|\gamma}(x_i|\gamma^*)}\Delta y_i\right)^2|\hat{\gamma}\right] \tag{57}$$

$$= \varepsilon^{-2}\left(n^{-1}n\right)E\left[\frac{\left(f_{x|\gamma}(x_i|\gamma^*)-f_{x|\gamma}(x_i|\hat{\gamma})\right)^2}{f_{x|\gamma}^2(x_i|\hat{\gamma})f_{x|\gamma}^2(x_i|\gamma^*)}(\Delta y_i)^2|\hat{\gamma}\right] \tag{58}$$

where $E\left[f_{x|\gamma}^{-2}(x_i|\hat{\gamma})f_{x|\gamma}^{-2}(x_i|\gamma^*)\left(f_{x|\gamma}(x_i|\gamma^*)-f_{x|\gamma}(x_i|\hat{\gamma})\right)^2y_i^2|\hat{\gamma}\right]\xrightarrow{p}0$ as[11] $\hat{\gamma}\xrightarrow{p}\gamma^*$, by the continuity of

[11] This expectation converges *in probability* because it is a function of $\hat{\gamma}$, which is random.

$f_{x|\gamma}(x_i|\gamma)$ in γ (from Lemma 12), the fact that $f_{x|\gamma}(x_i|\gamma)$ is bounded away from zero for $\gamma \in \Gamma$ and that

$$E\left[\frac{\left(f_{x|\gamma}(x_i|\gamma^*) - f_{x|\gamma}(x_i|\hat{\gamma})\right)^2}{f^2_{x|\gamma}(x_i|\hat{\gamma}) f^2_{x|\gamma}(x_i|\gamma^*)} (\Delta y_i)^2 |\hat{\gamma}\right] \leq \frac{4 \sup_{\gamma\in\Gamma} \sup_{x\in\mathcal{F}_\gamma} f^2_{x|\gamma}(x|\gamma)}{\inf_{\gamma\in\Gamma} \inf_{x\in\mathcal{F}_\gamma} f^2_{x|\gamma}(x|\gamma)} E\left[\frac{(\Delta y_i)^2}{f^2_{x|\gamma}(x|\gamma^*)}\right] < \infty$$

(since $\sup_{\gamma\in\Gamma} \sup_{x\in\mathcal{F}_\gamma} f^2_{x|\gamma}(x|\gamma)$ must be finite by the Hölder continuity of $f_{x|\gamma}(x|\gamma)$). Equation (58) then implies that $R_{11} \xrightarrow{p} 0$.

Similarly, $P[|R_{12}| \geq \varepsilon \mid \hat{\gamma}] \leq \varepsilon^{-2} E[(E[y_i|x_i = X(w_i, \hat{\gamma})] - E[y_i|x_i = X(w_i, \gamma^*)])^2 / f^2_{x|\gamma}(x_i|\hat{\gamma}) |\hat{\gamma}]$. Since $E[y_i|x_i = X(w_i, \gamma)] = \int f^{-1}_{x|\gamma}(X(w,\gamma)|\gamma) E[y|w] \delta(x_i - X(w_i, \gamma)) dw$, continuity of $f_{x|\gamma}(x|\gamma)$ and $X(w,\gamma)$ in γ and $\inf_{\gamma\in\Gamma} \inf_x f_{x|\gamma}(x|\gamma) > 0$ implies that $E[y_i|x_i = X(w_i,\gamma)]$ is continuous in γ. Then, as for R_{11}, we have $R_{12} \xrightarrow{p} 0$.

(ii) The remainder R_2 can then be bounded using the continuity of $\partial\theta(\gamma)/\partial\gamma'$ (from Assumption 11), the assumed root n consistency of $\hat{\gamma}$ and the fact that $\dot{\gamma} \xrightarrow{p} \gamma^*$: $R_2 = (\partial\theta(\gamma^*)/\partial\gamma' - \partial\theta(\dot{\gamma})/\partial\gamma') n^{1/2}(\hat{\gamma} - \gamma) = o_p(1) n^{1/2} O_p\left(n^{-1/2}\right) = o_p(1)$.

(iii) From the proofs of Theorem 3, the N_1 term can be written as $N_1 = n^{-1/2} \sum_{i=1}^n f^{-1}_{x|\gamma}(x|\gamma^*) \Delta y_i + o_p(1)$, where the sum is asymptotically normal.

(iv) By the assumption of the existence of an influence representation for $\hat{\gamma}$, the N_2 term can be written as $N_2 = B' n^{-1/2} \sum_{i=1}^n \Psi_i + o_p(1)$, where $B = \partial\theta(\gamma^*)/\partial\gamma$ and where the sum is asymptotically normal.

(v) The estimation error can then be written as $n^{1/2}\left(\tilde{\theta}_{k_n}(\hat{\gamma}) - \theta(\gamma^*)\right) = n^{-1/2} \sum_{i=1}^n (f^{-1}_{x|\gamma}(x|\gamma^*) \Delta y_i + B'\Psi_i) + o_p(1)$ and straightforward calculations provide the expression of the asymptotic variance V of this sum. An explicit expression for B can be given. First note that

$$\theta(\gamma) = E\left[\frac{y}{f_{x|\gamma}(X(w,\gamma)|\gamma)}\right] = \int \frac{E[y|w]}{f_{x|\gamma}(X(w,\gamma)|\gamma)} f_w(w) dw. \tag{59}$$

This shows that the only dependence of this expression on γ comes from the denominator $f_{x|\gamma}(X(w,\gamma)|\gamma)$. We can then write (the required expectations and derivatives commute by Assumption 11):

$$\frac{\partial\theta(\gamma)}{\partial\gamma} = -E\left[\frac{y}{\left(f_{x|\gamma}(X(w,\gamma)|\gamma)\right)^2} \frac{\partial f_{x|\gamma}(X(w,\gamma)|\gamma)}{\partial\gamma}\right]. \tag{60}$$

The quantity $f_{x|\gamma}(X(w,\gamma)|\gamma)$ depends on γ through two different paths: (i) the function $f_{x|\gamma}(\cdot|\gamma)$ depends on γ and (ii) the point of evaluation $X(w,\gamma)$ depends on γ. We then have

$$\frac{\partial\theta(\gamma)}{\partial\gamma} = -E\left[\frac{y}{\left(f_{x|\gamma}(X(w,\gamma)|\gamma)\right)^2} \left(\left[\frac{\partial f_{x|\gamma}(x|\gamma)}{\partial\gamma}\right]_{x=X(w,\gamma)} + \left[\frac{\partial f_{x|\gamma}(x|\gamma)}{\partial x}\right]_{x=X(w,\gamma)} \frac{\partial X(w,\gamma)}{\partial\gamma}\right)\right]. \tag{61}$$

Using Lemma 12 to evaluate $\partial f_{x|\gamma}(x|\gamma)/\partial\gamma$, we obtain

$$\begin{aligned}\frac{\partial\theta(\gamma)}{\partial\gamma} &= E\left[\frac{y}{f_{x|\gamma}(x|\gamma)}\frac{\partial}{\partial x}E\left[\frac{\partial X(w,\gamma)}{\partial\gamma}|x\right]\right]+ \\ &\quad -E\left[\frac{y}{f_{x|\gamma}(x|\gamma)}\frac{\partial\left(\ln\left(f_{x|\gamma}(x|\gamma)\right)\right)}{\partial x}\left(\frac{\partial X(w,\gamma)}{\partial\gamma}-E\left[\frac{\partial X(w,\gamma)}{\partial\gamma}|x\right]\right)\right]\end{aligned} \tag{62}$$

In the special case mentioned in the second part of the theorem, it can be verified that $E[\partial X(w,\gamma^*)/\partial\gamma|x,y] = E[\partial X(w,\gamma^*)/\partial\gamma]$ implies that $\partial\theta(\gamma)/\partial\gamma|_{\gamma=\gamma^*} = 0$. ■

Proof. (of Corollary 7) Since the estimator $\tilde{\theta} = \sum_{i=1}^{n-1} v_{[i]}\left(x_{[i+1]} - x_{[i]}\right)$, cannot be written as a differentiable functional of the joint cdf of y and x, an "influence function" cannot be defined. However, this estimator can be written as a differentiable functional of the joint density of y, x and an auxiliary random variable g_i, which is an iid sequence independent from y_i and x_i and drawn from a $e^{-g_i}1(g_i \geq 0)$ distribution. We can then consider a "pseudo influence function" $\tilde{\Psi}(y_i,x_i) = (y_i - E[y_i|x_i])\,g_i/f(x_i)$. The "pseudo influence function" for the more efficient estimator $\hat{\theta} = \sum_{i=1}^{n-1}\left(v_{[i]}+v_{[i+1]}\right)\left(x_{[i+1]}-x_{[i]}\right)/2$ is $\hat{\Psi}(y_i,x_i) = (y_i - E[y_i|x_i])(g_i+h_i)/(2f(x_i))$ where h_i is independent from g_i and has the same distribution.

For the binary choice estimator $\widehat{\beta}$, the appropriate pseudo influence function is

$$n^{1/2}\left(\widehat{\beta}-\beta\right) = n^{-1/2}\Delta\sum_{i=1}^{n}\left(\frac{(y_i - E[y_i|x_i])}{f(x_i)}\frac{(g_i+h_i)}{2} - \left(r_i z_i' - \overline{rz'}\right)\beta\right) + o_p(1) \tag{63}$$

(with g_i and h_i defined above) and so the asymptotic variance is $\Delta E\left[(U_i - W_i)(U_i-W_i)'\right]\Delta'$ where

$$U_i = \frac{(y_i - E[y_i|x_i])}{f(x_i)}\frac{(g_i+h_i)}{2} \tag{64}$$

$$W_i = \left(r_i z_i' - \overline{rz'}\right)\beta. \tag{65}$$

Note that $E[W_i W_i']$ is just a standard variance expression. However,

$$\begin{aligned}E[U_i U_i'] &= E\left[\left((y_i - E[y_i|x_i])(y_i-E[y_i|x_i])'/f^2(x_i)\right)(g_i+h_i)^2/4\right] \\ &= E\left[\left((y_i - E[y_i|x_i])(y_i-E[y_i|x_i])'/f^2(x_i)\right)\right]E\left[(g_i+h_i)^2/4\right] \\ &= E\left[\left((y_i - E[y_i|x_i])(y_i-E[y_i|x_i])'/f^2(x_i)\right)\right]\left(E\left[g_i^2\right]+2E[g_i]E[h_i]+\left[h_i^2\right]\right)/4 \\ &= E\left[\left((y_i - E[y_i|x_i])(y_i-E[y_i|x_i])'/f^2(x_i)\right)\right]\frac{6}{4} \\ &= (3/2)E\left[\mathrm{Var}(y_i|x_i)/f^2(x_i)\right]\end{aligned} \tag{66}$$

where $E\left[\mathrm{Var}\left(y_i|x_i\right)/f^2\left(x_i\right)\right]$ can be estimated by $\frac{n}{4}\sum_{i=1}^{n-1}\left(y_{[i+1]}-y_{[i]}\right)\left(y_{[i+1]}-y_{[i]}\right)'\left(x_{[i+1]}-x_{[i]}\right)^2$. Finally,

$$E\left[U_iW_i'\right]=E\left[\left(y_i-E\left[y_i|x_i\right]\right)W_i'/f\left(x_i\right)\right]E\left[\left(g_i+h_i\right)/2\right]=E\left[\left(y_i-E\left[y_i|x_i\right]\right)W_i'/f\left(x_i\right)\right] \tag{67}$$

and $E\left[U_iW_i'\right]$ can be estimated by $\sum_{i=1}^{n}\left(y_{[i]}-y_{[i+1]}\right)\left(x_{[i+1]}-x_{[i]}\right)nW_{[i]}$ since,

$$\begin{aligned}
& E\left[g_{[i]}W_{[i]}\left(y_{[i]}-y_{[i+1]}\right)/f\left(x_{[i]}\right)\right] \\
= \;& E\left[g_{[i]}W_{[i]}\left(y_{[i]}-E\left[y_{[i]}|x_{[i]}\right]-\left(y_{[i+1]}-E\left[y_{[i]}|x_{[i]}\right]\right)\right)/f\left(x_{[i]}\right)\right] \\
= \;& E\left[g_{[i]}W_{[i]}\left(y_{[i]}-E\left[y_{[i]}|x_{[i]}\right]-\left(y_{[i+1]}-E\left[y_{[i+1]}|x_{[i+1]}\right]\right)\right)/f\left(x_{[i]}\right)\right]+o_p\left(n^{-1/2}\right) \\
= \;& E\left[g_{[i]}W_{[i]}\left(y_{[i]}-E\left[y_{[i]}|x_{[i]}\right]\right)/f\left(x_{[i]}\right)\right]+ \\
& -E\left[E\left[\left(y_{[i+1]}-E\left[y_{[i+1]}|x_{[i+1]}\right]\right)|x_{[i+1]}\right]g_{[i]}W_{[i]}/f\left(x_{[i]}\right)\right]+o_p\left(n^{-1/2}\right) \\
= \;& E\left[g_{[i]}W_{[i]}\left(y_{[i]}-E\left[y_{[i]}|x_{[i]}\right]\right)/f\left(x_{[i]}\right)\right]+0+o_p\left(n^{-1/2}\right) \\
= \;& E\left[g_{[i]}\right]E\left[W_{[i]}\left(y_{[i]}-E\left[y_{[i]}|x_{[i]}\right]\right)/f\left(x_{[i]}\right)\right]+o_p\left(n^{-1/2}\right) \\
= \;& E\left[W_{[i]}\left(y_{[i]}-E\left[y_{[i]}|x_{[i]}\right]\right)/f\left(x_{[i]}\right)\right]+o_p\left(n^{-1/2}\right).
\end{aligned} \tag{68}$$

The expression for $\widehat{\Omega}$ given in the theorem thus follows. ■

Acknowledgements

This research was supported in part by the National Science Foundation through grants SES-9905010 (Lewbel) and SES-0214068 (Schennach). The authors would like to thank Peter Robinson and two anonymous referees for helpful comments, Eric Maurin for providing data, Zhihong Chen for research assitance, and Daniel McFadden for suggesting this class of estimators.

References

Abadie, A. and G. Imbens, 2002, Simple and Bias-Corrected Matching Estimators for Average Treatment Effects, Econometrica, forthcoming.

Barbe, P., 1994, Joint approximation of processes based on spacings and order statistics, Stochastic Processes and their Applications 53, 339–349.

Bloch, D. A. and J. L. Gastwirth, 1968, On a Simple Estimate of the Reciprocal of the Density Function, Annals of Mathematical Statistics 39, 1083–1085.

Cheng, M.-Y., J. Fan, and J. S. Marron, 1997, On automatic boundary corrections, Annals of Statistics 25, 1691–1708.

Cogneau, D. and E. Maurin, 2002, Parental Income and School Attendence in a Low-Income Country: a semi-parametric analysis, Unpublished Manuscript.

Hall, P. and J. L. Horowitz, 1990, Bandwidth Selection in Semiparametric Estimation of Censored Linear Regression Models, Econometric Theory 6, 123–150.

Hall, P. and A. Yatchew, 2005, Unified approach to testing functional hypotheses in semiparametric contexts, Journal of Econometrics, forthcoming.

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.

Härdle, W., J. Hart, J. S. Marron and A. B. Tsybakov, 1992, Bandwidth Choice for Average Derivative Estimation, Journal of the American Statistical Association, 87, 218–226.

Hausman, J. A. and W. Newey, 1995, Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss, Econometrica 63, 1445–1476.

Hirano, K., G. Imbens, and G. Ridder, 2003, Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, Econometrica 71, 1161–1189.

Hong, Y. and H. White, 2005, Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence, Econometrica 73, 837–901.

Klein, R. and R. H. Spady, 1993, An efficient Semiparametric Estimator for Binary Response Models, Econometrica 61, 387–421.

Lewbel, A., 1997, Semiparametric Estimation of Location and Other Discrete Choice Moments, Econometric Theory 13, 32–51.

Lewbel, A., 1998, Semiparametric Latent Variable Model Estimation With Endogenous or Mismeasured Regressors, Econometrica 66, 105–121.

Lewbel, A., 2000, Semiparametric Qualitative Response Model Estimation With Unknown Heteroskedasticity or Instrumental Variables, Journal of Econometrics 97, 145–177.

Lewbel, A., 2002, Endogeneous Selection or Treatment Model Estimation, Unpublished Manuscript.

Lewbel, A., O. Linton, and D. L. McFadden, 2001, Estimating Features of a Distribution From Binomial Data, unpublished manuscript.

Magnac, T. and E. Maurin, 2003, Identification and Information in Monotone Binary Models, unpublished manuscript.

McFadden, D. L., 1999, Computing Willingness-to-Pay in Random Utility Models, in: J. Moore, R. Riezman, and J. Melvin, eds., Trade Theory, and Econometrics: Essays in Honour of John S. Chipman (Routledge, New York) 253–274.

Mack, Y. P. and H. G. Müller, 1988, Convolution Type Estimators for Nonparametric Regression, Statistics and Probability Letters 7, 229–239.

Newey, W. K., 1994, The Asymptotic Variance of Semiparametric Estimators, Econometrica 62, 1349–1382.

Newey, W. K., 1997, Convergence rates and asymptotic normality for series estimators, Journal of econometrics 79, 147–168.

Newey, W. K. and P. A. Ruud, 1994, Density Weighted Linear Least Squares, University of California at Berkeley working paper.

Priestley, M. B. and M. T. Chao, 1972, Nonparametric Function Fitting, Journal of the Royal Statistical Society, Series B 34, 385–392.

Robinson, P. M., 1987, Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form, Econometrica 55, 875–891.

Staiger D., and J. H. Stock, 1997, Instrumental Variables Regression with Weak Instruments, Econometrica 65, 557–586.

Wand, M. P. and M. C. Jones, 1995, Kernel Smoothing (CRC Press, Boston).

Weiss, L., 1958, Limiting distributions of homogenous functions of sample spacings, Annals of Mathematical Statistics 29, 310–312.

Yatchew, A., 1997, An Elementary Estimator of the Partial Linear Model, Economics Letters 57, 135–43.

TABLE 1: MONTE CARLO RESULTS

	mean	stddev	lower	median	upper	rmse	mae	mdae
true density $\widehat{\theta}_0$	.9959	.2834	.7967	.9820	1.177	.2835	.2264	.1919
ordered data 1 $\widehat{\theta}_1$	1.000	.2663	.8209	1.001	1.175	.2663	.2112	.1770
ordered data 2 $\widehat{\theta}_2$	1.003	.2470	.8374	1.001	1.164	.2470	.1954	.1634
ordered data 3 $\widehat{\theta}_3$	1.006	.2429	.8441	1.002	1.166	.2430	.1921	.1608
Silverman b kernel $\widehat{\theta}_4$	.9925	.2338	.8334	.9891	1.151	.2339	.1866	.1588
$b/2$ kernel $\widehat{\theta}_5$	.9679	.2262	.8170	.9647	1.121	.2285	.1820	.1528
$2b$ kernel $\widehat{\theta}_6$	1.029	.2546	.8526	1.025	1.199	.2562	.2048	.1744

TABLE 2: SCHOOL ATTENDANCE ESTIMATES

	Kernel OLS	Ordered OLS	Kernel 2SLS	Ordered 2SLS
Constant	−2.758	−2.251	−12.394	−8.394
	(.474)	(.757)	(2.195)	(3.324)
Boy	.030	−.007	.033	−.005
	(.041)	(.067)	(.047)	(.070)
Log Income	.173	.144	.882	.597
	(.035)	(.056)	(.161)	(.244)
Mother's Education	.194	.036	−.089	−.068
	(.026)	(.039)	(.061)	(.094)

Note: Standard errors are in parentheses.