Econometrica Supplementary Material

SUPPLEMENT TO “INSTRUMENTAL VARIABLE ESTIMATION OF NONLINEAR ERRORS-IN-VARIABLES MODELS”: REVIEW, PROOFS, EXTENSIONS, AND EXAMPLE OF APPLICATION

(*Econometrica*, Vol. 75, No. 1, January, 2007, 201–239)

BY SUSANNE M. SCHENNACH[1]

This supplementary material contains some of the more technical details omitted from the main paper. First, a brief review of the theory of generalized functions is presented. Second, proofs of some basic properties of Fourier transforms as well as the asymptotics of the proposed generalized method of moments estimator are given. Third, the proposed estimator is compared with that suggested by Hausman, Newey, Ichimura, Powell (1991). Fourth, an alternative derivation of the moment conditions that necessitates weaker regularity conditions is provided. Fifth, the details of the Monte Carlo simulations are described. Sixth, an application of the proposed methodology to the estimation of the black–white income gap is presented. Finally, some computational aspects of the implementation of the estimator are described.

KEYWORDS: Errors-in-variables model, Fourier transform, generalized function, semiparametric model.

S.1. REVIEW OF THE THEORY OF GENERALIZED FUNCTIONS

The concept of generalized functions, also called tempered distributions (Lighthill (1962), Gel’fand and Shilov (1964), Schwartz (1966)), is central to the present paper, because most results rely on Fourier transforms, which often do not exist within the set of ordinary functions. Because generalized functions are not widely used in the econometrics literature (Phillips (1991) and Zinde-Walsh and Phillips (2003) are notable exceptions), this section recalls the definitions and known results that are relevant to our problem. Our summary of the theory of generalized functions most closely follows the treatment described by Lighthill (1962), which is, of course, equivalent to the other treatments found in the literature. We focus on the case of scalar-valued generalized functions of a scalar variable.

To define generalized functions,[2] we first need the following definition:

DEFINITION S.1: *Let $\mathcal{T}$ be the set of all functions $s:\mathbb{R}\mapsto\mathbb{R}$ that* (i) *are everywhere differentiable any number of times and* (ii) *are such that*[3] $|(d^k s(t))/dt^k| = O(|t|^{-\ell})$ *as* $|t|\to\infty$ *for all* $k,\ell\in\mathbb{N}^+$. *Functions in $\mathcal{T}$ are called test functions.*

[1] This work was made possible in part through financial support from the National Science Foundation via Grant SES-0452089. The author thanks Jeremy Fox, Ricardo Mayer, Derek Neal, and Xiaohong Chen, as well as participants at seminars given at the Universities of Rochester, Chicago, Maryland, Michigan, UCSD, and UC-Riverside, the 2004 summer meetings of the Econometric Society, and the CIRANO/CIREQ “Operator Methods in Microeconometrics, Time Series and Finance” conference for their helpful comments.

[2] We adopt the term “generalized function” instead of “distribution” to avoid any potential confusion with the concept of probability distribution function.

[3] By convention, $d^k s(t)/dt^k = s(t)$ for $k = 0$.

Intuitively, functions in $\mathcal{T}$ are both extremely smooth and extremely thin-tailed.

DEFINITION S.2: *A generalized function b is a sequence of functions b_k in $\mathcal{T}$ such that $\lim_{k\to\infty} \int b_k(t)s(t)\,dt$ exists for all $s \in \mathcal{T}$. (Generalized functions can also be defined as bounded linear functionals on $\mathcal{T}$, but this definition is less convenient for our purposes.)*

Note that the limit of the sequence $b_k(t)$ may not be part of $\mathcal{T}$, which enables the concept of generalized functions to be more general than a function. The value of the integral $\int b(t)s(t)\,dt$ for a given $s \in \mathcal{T}$ is then defined as $\lim_{k\to\infty} \int b_k(t)s(t)\,dt$. Perhaps the best known example of a generalized function is the Dirac delta function $\delta(t)$, defined, for instance, by the sequence

$$b_k(t) = \sqrt{\frac{k}{2\pi}} \exp\left(-\frac{kt^2}{2}\right). \tag{S.1}$$

Another important example of a generalized function is the jth derivative of the delta function, denoted by $\delta^{(j)}(t)$ and defined by the sequence $d^j b_k(t)/dt^j$, where $b_k(t)$ is as in Equation (S.1). The generalized function $\delta^{(j)}(t)$ has the property that $\delta^{(0)}(t) \equiv \delta(t)$ and

$$\int \delta^{(j)}(t)s(t)\,dt = (-1)^j \left.\frac{d^j s(t)}{dt^j}\right|_{t=0} \quad \text{for} \quad j \in \mathbb{N}.$$

DEFINITION S.3: *Two generalized functions $a(t)$ and $b(t)$ are said to be equal if their associated sequences $a_k(t)$ and $b_k(t)$, respectively, are such that $\lim_{k\to\infty} \int a_k(t)s(t)\,dt = \lim_{k\to\infty} \int b_k(t)s(t)\,dt$ for all $s \in \mathcal{T}$.*

Note that Definition S.3 does not require that $a_k(t) = b_k(t)$ for all k and, hence, a given generalized function can be defined in terms of more than one sequence. The set of generalized functions is closed under addition, subtraction, and differentiation. The product of a generalized function with an ordinary function is guaranteed to be a generalized function if all of the ordinary function's derivatives exist and diverge no faster than a power of t as $|t| \to \infty$. However, the product of two generalized functions may not be a generalized function.

Ordinary functions can be viewed as particular cases of generalized functions. For instance, if we let $\mathcal{I}$ be the set of all ordinary functions $c(t)$ such that $\int (1+t^2)^{-\ell} |c(t)|\,dt$ is finite for some $\ell \in \mathbb{N}$, then all ordinary functions in $\mathcal{I}$ are also generalized functions. A generalized function $b(t)$ is said to be equal to an ordinary function $c(t)$ in an interval I if, for all $s \in \mathcal{T}$ that are supported on I, we have $\int b(t)s(t)\,dt = \int c(t)s(t)\,dt$. In the case of the Dirac delta function, $\delta(t)$ is equal to the 0 function over any interval that does not contain 0.

However, $\delta(t)$ is not equal to any ordinary function over any interval that includes 0. This concept is important because it will allow us to treat generalized functions as ordinary functions, as long as we stay away from their "singular" points. More generally, two generalized functions $b(t)$ and $c(t)$ are also said to be equal over an interval I if, for all $s \in \mathcal{T}$ that are supported on I, we have $\int b(t)s(t)\,dt = \int c(t)s(t)\,dt$.

Perhaps the most important result for our purpose is that the Fourier transform of a generalized function is a generalized function. As a particular case of this result, the Fourier transform of any function in $\mathcal{I}$ is a generalized function. Hence, in general, the Fourier transform of an ordinary function will not necessarily be an ordinary function, but rather will be a generalized function.

A generalized function $b(t)$ can always be decomposed as

$$b(t) = b_o(t) + b_s(t), \tag{S.2}$$

where $b_o(t)$ is an ordinary function while $b_s(t)$ is purely singular, consisting solely of a linear combination of delta function derivatives of a finite order.[4] This result directly follows from the fact that every generalized function can be written as the derivative of order $k \in \mathbb{N}$ of some continuous function $c(t)$ (Theorem III in Temple (1963) establishes this for a class of generalized functions including those considered here as a particular case). At every point t where $c(t)$ is k times differentiable in the usual sense, the generalized function can be written as an ordinary function, while at every point where $c(t)$ is not k times differentiable, a delta function derivative is created in the differentiation process. The fact that the two pieces are additively separable follows from the linear nature of the space of generalized functions. The decomposition (S.2) is unique because there exists no ordinary function $b_o(t)$ such that, for $k \in \mathbb{N}$, $\int b_o(t)s(t)\,dt = \int \delta^{(k)}(t)s(t)\,dt$ for all test function $s(t)$.

Moreover, the product of a generalized function $b(t)$ with an ordinary function $a_o(t)$ can be decomposed as

$$b(t)a_o(t) = b_o(t)a_o(t) + b_s(t)a_o(t), \tag{S.3}$$

where $b_o(t)a_o(t)$ is an ordinary function and where $b_s(t)a_o(t)$ is purely singular, as implied by Lemma 2 (see the Appendix in the main paper). Of course, $b(t)a_o(t)$ will only be well defined if $a_o(t)$ admits a sufficient number of continuous derivatives at the points where the delta function derivatives contained in $b(t)$ are located.

Although this review focuses on so-called *tempered* distributions, there exist more general classes of generalized functions. For instance, as described in Gel'fand and Shilov (1964), the set $\mathcal{T}$ can be limited to compactly supported

[4] The linear combination can consist of an infinite number of terms (with delta function derivatives at different locations).

infinitely differentiable functions, which expands the set of generalized functions for which the limit $\lim_{k\to\infty}\int a_k(t)s(t)\,dt$ exists for any $s \in \mathcal{T}$. However, in this work, we focus on functions $a(t)$ whose Fourier transforms $\alpha(\tau)$ are tempered distributions, therefore limiting ourselves to functions $a(t)$ that diverge no faster than some power of t as $|t| \to \infty$.

S.2. SIMPLE RESULTS ABOUT FOURIER TRANSFORMS AND GENERALIZED FUNCTIONS

DEFINITION S.4: *For some function $\psi(\zeta)$, let $d^{-1}\psi(\zeta)/d\zeta^{-1} \equiv \int_a^\zeta \psi(\xi)\,d\xi$ for some arbitrary constant a. For $k \geq 1$, define, by recursion,*

$$\frac{d^{-k-1}}{d\zeta^{-k-1}}\psi(\zeta) \equiv \frac{d^{-1}}{d\zeta^{-1}}\frac{d^{-k}}{d\zeta^{-k}}\psi(\zeta).$$

COMPLETE PROOF OF LEMMA 2: Let ψ be some test function in $\mathcal{T}$ as given in Definition S.1. By k repeated integration by parts, we have

$$\begin{aligned}&\int(\delta^{(k)}(\zeta)\phi(\zeta))\psi(\zeta)\,d\zeta\\&\quad=(-1)^k\int\left(\frac{d^{-k}}{d\zeta^{-k}}\delta^{(k)}(\zeta)\right)\frac{d^k}{d\zeta^k}(\phi(\zeta)\psi(\zeta))\,d\zeta,\end{aligned}$$

after noting that the boundary terms vanish due to the thin tails of $\psi(\zeta)$ and all of its derivatives. Next,

$$\begin{aligned}&\int(\delta^{(k)}(\zeta)\phi(\zeta))\psi(\zeta)\,d\zeta\\&\quad=(-1)^k\int\delta(\zeta)\left(\frac{d^k}{d\zeta^k}(\phi(\zeta)\psi(\zeta))\right)d\zeta\\&\quad=(-1)^k\int\delta(\zeta)\sum_{j=0}^{k}\binom{k}{j}\frac{d^{k-j}\phi(\zeta)}{d\zeta^{k-j}}\frac{d^j\psi(\zeta)}{d\zeta^j}\,d\zeta\\&\quad=(-1)^k\sum_{j=0}^{k}\binom{k}{j}\frac{d^{k-j}\phi(0)}{d\zeta^{k-j}}\frac{d^j\psi(0)}{d\zeta^j}\\&\quad=(-1)^k\sum_{j=0}^{k}\binom{k}{j}\frac{d^{k-j}\phi(0)}{d\zeta^{k-j}}\int\delta^{(j)}(\zeta)\psi(\zeta)\,d\zeta\\&\quad=\int\left((-1)^k\sum_{j=0}^{k}\binom{k}{j}\frac{d^{k-j}\phi(0)}{d\zeta^{k-j}}\delta^{(j)}(\zeta)\right)\psi(\zeta)\,d\zeta.\end{aligned}$$

Q.E.D.

COMPLETE PROOF OF LEMMA 4: Substituting Equations (55)–(58) into Equations (8) and (9), we obtain

$$\begin{aligned}
&\varepsilon_{y,o}(\zeta) + 2\pi \sum_{k=0}^{\bar{k}} \varepsilon_{y,k}(-\mathbf{i})^k \delta^{(k)}(\zeta) \\
&\quad = \left(\gamma_o(\zeta, \theta) + 2\pi \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^k \delta^{(k)}(\zeta) \right) \phi(\zeta), \\
&\mathbf{i}\varepsilon_{xy,o}(\zeta) + 2\pi\mathbf{i} \sum_{k=-1}^{\bar{k}} \varepsilon_{xy,k}(-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta) \\
&\quad = \left(\dot{\gamma}_o(\zeta, \theta) + 2\pi \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^k \delta^{(k+1)}(\zeta) \right) \phi(\zeta).
\end{aligned}$$

Equating the ordinary functions part of each expression yields

$$\begin{aligned}
&\varepsilon_{y,o}(\zeta) = \gamma_o(\zeta, \theta)\phi(\zeta), \\
&\mathbf{i}\varepsilon_{xy,o}(\zeta) = \dot{\gamma}_o(\zeta, \theta)\phi(\zeta),
\end{aligned}$$

while equating the singular parts yields

$$\begin{aligned}
&\sum_{k=0}^{\bar{k}} \varepsilon_{y,k}(-\mathbf{i})^k \delta^{(k)}(\zeta) = \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^k \delta^{(k)}(\zeta)\phi(\zeta), \\
&\sum_{k=-1}^{\bar{k}} \mathbf{i}\varepsilon_{xy,k}(-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta) = \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^k \delta^{(k+1)}(\zeta)\phi(\zeta).
\end{aligned}$$

By Lemma 2, we have

$$\begin{aligned}
&\sum_{k=0}^{\bar{k}} \varepsilon_{y,k}(-\mathbf{i})^k \delta^{(k)}(\zeta) = \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^k \sum_{j=0}^{k} \binom{k}{j} \phi^{(k-j)}(0)\delta^{(j)}(\zeta), \\
&\sum_{k=-1}^{\bar{k}} \mathbf{i}\varepsilon_{xy,k}(-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta) \\
&\quad = \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^k \sum_{j=0}^{k+1} \binom{k+1}{j} \phi^{(k+1-j)}(0)\delta^{(j)}(\zeta).
\end{aligned}$$

Simple manipulations then give

$$\begin{aligned}
&\sum_{k=0}^{\bar{k}} \varepsilon_{y,k}(-\mathbf{i})^{k}\delta^{(k)}(\zeta) \\
&\quad = \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^{k} \sum_{j=0}^{\bar{k}} \binom{k}{j} \mathbb{1}(j \le k)\phi^{(k-j)}(0)\delta^{(j)}(\zeta), \\
&\sum_{k=-1}^{\bar{k}} \mathbf{i}\varepsilon_{xy,k}(-\mathbf{i})^{k+1}\delta^{(k+1)}(\zeta) \\
&\quad = \sum_{k=0}^{\bar{k}} \gamma_k(\theta)(-\mathbf{i})^{k} \sum_{j=0}^{\bar{k}+1} \binom{k+1}{j} \mathbb{1}(j \le k+1)\phi^{(k+1-j)}(0)\delta^{(j)}(\zeta), \\
&\sum_{j=0}^{\bar{k}} \varepsilon_{y,j}(-\mathbf{i})^{j}\delta^{(j)}(\zeta) \\
&\quad = \sum_{j=0}^{\bar{k}} \sum_{k=0}^{\bar{k}} \binom{k}{j} \gamma_k(\theta)(-\mathbf{i})^{k} \mathbb{1}(j \le k)\phi^{(k-j)}(0)\delta^{(j)}(\zeta), \\
&\sum_{j=-1}^{\bar{k}} \mathbf{i}\varepsilon_{xy,j}(-\mathbf{i})^{j+1}\delta^{(j+1)}(\zeta) \\
&\quad = \sum_{j=0}^{\bar{k}+1} \sum_{k=0}^{\bar{k}} \binom{k+1}{j} \gamma_k(\theta)(-\mathbf{i})^{k} \mathbb{1}(j \le k+1)\phi^{(k+1-j)}(0)\delta^{(j)}(\zeta) \\
&\quad = \sum_{j=-1}^{\bar{k}} \sum_{k=0}^{\bar{k}} \binom{k+1}{j+1} \gamma_k(\theta)(-\mathbf{i})^{k} \mathbb{1}(j \le k)\phi^{(k-j)}(0)\delta^{(j+1)}(\zeta).
\end{aligned}$$

Equating the coefficients of the delta function derivatives of the same order gives

$$\begin{aligned}
\varepsilon_{y,j}(-\mathbf{i})^{j} &= \sum_{k=0}^{\bar{k}} \binom{k}{j} \gamma_k(\theta)(-\mathbf{i})^{k} \mathbb{1}(j \le k)\phi^{(k-j)}(0), \\
\mathbf{i}\varepsilon_{xy,j}(-\mathbf{i})^{j+1} &= \sum_{k=0}^{\bar{k}} \binom{k+1}{j+1} \gamma_k(\theta)(-\mathbf{i})^{k} \mathbb{1}(j \le k)\phi^{(k-j)}(0),
\end{aligned}$$

$$
\begin{aligned}
\varepsilon_{y,j}(-\mathbf{i})^j &= \sum_{l=-j}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta)(-\mathbf{i})^{j+l}\mathbb{1}(j \le j+l)\phi^{(j+l-j)}(0) \\
&= \sum_{l=-j}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta)(-\mathbf{i})^{j+l}\mathbb{1}(0 \le l)\phi^{(l)}(0) \\
&= \sum_{l=0}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta)(-\mathbf{i})^{j+l}\phi^{(l)}(0) \\
&= \sum_{l=0}^{\bar{k}} \binom{j+l}{j} \gamma_{j+l}(\theta)(-\mathbf{i})^{j+l}\mathbb{1}(l \le \bar{k}-j)\phi^{(l)}(0) \\
&= \sum_{k=0}^{\bar{k}} \binom{k+j}{j} \gamma_{k+j}(\theta)(-\mathbf{i})^{k+j}\mathbb{1}(k \le \bar{k}-j)\phi^{(k)}(0),
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{i}\varepsilon_{xy,j}(-\mathbf{i})^{j+1} \\
&\quad = \sum_{l=-j}^{\bar{k}-j} \binom{j+l+1}{j+1} \gamma_{j+l}(\theta)(-\mathbf{i})^{j+l}\mathbb{1}(j \le j+l)\phi^{(j+l-j)}(0) \\
&\quad = \sum_{l=0}^{\bar{k}-j} \binom{j+l+1}{j+1} \gamma_{j+l}(\theta)(-\mathbf{i})^{j+l}\phi^{(l)}(0) \\
&\quad = \sum_{l=0}^{\bar{k}+1} \binom{j+l+1}{j+1} \gamma_{j+l}(\theta)(-\mathbf{i})^{j+l}\mathbb{1}(l \le \bar{k}-j)\phi^{(l)}(0) \\
&\quad = \sum_{k=0}^{\bar{k}} \binom{k+j+1}{j+1} \gamma_{k+j}(\theta)(-\mathbf{i})^{k+j}\mathbb{1}(k \le \bar{k}-j)\phi^{(k)}(0) \\
&\quad\quad \text{for} \quad j \ge 0,
\end{aligned}
$$

$$
\begin{aligned}
\varepsilon_{y,j} &= \sum_{k=0}^{\bar{k}} \binom{k+j}{j} \gamma_{k+j}(\theta)\mathbb{1}(k \le \bar{k}-j)(-\mathbf{i})^k\phi^{(k)}(0), \\
\varepsilon_{xy,j} &= \sum_{k=0}^{\bar{k}} \binom{k+j+1}{j+1} \gamma_{k+j}(\theta)\mathbb{1}(k \le \bar{k}-j)(-\mathbf{i})^k\phi^{(k)}(0).
\end{aligned}
$$

Q.E.D.

S.3. ASYMPTOTICS OF THE GENERALIZED METHOD OF MOMENTS ESTIMATOR

S.3.1. *Definitions*

Let (X_j, Y_j, W_j) for $j = 1, \dots, n$ be a given sample. First, the variable Z_j needs to be constructed from the instruments W_j (see Equation (4) in the main text). To this effect, the parameter vector α in model (2) is estimated using standard (nonlinear) least squares on the specification

$$X_j = m(W_j, \alpha) + (\Delta X_j^* + \Delta X_j), \tag{S.4}$$

where $E[(\Delta X_j^* + \Delta X_j)|W_j] = 0$ by the assumptions of model (2). The resulting $\hat{\alpha}$ is used to define the variable $\hat{Z}_j$ as

$$\hat{Z}_j = m(W_j, \hat{\alpha}). \tag{S.5}$$

The variable $\hat{Z}_j$ estimates the true $Z_j = m(W_j, \alpha^*)$, where α^* denotes the true value of α. Let $p(\cdot|\alpha)$ denote the density of the quantity $m(W_j, \alpha)$ for a given α and let $p(z) = p(z|\alpha^*)$. Next, a nonparametric kernel density estimate of $p(\cdot|\hat{\alpha})$ at point $\hat{Z}_j$ can be obtained from

$$\hat{p}(\hat{Z}_j|\hat{\alpha}) = (nh)^{-1} \sum_{i=1, i\neq j}^{n} K((\hat{Z}_i - \hat{Z}_j)/h)$$

for some kernel $K(\cdot)$ and some bandwidth sequence $h \to 0$ as $n \to \infty$.

Finally, $\hat{\theta}$ is defined as the solution to $\hat{Q}(\theta, \hat{\alpha}) = 0$, where

$$\hat{Q}(\theta, \alpha) \equiv n^{-1} \sum_{j=1}^{n} \left(\frac{Y(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}(m(W_j, \alpha)|\alpha)} - \mathbf{e} \right) \mathbb{1}(\hat{p}(m(W_j, \alpha)|\alpha) \geq \tau), \tag{S.6}$$

$$Y(x, y, w, \theta, \alpha) = \begin{bmatrix} yr_y(m(w, \alpha), \theta) + xyr_{xy}(m(w, \alpha), \theta) \\ yr_{1y}(m(w, \alpha), \theta) \end{bmatrix}, \tag{S.7}$$

$$\mathbf{e} = (\underbrace{0, \dots, 0}_{N_\theta - N_s}, \underbrace{1, \dots, 1}_{N_s})', \tag{S.8}$$

where $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1 when the event $\cdot$ occurs, and τ is some trimming threshold such that $\tau \to 0$ as $n \to \infty$, which is designed to keep divisions by zero under control.[5] The scalar N_s is the dimension of the

[5] The trimming is not introduced to ensure that expectations such as $E[Yr_y(Z, \theta)/p(Z)]$ or $E[(Yr_y(Z, \theta)/p(Z))^2]$ exist, but rather to show that remainder terms are asymptotically negligible. If $E[(Yr_y(Z, \theta)/p(Z))^2]$, for instance, did not exist, no trimming scheme would restore the root n consistent estimation of the moment $E[Yr_y(Z, \theta)/p(Z)]$.

range of $r_{1y}(z,\theta)$ and can therefore be 0, 1, or 2. The true value of θ, denoted θ^*, is the solution to $Q(\theta,\alpha^*)=0$, where

$$Q(\theta,\alpha)=E\big[Q(X,Y,W,\theta,p(\cdot|\alpha))\big] \tag{S.9}$$

and where $Q(x,y,w,\theta,p)$ is defined in Equation (16).

S.3.2. *Proofs*

Although the following lemma may seem familiar, we were not able to find this result at the required level of generality in the existing literature (Theorems 1 and 3 in Andrews (1995) and Theorem 2.8 in Pagan and Ullah (1999) come very close, however).

LEMMA S.1: *Under Assumptions* 8, 14, *and* 15,

$$\sup_{\alpha\in\mathcal{A}}\sup_{z\in\mathbb{R}}|\tilde{p}(z|\alpha)-p(z|\alpha)|=O_p(n^{-1/2}h^{-1})+O(h^{N_K}),$$

where $\tilde{p}(z|\alpha)=(nh)^{-1}\sum_{j=1}^{n}K((Z_j-z)/h)$ *and* $p(z|\alpha)$ *is the density of* $Z=m(W,\alpha)$ *for a given function* $m(W,\alpha)$ *of some random vector* W. *The same result holds with* $\tilde{p}(z|\alpha)$ *replaced by* $\hat{p}(z|\alpha)=(nh)^{-1}\sum_{j=1}^{n}K((Z_j-z)/h)\mathbb{1}(Z_j\neq z)$.

PROOF: This proof is based in part on the proof of Theorem 2.8 in Pagan and Ullah (1999). Note that $\sup_{\alpha\in\mathcal{A}}\sup_{z\in\mathbb{R}}|\tilde{p}(z|\alpha)-p(z|\alpha)|\leq R+B$, where

$$\begin{aligned} R&=\sup_{\alpha\in\mathcal{A}}\sup_{z\in\mathbb{R}}\big|\tilde{p}(z|\alpha)-E[\tilde{p}(z|\alpha)]\big|,\\ B&=\sup_{\alpha\in\mathcal{A}}\sup_{z\in\mathbb{R}}\big|E[\tilde{p}(z|\alpha)]-p(z|\alpha)\big|.\end{aligned}$$

By the convolution theorem,

$$\begin{aligned} R&=\sup_{\alpha\in\mathcal{A}}\sup_{z\in\mathbb{R}}\left|\int\kappa(h\zeta)n^{-1}\sum_{j=1}^{n}(e^{\mathbf{i}\zeta Z_j}-E[e^{\mathbf{i}\zeta Z_j}])e^{-\mathbf{i}\zeta z}\,d\zeta\right|\\ &\leq\sup_{\alpha\in\mathcal{A}}\sup_{z\in\mathbb{R}}\int|\kappa(h\zeta)|\left|n^{-1}\sum_{j=1}^{n}(e^{\mathbf{i}\zeta Z_j}-E[e^{\mathbf{i}\zeta Z_j}])\right|d\zeta\\ &=\sup_{\alpha\in\mathcal{A}}\int|\kappa(h\zeta)|\left|n^{-1}\sum_{j=1}^{n}(e^{\mathbf{i}\zeta Z_j}-E[e^{\mathbf{i}\zeta Z_j}])\right|d\zeta,\end{aligned}$$

where $\kappa(\zeta)$ denotes the Fourier transform of $K(z)$. We then have

$$E[R]\leq\sup_{\alpha\in\mathcal{A}}\int|\kappa(h\zeta)|E\left[\left|n^{-1}\sum_{j=1}^{n}(e^{\mathbf{i}\zeta Z_j}-E[e^{\mathbf{i}\zeta Z_j}])\right|\right]d\zeta$$

$$
\begin{aligned}
&\leq \sup_{\alpha\in\mathcal{A}} \int |\kappa(h\zeta)| \left(E\left[\left| n^{-1} \sum_{j=1}^{n} (e^{\mathbf{i}\zeta Z_j} - E[e^{\mathbf{i}\zeta Z_j}]) \right|^2 \right] \right)^{1/2} d\zeta \\
&= \sup_{\alpha\in\mathcal{A}} \int |\kappa(h\zeta)| \big(n^{-1} E\big[(e^{\mathbf{i}\zeta Z_j} - E[e^{\mathbf{i}\zeta Z_j}]) \\
&\qquad \times (e^{-\mathbf{i}\zeta Z_j} - E[e^{-\mathbf{i}\zeta Z_j}]) \big] \big)^{1/2} d\zeta \\
&= \sup_{\alpha\in\mathcal{A}} n^{-1/2} \int |\kappa(h\zeta)| \big(E\big[(e^{\mathbf{i}\zeta Z_j} - E[e^{\mathbf{i}\zeta Z_j}]) \\
&\qquad\qquad \times (e^{-\mathbf{i}\zeta Z_j} - E[e^{-\mathbf{i}\zeta Z_j}]) \big] \big)^{1/2} d\zeta \\
&\leq n^{-1/2} 2^{1/2} \int |\kappa(h\zeta)| \, d\zeta \\
&= n^{-1/2} h^{-1} 2^{1/2} \int |\kappa(\zeta)| \, d\zeta \\
&= O(n^{-1/2} h^{-1})
\end{aligned}
$$

and $R = O_p(n^{-1/2} h^{-1})$ by Markov's inequality. Next,

$$
B = \sup_{\alpha\in\mathcal{A}} \sup_{z\in\mathbb{R}} \left| \int h^{-1} K(h^{-1} v) (p(z+v|\alpha) - p(z|\alpha)) \, dv \right|.
$$

By a Taylor expansion,

$$
\begin{aligned}
B &= \sup_{\alpha\in\mathcal{A}} \sup_{z\in\mathbb{R}} \left| \int h^{-1} K(h^{-1} v) \left(\sum_{j=1}^{N_k - 1} p^{(j)}(z|\alpha) \frac{v^j}{j!} + p^{(N_k)}(\tilde{z}|\alpha) \frac{v^{N_k}}{N_k!} \right) dv \right| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(for } \tilde{z} \in [z, z+v]) \\
&= \sup_{\alpha\in\mathcal{A}} \sup_{z\in\mathbb{R}} \left| \int h^{-1} K(h^{-1} v) p^{(N_k)}(\tilde{z}|\alpha) \frac{v^{N_k}}{N_k!} \, dv \right|
\end{aligned}
$$

by Assumption 14(iii). Then, by a change of variable,

$$
\begin{aligned}
B &= \sup_{\alpha\in\mathcal{A}} \sup_{z\in\mathbb{R}} \left| \int K(u) p^{(N_k)}(\tilde{z}|\alpha) \frac{u^{N_k} h^{N_k}}{N_k!} \, du \right| \\
&\leq h^{N_k} \Big(\sup_{\alpha\in\mathcal{A}} \sup_{z\in\mathbb{R}} |p^{(N_k)}(\tilde{z}|\alpha)| \Big) \frac{1}{N_k!} \left| \int |K(u)| |u|^{N_k} \, du \right| \\
&= O(h^{N_k})
\end{aligned}
$$

by Assumptions 14(iv) and 15.

The second assertion is shown by noting that the difference between $\tilde{p}(z)$ and $\hat{p}(z)$ is at most $K(0)n^{-1}h^{-1}$, which is of an order less than $n^{-1/2}h^{-1}$ and can therefore be absorbed in the $O_p(n^{-1/2}h^{-1})$ remainder. *Q.E.D.*

PROOF OF THEOREM 3: Let $\Upsilon(x, y, w, \theta, \alpha)$, $\hat{Q}(\theta, \alpha)$, and $Q(\theta, \alpha)$ be as defined in Section S.3.1. We first show consistency of $\hat{\theta}$. This involves establishing the uniform convergence of $\hat{Q}(\theta, \hat{\alpha})$ to $Q(\theta, \alpha^*)$ for $\theta \in \Theta$. We first note that $\hat{\alpha} \xrightarrow{p} \alpha^*$ by Lemma 2.4 and Theorem 2.1 in Newey and McFadden (1994), under Assumptions 8 and 9. Hence $\hat{\alpha} \in \mathcal{A}$ with probability approaching 1. We can then write, with probability approaching 1,

$$\begin{aligned}
&\sup_{\theta\in\Theta} \|\hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\| \\
&\quad \le \sup_{\theta\in\Theta} \|\hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \hat{\alpha})\| + \sup_{\theta\in\Theta} \|Q(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\| \\
&\quad \le \sup_{\alpha\in\mathcal{A}} \sup_{\theta\in\Theta} \|\hat{Q}(\theta, \alpha) - Q(\theta, \alpha)\| + \sup_{\theta\in\Theta} \|Q(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\|,
\end{aligned}$$

where $\sup_{\theta\in\Theta} \|Q(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\| \xrightarrow{p} 0$ by $\hat{\alpha} \xrightarrow{p} \alpha^*$ and Assumption 18. Next,

$$\sup_{\alpha\in\mathcal{A}} \sup_{\theta\in\Theta} \|\hat{Q}(\theta, \alpha) - Q(\theta, \alpha)\| \le R_A + R_I + R_D,$$

where

$$R_A = \sup_{\alpha\in\mathcal{A}} \sup_{\theta\in\Theta} \left\| n^{-1} \sum_{j=1}^{n} \frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j, \alpha)|\alpha)} - E\left[\frac{\Upsilon(X, Y, W, \theta, \alpha)}{p(m(W, \alpha)|\alpha)}\right] \right\|,$$

$$\begin{aligned}
R_I = \sup_{\alpha\in\mathcal{A}} \sup_{\theta\in\Theta} \Bigg\| n^{-1} \sum_{j=1}^{n} & \frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j, \alpha)|\alpha)} \\
& \times \big(\mathbb{1}\big(\hat{p}(m(W_j, \alpha)|\alpha) \ge \tau\big) - 1\big) \Bigg\|,
\end{aligned}$$

$$\begin{aligned}
R_D = \sup_{\alpha\in\mathcal{A}} \sup_{\theta\in\Theta} \Bigg\| n^{-1} \sum_{j=1}^{n} & \Upsilon(X_j, Y_j, W_j, \theta, \alpha) \\
& \times \frac{p(m(W_j, \alpha)|\alpha) - \hat{p}(m(W_j, \alpha)|\alpha)}{\hat{p}(m(W_j, \alpha)|\alpha) p(m(W_j, \alpha)|\alpha)} \\
& \times \mathbb{1}\big(\hat{p}(m(W_j, \alpha)|\alpha) \ge \tau\big) \Bigg\|.
\end{aligned}$$

We then have $\sup_{\theta\in\Theta}\|R_A\| \xrightarrow{p} 0$ by Assumptions 8, 10, and 11 and Lemma 2.4 in Newey and McFadden (1994). Next, by Lemma S.1, we have

$$
\begin{aligned}
R_I &\leq \sup_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} n^{-1}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha)|\alpha)}\big|\mathbb{1}\big(\hat{p}(m(W_j, \alpha)|\alpha) < \tau\big)\big| \\
&\leq \sup_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} n^{-1}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha)|\alpha)} \\
&\qquad\times\big|\mathbb{1}\big(p(m(W_j, \alpha)|\alpha) - Cn^{\epsilon-1/2}h^{-1} < \tau\big)\big| \\
&\qquad\qquad\text{(with probability approaching 1, for } \epsilon \in \left]0, 1/4\right[\text{)} \\
&= \sup_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} n^{-1}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha)|\alpha)} \\
&\qquad\times\big|\mathbb{1}\big(p(m(W_j, \alpha)|\alpha) < \tau(1 + Cn^{\epsilon-1/2}h^{-1}/\tau)\big)\big| \\
&\leq \sup_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} n^{-1}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha)|\alpha)}\big|\mathbb{1}\big(p(m(W_j, \alpha)|\alpha) < 2\tau\big)\big| \\
&\qquad\qquad\text{(by Assumption 16)}
\end{aligned}
$$

and $E[R_I] \leq E[\sup_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta}\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\||\mathbb{1}(p(Z_j) < 2\tau)|/p(m(W_j, \alpha)|\alpha)] = o(n^{-1/2})$ by Assumption 17, thus implying that $R_I = o_p(n^{-1/2})$, by Markov's inequality. Next,

$$
\begin{aligned}
R_D &\leq \sup_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} n^{-1}\sum_{j=1}^{n}\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\| \\
&\qquad\times\left(\frac{|p(m(W_j, \alpha)|\alpha) - \hat{p}(m(W_j, \alpha)|\alpha)|}{\hat{p}(Z_j)p(m(W_j, \alpha)|\alpha)}\right)\hat{I}_j \\
&\leq \sup_{\alpha\in\mathcal{A}}\sup_{\theta\in\Theta} \tau^{-1}n^{-1}\sum_{j=1}^{n}\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\| \\
&\qquad\times\left(\frac{|p(m(W_j, \alpha)|\alpha) - \hat{p}(m(W_j, \alpha)|\alpha)|}{p(m(W_j, \alpha)|\alpha)}\right)\hat{I}_j \\
&\leq \sup_{z\in\mathbb{R}}|p(z) - \hat{p}(z)|\tau^{-1}n^{-1}\sum_{j=1}^{n}\left(\frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha)|\alpha)}\right) \\
&= (O_p(n^{-1/2}h^{-1}) + O(h^{N_K}))\tau^{-1}O_p(1)
\end{aligned}
$$

by Lemma S.1, and Lemma 2.4 in Newey and McFadden (1994), under Assumptions 8, 10, and 11. By Assumption 16, $n^{-1/2}h^{-1}\tau^{-1} \to 0$ and $h^{N_k} \to 0$, and it follows that $R_D \xrightarrow{p} 0$.

Having shown that $\sup_{\theta\in\Theta} \|\hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\| \xrightarrow{p} 0$, we now establish that this implies[6] that $\hat{\theta}$ converges to θ^*. Because $\hat{Q}(\hat{\theta}, \hat{\alpha}) = 0$ and $\sup_{\theta\in\Theta} \|\hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\| \xrightarrow{p} 0$, it follows that $\operatorname{plim}_{n\to\infty} Q(\hat{\theta}, \alpha^*) = 0$. Because $\hat{Q}(\theta, \hat{\alpha})$ is continuous in θ (because $Y(x_j, y_j, w_j, \theta, \alpha)$ is) and its convergence to $Q(\theta, \alpha^*)$ is uniform in θ, $Q(\theta, \alpha^*)$ must be continuous in θ. Combining these two results yields $\operatorname{plim}_{n\to\infty} Q(\hat{\theta}, \alpha^*) = Q(\operatorname{plim}_{n\to\infty} \hat{\theta}, \alpha^*) = 0$. Given that $\theta = \theta^*$ is the only solution to $Q(\theta, \alpha^*) = 0$ by Assumption 6, we conclude that $\operatorname{plim}_{n\to\infty} \hat{\theta} = \theta^*$.

Having shown consistency, we turn to asymptotic normality and root n consistency. By a standard mean value expansion of the first-order conditions $\hat{Q}(\hat{\theta}, \hat{\alpha}) = 0$ around θ^* and the usual manipulations,

$$n^{1/2}(\hat{\theta} - \theta^*) = -\left(\frac{\partial \hat{Q}(\bar{\theta}, \hat{\alpha})}{\partial \theta'}\right)^{-1} n^{1/2}\hat{Q}(\theta^*, \hat{\alpha}) \tag{S.10}$$

for some mean value $\bar{\theta}$. Following the same steps used previously to show uniform convergence in probability of $\hat{Q}(\theta, \hat{\alpha})$, we can show that $\sup_{\theta\in\mathcal{N}} \|\partial\hat{Q}(\theta, \hat{\alpha})/\partial\theta' - \partial Q(\theta, \alpha^*)/\partial\theta'\| \xrightarrow{p} 0$ and $\partial Q(\theta, \alpha^*)/\partial\theta'$ is continuous in θ by simply replacing Assumption 11 with Assumption 12. Given that $\hat{\theta} \xrightarrow{p} \theta^*$, it follows that $\bar{\theta} \xrightarrow{p} \theta^*$ and that $\partial Q(\bar{\theta}, \alpha^*)/\partial\theta' \xrightarrow{p} \partial Q(\theta^*, \alpha^*)/\partial\theta'$, thus implying that

$$\frac{\partial \hat{Q}(\bar{\theta}, \hat{\alpha})}{\partial \theta'} \xrightarrow{p} \frac{\partial Q(\theta^*, \alpha^*)}{\partial \theta'}. \tag{S.11}$$

Next, we let $Y_j = Y(X_j, Y_j, W_j, \theta^*, \alpha^*)$, $Z_j = m(W_j, \alpha^*)$, $\hat{p}(Z_j) = \hat{p}(m(W_j, \alpha^*)|\alpha^*)$, $p(Z_j) = p(m(W_j, \alpha^*)|\alpha^*)$, $\hat{I}_j = \mathbb{1}(\hat{p}(Z_j) \geq \tau)$, and $I_j = \mathbb{1}(p(Z_j) \geq \tau)$, and decompose the term $n^{1/2}\hat{Q}(\theta^*, \hat{\alpha})$ in Equation S.10 as

$$n^{1/2}\hat{Q}(\theta^*, \hat{\alpha}) = N + N_\alpha + R_{T1} + R_{T2} + R_{T3} + R_L + R_U + R_B + R_{\text{sec}},$$

where the asymptotically normal terms are given by

$$N = n^{-1/2} \sum_{j=1}^{n} \frac{Y_j - E[Y_j|Z_j]}{p(Z_j)},$$

$$N_\alpha = n^{1/2}(Q(\theta^*, \hat{\alpha}) - Q(\theta^*, \alpha^*)),$$

[6] This would be obvious if $\hat{\theta}$ were defined as the maximizer of a random function. Here $\hat{\theta}$ is the solution to a set of equations and the usual consistency result (e.g., Theorem 2.1 in Newey and McFadden (1994)) does not directly apply.

while the remainder terms associated with trimming are

$$\begin{aligned}
R_{T1} &= n^{-1/2}\sum_{j=1}^{n}\frac{Y_j}{\hat{p}(Z_j)}(\hat{I}_j - I_j),\\
R_{T2} &= n^{1/2}E\left[\frac{Y_j}{p(Z_j)}(1 - I_j)\right],\\
R_{T3} &= n^{-1/2}\sum_{j=1}^{n}\frac{(Y_j - E[Y_j|Z_j])}{p(Z_j)}(I_j - 1),
\end{aligned}$$

the remainder from the linearization is given by

$$R_L = n^{-1/2}\sum_{j=1}^{n}\frac{Y_j}{\hat{p}(Z_j)p^2(Z_j)}(\hat{p}(Z_j) - p(Z_j))^2 I_j,$$

the "U-statistic" term is

$$\begin{aligned}
R_U = -n^{-1/2}\sum_{j=1}^{n}\Bigg(&\frac{Y_j}{p^2(Z_j)}\big(\hat{p}(Z_j) - E[\hat{p}(Z_j)|Z_j]\big)I_j\\
&- \left(\frac{E[Y_j|Z_j]}{p(Z_j)}I_j - E\left[\frac{Y_j}{p(Z_j)}I_j\right]\right)\Bigg),
\end{aligned}$$

the "bias" term is

$$R_B = n^{-1/2}\sum_{j=1}^{n}\frac{Y_j}{p^2(Z_j)}\big(p(Z_j) - E[\hat{p}(Z_j)|Z_j]\big)I_j,$$

and the "stochastic equicontinuity" remainder term is

$$R_{\text{sec}} = n^{1/2}\big((\hat{Q}(\theta^*, \hat{\alpha}) - Q(\theta^*, \hat{\alpha})) - (\hat{Q}(\theta^*, \alpha^*) - Q(\theta^*, \alpha^*))\big).$$

We consider each remainder in turn:

$$\begin{aligned}
|R_{T1}| &\le n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{\hat{p}(Z_j)}\big|\mathbb{1}(\hat{p}(Z_j) \ge \tau) - \mathbb{1}(p(Z_j) \ge \tau)\big|\\
&\le n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j) - Cn^{\epsilon-1/2}h^{-1}}\big|\mathbb{1}(\hat{p}(Z_j) \ge \tau) - 1(p(Z_j) \ge \tau)\big|
\end{aligned}$$

(with probability approaching 1, for $\epsilon \in \,]0, 1/4[$)

$$
\begin{aligned}
&\le n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)-\frac{p(Z_j)}{\tau-Cn^{\epsilon-1/2}h^{-1}}Cn^{\epsilon-1/2}h^{-1}} \\
&\qquad \times\left|\mathbb{1}(\hat{p}(Z_j)\ge\tau)-\mathbb{1}(p(Z_j)\ge\tau)\right| \\
&\hspace{20em}\text{(with probability approaching 1)} \\
&=\frac{1}{1-\frac{1}{C\tau n^{1/2-\epsilon}h-1}}n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\left|\mathbb{1}(\hat{p}(Z_j)\ge\tau)-\mathbb{1}(p(Z_j)\ge\tau)\right| \\
&=O(1)n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\Big|\mathbb{1}(\hat{p}(Z_j)\ge\tau\text{ and }p(Z_j)<\tau) \\
&\hspace{10em}-\mathbb{1}(p(Z_j)\ge\tau\text{ and }\hat{p}(Z_j)<\tau)\Big| \\
&=O(1)n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(\hat{p}(Z_j)\ge\tau\text{ and }p(Z_j)<\tau) \\
&\quad+O(1)n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)\ge\tau\text{ and }\hat{p}(Z_j)<\tau) \\
&\le O(1)n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau) \\
&\quad+O(1)n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(\hat{p}(Z_j)<\tau) \\
&\le O(1)n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau) \\
&\quad+O(1)n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau-Cn^{\epsilon-1/2}h^{-1}) \\
&\hspace{20em}\text{(with probability approaching 1)}
\end{aligned}
$$

where

$$
\begin{aligned}
E\left[n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau)\right]&=n^{1/2}E\left[\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau)\right] \\
&=o(1),
\end{aligned}
$$

$$\begin{aligned}
&E\left[n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau-Cn^{\epsilon-1/2}h^{-1})\right]\\
&\quad=n^{1/2}E\left[\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau-Cn^{\epsilon-1/2}h^{-1})\right]\\
&\quad=n^{1/2}E\left[\frac{|Y_j|}{p(Z_j)}\mathbb{1}\left(p(Z_j)<\tau\left(\frac{1-Cn^{\epsilon-1/2}h^{-1}}{\tau}\right)\right)\right]\\
&\quad=n^{1/2}E\left[\frac{|Y_j|}{p(Z_j)}\mathbb{1}\big(p(Z_j)<\tau(1-o(1))\big)\right]\\
&\quad\to n^{1/2}E\left[\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)<\tau)\right]=o(1),
\end{aligned}$$

and, by Markov's inequality, $R_{T1}=o_p(1)$. Next,

$$\begin{aligned}
|R_{T2}|&\le n^{1/2}E\left[\frac{|Y_j|}{p(Z_j)}|I_j-1|\right]\\
&=n^{1/2}E\left[\frac{|Y_j|}{p(Z_j)}\mathbb{1}(p(Z_j)\le\tau)\right]\\
&=n^{1/2}o(n^{-1/2})=o(1)
\end{aligned}$$

and

$$\begin{aligned}
E[|R_{T3}|]&=E\left[\left|n^{-1/2}\sum_{j=1}^{n}\frac{Y_j-E[Y_j|Z_j]}{p(Z_j)}(I_j-1)\right|\right]\\
&\le n^{1/2}2E\left[\frac{|Y_j|}{p(Z_j)}|I_j-1|\right]\\
&=n^{1/2}o(n^{-1/2})=o(1),
\end{aligned}$$

implying that $|R_{T3}|=o_p(1)$ as well by the Markov inequality. The linearization remainder is then

$$\begin{aligned}
|R_L|&=\left|n^{-1/2}\sum_{j=1}^{n}\frac{Y_j}{\hat{p}(Z_j)p^2(Z_j)}(\hat{p}(Z_j)-p(Z_j))^2I_j\right|\\
&\le n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{\hat{p}(Z_j)p^2(Z_j)}|\hat{p}(Z_j)-p(Z_j)|^2I_j
\end{aligned}$$

$$
\begin{aligned}
&\le n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{(p(Z_j)-Cn^{-1/2}h^{-1})p^2(Z_j)}|\hat p(Z_j)-p(Z_j)|^2 I_j\\
&\le n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{(\tau-Cn^{-1/2}h^{-1})\tau p(Z_j)}|\hat p(Z_j)-p(Z_j)|^2 I_j\\
&=\frac{1}{\tau^2}\left(1-\frac{Cn^{-1/2}h^{-1}}{\tau}\right)^{-1} n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}|\hat p(Z_j)-p(Z_j)|^2 I_j\\
&\le\frac{2}{\tau^2}n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}|\hat p(Z_j)-p(Z_j)|^2\\
&\le\frac{2Cn^{-1}h^{-2}}{\tau^2}n^{1/2}n^{-1}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\\
&\le\frac{2Cn^{-1}h^{-2}}{\tau^2}n^{1/2}\left(n^{-1}\sum_{j=1}^{n}\frac{|Y_j|^2}{p^2(Z_j)}\right)^{1/2}\\
&=\frac{2Cn^{-1}h^{-2}}{\tau^2}n^{1/2}O_p(1)\\
&=o(n^{-1/2})n^{1/2}O_p(1)=o_p(1).
\end{aligned}
$$

The “U-statistic” term can be written as

$$
\begin{aligned}
-R_U&=n^{-1/2}\sum_{j=1}^{n}(n-1)^{-1}\\
&\quad\times\sum_{i\neq j}\biggl(\frac{Y_jI_j}{p^2(Z_j)}\bigl(K_h(Z_i-Z_j)-E[K_h(Z_i-Z_j)|Z_j]\bigr)\\
&\qquad\qquad-\left(\frac{E[Y_j|Z_j]}{p(Z_j)}I_j-E\left[\frac{Y_j}{p(Z_j)}I_j\right]\right)\biggr)\\
&=n^{-1/2}\sum_{j=1}^{n}(n-1)^{-1}\\
&\quad\times\sum_{i\neq j}\left(\frac{Y_jI_j}{2p^2(Z_j)}+\frac{Y_iI_i}{2p^2(Z_i)}\right)\\
&\quad\times\bigl(K_h(Z_i-Z_j)-E[K_h(Z_i-Z_j)|Z_j]\bigr)
\end{aligned}
$$

$$
\begin{aligned}
&-\left(\frac{E[Y_i|Z_i]}{p(Z_i)}I_i - E\left[\frac{Y_i}{p(Z_i)}I_i\right]\right)\\
&= n^{1/2}\binom{n}{2}^{-1}\sum_{j=1}^{n}\sum_{i=j+1}^{n} U((Y_j, Z_j), (Y_i, Z_i)),
\end{aligned}
$$

where $K_h(z) = h^{-1}K(z/h)$ and

$$
\begin{aligned}
U((Y_j, Z_j), (Y_i, Z_i)) &= \left(\frac{Y_j I_j}{2p^2(Z_j)} + \frac{Y_i I_i}{2p^2(Z_i)}\right)\\
&\quad \times \left(K_h(Z_i - Z_j) - E[K_h(Z_i - Z_j)|Z_j]\right)\\
&\quad - \left(\frac{E[Y_i|Z_i]}{p(Z_i)}I_i - E\left[\frac{Y_i}{p(Z_i)}I_i\right]\right).
\end{aligned}
$$

Using the U-statistic projection theorem (e.g., Lemma 3.1 in Powell, Stock, and Stoker (1989)), standard but tedious manipulations show that $R_U = o_p(1)$ under Assumptions 14 and 16. Finally, the bias term is

$$
\begin{aligned}
|R_B| &\le n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p^2(Z_j)}\left|p(Z_j) - E[\hat{p}(Z_j)|Z_j]\right|I_j\\
&\le \tau^{-1}n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\left|p(Z_j) - E[\hat{p}(Z_j)|Z_j]\right|I_j\\
&\le \tau^{-1}n^{-1/2}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}\left|p(Z_j) - E[\hat{p}(Z_j)|Z_j]\right|\\
&\le \tau^{-1}n^{1/2}n^{-1}\sum_{j=1}^{n}\frac{|Y_j|}{p(Z_j)}Ch^{N_K} \qquad \text{by Lemma S.1}
\end{aligned}
$$

and $|R_B| = O_p(n^{1/2}h^{N_K}\tau^{-1}) = o_p(1)$ because $n^{1/2}h^{N_K}\tau^{-1} \to 0$ by Assumption 16.

To bound the R_{sec} term, let $S_\tau(t)$ be continuously differentiable in t for all $\tau \neq 0$ and such that (i) $\mathbb{1}(t \ge \tau) = 0 \Leftrightarrow S_\tau(t) = 0$, (ii) $\mathbb{1}(t \ge \tau) = 1 \Leftrightarrow S_{2\tau}(t) = 1$, (iii) $0 \le S_\tau(t) \le 1$, and (iv) $\sup_{t\in\mathbb{R}} |dS_\tau(t)/dt| = O(\tau)$. We then decompose $\hat{Q}(\theta^*, \alpha)$ as

$$
\hat{Q}(\theta^*, \alpha) = \hat{Q}_S(\theta^*, \alpha) + R_S(\alpha),
$$

where $\hat{Q}_S(\theta^*, \alpha)$ is continuous in α, while $R_S(\alpha)$ may not be, and they are given by

$$\hat{Q}_S(\theta^*, \alpha) = n^{-1} \sum_{j=1}^{n} \frac{Y(X_j, Y_j, W_j, \theta^*, \alpha)}{\hat{p}(m(W_j, \alpha)|\alpha)} S_\tau\big(\hat{p}(m(W_j, \alpha)|\alpha)\big),$$

$$\begin{aligned} R_S(\alpha) &= n^{-1} \sum_{j=1}^{n} \frac{Y(X_j, Y_j, W_j, \theta^*, \alpha)}{\hat{p}(m(W_j, \alpha)|\alpha)} \\ &\qquad \times \big(\mathbb{1}\big(\hat{p}(m(W_j, \alpha)|\alpha) \geq \tau\big) - S_\tau\big(\hat{p}(m(W_j, \alpha)|\alpha)\big)\big). \end{aligned}$$

The remainder $R_S(\alpha)$ satisfies $\sup_{\alpha \in \mathcal{A}} \|R_S(\alpha)\| = o_p(n^{-1/2})$ because

$$\begin{aligned} &\sup_{\alpha \in \mathcal{A}} \|R_S(\alpha)\| \\ &\quad \leq \sup_{\alpha \in \mathcal{A}} n^{-1} \sum_{j=1}^{n} \frac{\|Y(X_j, Y_j, W_j, \theta^*, \alpha)\|}{p(m(W_j, \alpha)|\alpha) - n^{-1/2}h^{-1}} \mathbb{1}\big(\hat{p}(m(W_j, \alpha)|\alpha) < 2\tau\big) \\ &\quad = o_p(n^{-1/2}) \end{aligned}$$

by Assumption 17 and the same arguments as used for R_{T1}. We can then write R_{sec} as

$$\begin{aligned} R_{\text{sec}} &= n^{1/2}\big((\hat{Q}(\theta^*, \hat{\alpha}) - Q(\theta^*, \hat{\alpha})) - (\hat{Q}(\theta^*, \alpha^*) - Q(\theta^*, \alpha^*))\big) \\ &= n^{1/2}\big((\hat{Q}_S(\theta^*, \hat{\alpha}) - Q(\theta^*, \hat{\alpha})) - (\hat{Q}_S(\theta^*, \alpha^*) - Q(\theta^*, \alpha^*))\big) \\ &\quad + o_p(1) \\ &= \left(\frac{\partial \hat{Q}(\theta^*, \bar{\alpha})}{\partial \alpha'} - \frac{\partial Q(\theta^*, \bar{\alpha})}{\partial \alpha'}\right) n^{1/2}(\hat{\alpha} - \alpha^*) + o_p(1) \end{aligned} \tag{S.12}$$

for some mean value $\bar{\alpha}$. We then decompose $(\partial/\partial\alpha')\hat{Q}_S(\theta^*, \alpha)$ as

$$\frac{\partial}{\partial \alpha'} \hat{Q}_S(\theta^*, \alpha) = D_1 + D_2 + R_{DS},$$

where

$$D_1 = n^{-1/2} \sum_{j=1}^{n} \left(\frac{\frac{\partial}{\partial \alpha'} Y(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}(m(W_j, \alpha)|\alpha)}\right) S_\tau\big(\hat{p}(m(W_j, \alpha)|\alpha)\big),$$

$$D_2 = -n^{-1/2} \sum_{j=1}^{n} \left(\frac{Y(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}^2(m(W_j, \alpha)|\alpha)} \frac{\partial}{\partial \alpha'} \hat{p}(m(W_j, \alpha)|\alpha)\right)$$

$$\times S_\tau\big(\hat{p}(m(W_j,\alpha)|\alpha)\big),$$

$$R_{DS} = n^{-1/2}\sum_{j=1}^{n}\frac{\Upsilon(X_j,Y_j,W_j,\theta,\alpha)}{\hat{p}(m(W_j,\alpha)|\alpha)}\frac{\partial S_\tau(\hat{p}(m(W_j,\alpha)|\alpha))}{\partial\alpha'}.$$

The R_{DS} term is negligible, because

$$\begin{aligned}
&\left\| n^{-1/2}\sum_{j=1}^{n}\frac{\Upsilon(X_j,Y_j,W_j,\theta,\alpha)}{\hat{p}(m(W_j,\alpha)|\alpha)}\frac{\partial S_\tau(\hat{p}(m(W_j,\alpha)|\alpha))}{\partial\alpha'}\right\| \\
&\quad\le n^{-1/2}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j,Y_j,W_j,\theta,\alpha)\|}{\tau}\left|\frac{\partial S_\tau(\hat{p}(m(W_j,\alpha)|\alpha))}{\partial\alpha'}\right| \\
&\quad\le n^{-1/2}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j,Y_j,W_j,\theta,\alpha)\|}{\tau}C\tau\mathbb{1}\big(p(m(W_j,\alpha)|\alpha)>\tau\big) \\
&\quad= Cn^{-1/2}\sum_{j=1}^{n}\|\Upsilon(X_j,Y_j,W_j,\theta,\alpha)\|\mathbb{1}\big(p(m(W_j,\alpha)|\alpha)>\tau\big) \\
&\quad= Cn^{-1/2}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j,Y_j,W_j,\theta,\alpha)\|}{p(m(W_j,\alpha)|\alpha)}p(m(W_j,\alpha)|\alpha) \\
&\qquad\times\mathbb{1}\big(p(m(W_j,\alpha)|\alpha)>\tau\big) \\
&\quad\le Cn^{1/2}n^{-1}\sum_{j=1}^{n}\frac{\|\Upsilon(X_j,Y_j,W_j,\theta,\alpha)\|}{p(m(W_j,\alpha)|\alpha)}\mathbb{1}\big(p(m(W_j,\alpha)|\alpha)>\tau\big) \\
&\qquad\qquad\text{(since } p(z|\alpha) \text{ is bounded by Assumption 15)} \\
&\quad= n^{1/2}o_p(n^{-1/2}) = o_p(1) \\
&\qquad\qquad\text{(by Markov's inequality and Assumption 17.)}
\end{aligned}$$

Now, the terms D_1 and D_2 can be handled through the same techniques as those used to show uniform convergence of $\hat{Q}(\theta,\hat{\alpha})$ after noting that trimming by $S_\tau(\hat{p}(m(W_j,\alpha)|\alpha))$ is asymptotically equivalent to trimming by $\mathbb{1}(\hat{p}(m(W_j,\alpha)|\alpha)\ge\tau)$. Under Assumption 19 and by using an expansion of the form

$$\begin{aligned}
D_1 &= n^{-1/2}\sum_{j=1}^{n}\frac{\frac{\partial}{\partial\alpha'}\Upsilon(X_j,Y_j,W_j,\theta,\alpha)}{p(m(W_j,\alpha)|\alpha)}S_\tau\big(\hat{p}(m(W_j,\alpha)|\alpha)\big) \\
&\quad - n^{-1/2}\sum_{j=1}^{n}\frac{\frac{\partial}{\partial\alpha'}\Upsilon(X_j,Y_j,W_j,\theta,\alpha)}{p(m(W_j,\alpha)|\alpha)}
\end{aligned}$$

$$\times \frac{(\hat{p}(m(W_j,\alpha)|\alpha) - p(m(W_j,\alpha)|\alpha))}{\hat{p}(m(W_j,\alpha)|\alpha)}$$

$$\times S_\tau\big(\hat{p}(m(W_j,\alpha)|\alpha)\big),$$

$$D_2 = n^{-1/2}\sum_{j=1}^{n} \frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j,\alpha)|\alpha)}$$

$$\times \left(1 - \frac{(\hat{p}(m(W_j,\alpha)|\alpha) - p(m(W_j,\alpha)|\alpha))}{\hat{p}(m(W_j,\alpha)|\alpha)}\right)$$

$$\times \left(\frac{\partial}{\partial\alpha'} p(m(W_j,\alpha)|\alpha)\right.$$

$$\left.+ \left(\frac{\partial}{\partial\alpha'}\hat{p}(m(W_j,\alpha)|\alpha) - \frac{\partial}{\partial\alpha'} p(m(W_j,\alpha)|\alpha)\right)\right)$$

$$\Big/\hat{p}(m(W_j,\alpha)|\alpha),$$

it can be shown that $D_1 \xrightarrow{p} E[(\partial\Upsilon(X_j, Y_j, W_j, \theta, \alpha)/\partial\alpha')/p(m(W_j,\alpha)|\alpha)]$ and $D_2 \xrightarrow{p} E[(\Upsilon(X_j, Y_j, W_j, \theta, \alpha) / p^2(m(W_j, \alpha)|\alpha))(\partial p(m(W_j, \alpha)|\alpha)/\partial\alpha')]$ uniformly in α for $\alpha \in \mathcal{A}$. (The convergence rate of $\partial\hat{p}(m(W_j,\alpha)|\alpha)/\partial\alpha' - \partial p(m(W_j,\alpha)|\alpha)/\partial\alpha'$ is obtained as in the proof of Lemma S.1, with N_K replaced by $N_K - 1$.) This implies by Assumption 18 that

$$\sup_{\alpha\in\mathcal{A}}\left(\frac{\partial\hat{Q}(\theta^*,\alpha)}{\partial\alpha'} - \frac{\partial Q(\theta^*,\alpha)}{\partial\alpha'}\right) \xrightarrow{p} 0,$$

and by Equation (S.12) and the fact that $\hat{\alpha} - \alpha^* = O_p(n^{-1/2})$, we have that $R_{\text{sec}} = o_p(1)$.

Having bounded all remainder terms, we note that the N term clearly satisfies

$$N = n^{-1/2}\sum_{j=1}^{n} \psi_\theta(X_j, Y_j, W),$$

where $E[\psi_\theta(X_j, Y_j, W)\psi'_\theta(X_j, Y_j, W)]$ is finite under Assumption 13.

By a mean-value expansion, the N_α term is equal to

$$N_\alpha = \frac{\partial Q(\theta^*, \bar{\alpha})}{\partial\alpha'} n^{1/2}(\hat{\alpha} - \alpha^*)$$

for some mean value $\bar{\alpha}$. Given that $\hat{\alpha} \xrightarrow{p} \alpha^*$ and therefore $\bar{\alpha} \xrightarrow{p} \alpha^*$, Assumption 18 implies that $\partial Q(\theta^*, \bar{\alpha})/\partial\alpha' \xrightarrow{p} \partial Q(\theta^*, \alpha)/\partial\alpha'$.

By standard results (such as Theorem 3.1 in Newey and McFadden (1994)), Assumptions 8 and 9 imply that the first-step estimate $\hat{\alpha}$ is a root n consistent estimator of α^* with influence function equal to

$$\psi_\alpha(x, w) = -\left(E\left[\frac{\partial m(W, \alpha^*)}{\partial \alpha}\frac{\partial m(W, \alpha^*)}{\partial \alpha'}\right]\right)^{-1} \times \frac{\partial m(w, \alpha^*)}{\partial \alpha}(x - m(w, \alpha^*))$$

and such that $E[\psi_\alpha(X, W)\psi'_\alpha(X, W)]$ exists. Hence, we can write

$$N_\alpha = n^{-1/2}\sum_{j=1}^{n}\frac{\partial Q(\theta^*, \alpha)}{\partial \alpha'}\psi_\alpha(X_j, W_j).$$

We have just established that

$$\begin{aligned} &n^{1/2}\hat{Q}(\theta^*, \hat{\alpha}) \\ &\quad = n^{-1/2}\sum_{j=1}^{n}\left(\psi_\theta(X_j, Y_j, W_j) + \frac{\partial Q(\theta^*, \alpha)}{\partial \alpha'}\psi_\alpha(X_j, W_j)\right) + o_p(1), \end{aligned}$$

and by the finiteness of $E[\psi_\theta(X_j, Y_j, W_j)\psi'_\theta(X_j, Y_j, W_j)]$ and $E[\psi_\alpha(X_j, W_j) \times \psi'_\alpha(X_j, W_j)]$, the Cauchy–Schwarz inequality, Assumption 8, and the Lindeberg–Levy central limit theorem, this sum is asymptotically normal. By Equations (S.10), (S.11) and the Slutzky theorem, the conclusion of the theorem follows. *Q.E.D.*

S.4. COMPARISON WITH HAUSMAN, NEWEY, ICHIMURA, AND POWELL

In the polynomial case, the proposed estimator can be shown to have the same influence function as the IV estimator described in Hausman, Newey, Ichimura, and Powell (1991) simply by choosing suitable weighting functions, because both estimators rely on the same functional equations as a starting point (Equations (6) and (7) in the text). More specifically, in Section 3.2.2, the weighting functions must be selected such that

$$\begin{aligned} V_y(z, \theta) &= p(z)E[\mathbf{Z}\mathbf{Z}']^{-1}\mathbf{z}, \\ V_{xy}(z, \theta) &= p(z)E[\mathbf{Z}_+\mathbf{Z}'_+]^{-1}\mathbf{z}_+, \end{aligned}$$

where $\mathbf{Z} = [1, Z, \dots, Z^{\bar{k}}]'$, $\mathbf{z} = [1, z, \dots, z^{\bar{k}}]'$, $\mathbf{Z}_+ = [1, Z, \dots, Z^{\bar{k}+1}]'$, and $\mathbf{z}_+ = [1, z, \dots, z^{\bar{k}+1}]'$. (In that special case, $p(z)$ would not need to be estimated, because it would cancel with the division by $p(z)$ in the moment conditions.) The fact that there exists one choice of weighting functions that reach the same asymptotic variance as in Hausman, Newey, Ichimura, and Powell (1991) shows that the proposed estimator can be at least as efficient.

S.5. ROOT n CONSISTENT ESTIMATION UNDER MORE GENERAL CONDITIONS

The construction of the moment conditions in Section 3.2 uses, as a starting point, a set of smooth and rapidly decaying functions $\mathcal{G}$, described in Definition 1. This supplement shows how it is possible to define a larger set $\mathcal{G}'$ that enables root n consistent estimation in even more general settings, for instance, allowing for distributions of the disturbance U whose moment generating function exists only over a finite interval.

In that case, the enlarged set $\mathcal{G}'$ should also contain functions that can be written as linear combinations of products of polynomials, sines, cosines, and functions of the form $\sigma(a\zeta + b)$ for $a, b \in \mathbb{R}$ and

$$\sigma(\zeta) = \exp(-\cos^{-2}(\zeta\pi/2))\mathbb{1}(|\zeta| \leq 1). \tag{S.13}$$

The function $\sigma(\zeta)$ is compactly supported and infinitely many times differentiable (including at $|\zeta| = 1$). It is a refinement over the well known function $\exp(-(1 - \zeta^2)^{-1})\mathbb{1}(|\zeta| \leq 1)$ that improves the rate of decay of the inverse Fourier transform of $\sigma(\zeta)$ to $\exp(-c|z|)$ for some $c > 0$ instead of merely faster than $|z|^{-k}$ for any $k \in \mathbb{N}$, as shown in Lemma S.2 and Theorem S.1 at the end of this section.

The treatment in Section 3.2 carries over with this alternative set $\mathcal{G}'$ with one exception. The fact that $\sigma(\zeta)$ is compactly supported (and therefore that there exists compactly supported $\lambda(\zeta)$ in the set $\mathcal{G}'$) enables the use of Lemma 5 in the case where the moment generating function of U exists only on a finite interval. In that case, the Taylor series of the characteristic function $\phi(\zeta)$ of U converges only in a finite interval and it is crucial that a compactly supported $\lambda(\zeta)$ be used to "eliminate" the region where the Taylor series does not converge. Furthermore, the fact that the inverse Fourier transform of $\sigma(\zeta)$ decays rapidly is helpful to ensure that the functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$, and $r_{1y}(z, \theta)$ are rapidly decaying in z so that expectations of the form $E[Yr_y(Z, \theta)/p(Z)]$, for instance, exist. We can then state the following corollary to Theorem 2.

ASSUMPTION S.1: *The function $E[e^{tU}]$ exists for t in some neighborhood of the origin.*

COROLLARY S.1: *Under Assumptions* 1, 2(i), 4–6, *and* S.1, *if $Q(x, y, z, \theta, p)$ is as defined in Equation* (16) *and Section* 3.2 (*with $\mathcal{G}$ replaced by $\mathcal{G}'$*), *then there exists a compact set $\Theta \subset \mathbb{R}^{N_\theta}$ that contains θ^* in its interior such that $\theta = \theta^*$ is the only solution to $E[Q(X, Y, Z, \theta, p)] = 0$ in Θ. Assumption* S.1 *is unnecessary when $r_{y,o}(z, \theta)$, $r_{xy,o}(z, \theta)$, and $r_{1y,o}(z, \theta)$ are empty or when $r_{y,s}(z, \theta)$, $r_{xy,s}(z, \theta)$, $r_{1y,s}(z, \theta)$, and $r_{1y,o}(z, \theta)$ are empty.*

The set $\mathcal{G}$ can also be enlarged by including functions that have an $(\mathbf{i}\zeta)^{-1}$ prefactor. The treatment in Section 3.2 again carries over to this case, except that the proof of Lemma 5 needs to be adapted (because the absolute integrability assumption made in Lemma 5 does not automatically hold) by employing the following technique. The left-hand side of Equation (67) can be decomposed as

$$\int_{-\eta}^{\eta} \lambda(\zeta)\phi(0)\,d\zeta = \int_{-\eta}^{\eta}\left(\lambda(\zeta) - \frac{C}{\mathbf{i}\zeta}\right)\phi(0)\,d\zeta + \int_{-\eta}^{\eta}\frac{C}{\mathbf{i}\zeta}\phi(0)\,d\zeta,$$

where C is a constant such that $(\lambda(\zeta) - C/\mathbf{i}\zeta)$ is absolutely integrable and where $\int_{-\eta}^{\eta}(C\phi(0)/(\mathbf{i}\zeta))\,d\zeta = 0$ in the Cauchy principal value sense. Next, $\lambda(\zeta)$ can be replaced by $(\lambda(\zeta) - C/\mathbf{i}\zeta)$ in the right-hand side of Equation (67). The remainder of the proof is unchanged.

LEMMA S.2: *Let $\sigma(\zeta)$ be the Fourier transform of $s(z)$. For $\alpha \in \mathbb{R}^+$ and $\gamma \in \mathbb{N}$, if*

$$\sum_{t=0}^{\infty}\frac{\alpha^t}{t!}\int\left|\frac{d^{\gamma t}\sigma(\zeta)}{d\zeta^{\gamma t}}\right|d\zeta < \infty,$$

then, for some $C > 0$,

$$|s(z)| \le C\exp(-\alpha|z|^\gamma).$$

PROOF: Let $T(z) = \exp(\alpha z^\gamma)$. Because the radius of convergence of the Taylor series of the exponential function is infinite, we can also write $T(z) = \sum_{t=0}^{\infty}\alpha^t z^{\gamma t}/t!$ for all $z \in \mathbb{R}$. Let Θ denote the linear operator defined by

$$\Theta\sigma(\zeta) = \sum_{t=0}^{\infty}\frac{\alpha^t}{t!}\frac{(-\mathbf{i})^{\gamma t}d^{\gamma t}\sigma(\zeta)}{d\zeta^{\gamma t}}.$$

Because the Fourier transform of $z^t s(z)$ is $(-\mathbf{i})^t\,d^t\sigma(\zeta)/d\zeta^t$, the Fourier transform of $T(z)s(z)$ is $\Theta\sigma(\zeta)$. We can then write, for $z \ge 0$,

$$\begin{aligned}|s(z)| &= \frac{1}{|T(z)|}|T(z)s(z)| = \frac{1}{|T(z)|}\left|\int \Theta\sigma(\zeta)e^{-\mathbf{i}\zeta z}\,d\zeta\right| \\ &\le \frac{1}{|T(z)|}\int|\Theta\sigma(\zeta)|\,d\zeta \\ &= \frac{1}{|T(z)|}\int\left|\sum_{t=0}^{\infty}\frac{\alpha^t}{t!}\frac{(-\mathbf{i})^{\gamma t}d^{\gamma t}\sigma(\zeta)}{d\zeta^{\gamma t}}\right|d\zeta\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{|T(z)|}\sum_{t=0}^{\infty}\frac{\alpha^t}{t!}\int\left|\frac{d^{\gamma t}\sigma(\zeta)}{d\zeta^{\gamma t}}\right|d\zeta\\
&= \frac{C}{|T(z)|} = C\exp(-\alpha|z|^{\gamma})
\end{aligned}
$$

with

$$C = \sum_{t=0}^{\infty}\frac{\alpha^t}{t!}\int\left|\frac{d^{\gamma t}\sigma(\zeta)}{d\zeta^{\gamma t}}\right|d\zeta < \infty.$$

For $z < 0$, we can similarly write

$$
\begin{aligned}
|s(z)| &= \frac{1}{|T(-z)|}|T(-z)s(z)| = \frac{1}{|T(-z)|}\left|\int \Theta\sigma(\zeta)e^{\mathrm{i}\zeta z}\,d\zeta\right|\\
&\le \frac{1}{|T(-z)|}\int|\Theta\sigma(\zeta)|\,d\zeta \le \frac{C}{|T(|z|)|} = C\exp(-\alpha|z|^{\gamma}). \qquad \textit{Q.E.D.}
\end{aligned}
$$

THEOREM S.1: *The inverse Fourier transform* $s(\zeta)$ *of the function*

$$\sigma(\zeta) = \exp(-\cos^{-2}(\zeta))\mathbb{1}(|\zeta| \le \pi/2)$$

is such that $|s(z)| \le C\exp(-\alpha|z|)$ *for* $\alpha \in [0, 1/3[$ *and some positive* $C < \infty$.

PROOF: The proof consists of verifying that $\sigma(\zeta)$ satisfies the hypothesis of Lemma S.2. The tth derivative of $\exp(-\cos^{-2}(\zeta))$ consists of a sum of at most 3^t terms of the form

$$\text{(S.14)}\qquad C\exp(-\cos^{-2}(\zeta))\cos^{-p}(\zeta)\sin^{q}(\zeta),$$

where $q \ge 0$, $0 \le p \le 2t$, and $|C| \le 1 + t$. Because $p \le 2t$, $|\sin(\zeta)| \le 1$, and $X^t\exp(-X) \le t^t\exp(-t)$ for all $X \in \mathbb{R}^+$ and all $t \in \mathbb{N}$, we have

$$
\begin{aligned}
&\left|\exp(-\cos^{-2}(\zeta))\cos^{-p}(\zeta)\sin^{q}(\zeta)\right|\\
&\quad\le \exp(-\cos^{-2}(\zeta))\cos^{-2t}(\zeta)\\
&\quad\le t^t\exp(-t).
\end{aligned}
$$

Consequently, for some $C > 0$,

$$
\begin{aligned}
\sum_{t=0}^{\infty}\frac{\alpha^t}{t!}\int\left|\frac{d^t\sigma(\zeta)}{d\zeta^t}\right|d\zeta &\le C\sum_{t=0}^{\infty}\frac{\alpha^t}{t!}3^t(1+2t)t^t\exp(-t)\\
&\le C\sum_{t=0}^{\infty}\alpha^t(3+\varepsilon_1)^t\frac{t^t\exp(-t)}{t!} \qquad (\text{for any } \varepsilon_1 > 0)
\end{aligned}
$$

$$\leq C\sum_{t=0}^{\infty}((3+\varepsilon_2)\alpha)^t \qquad \text{(for any } \varepsilon_2 > 0\text{)},$$

which converges if $\alpha < 1/3$, choosing $\varepsilon_2 < 1/\alpha - 3$. *Q.E.D.*

S.6. DETAILS OF THE MONTE CARLO SIMULATIONS

We consider three different specifications, namely, a polynomial, a rational fraction and a probit model. In all cases, the mismeasured regressor X is generated from

$$X = X^* + \Delta X,$$
$$X^* = Z - U$$

with Z, U, and ΔX drawn from the distributions

$$Z \sim N(0,1), \quad U \sim N(0,1/4), \quad \Delta X \sim N(0,1/4).$$

Note that the ratio of the standard deviation of the measurement error ΔX to the standard deviation of the true regressor X^* is $(1/2)/\sqrt{(1+1/4)} \approx 0.45$, so that the measurement error is fairly large. In addition, the R^2 of the equation $X = Z - U + \Delta X$ is 2/3, indicating that the "strength" of the instrument is of a magnitude that is fairly typical for applications.

The dependent variable Y is generated from

$$Y = g(X^*, \theta) + \Delta Y,$$

where the functional form of $g(x^*, \theta)$ and the distribution of ΔY differ for each model.

For the kernel density estimation of the density of Z, an infinite order kernel is used, which has the desirable property that the estimation bias decays faster than any power of the bandwidth h as $h \to 0$. The specific kernel $K(z)$ used is the inverse Fourier transform of

$$\kappa(\zeta) = \left(\int_{-\infty}^{\infty} \sigma\left(\frac{\xi+2}{1.9}\right) d\xi\right)^{-1} \int_{-\infty}^{\zeta} \left(\sigma\left(\frac{\xi+2}{1.9}\right) - \sigma\left(\frac{\xi-2}{1.9}\right)\right) d\xi, \tag{S.15}$$

where $\sigma(\zeta)$ is given by $\sigma(\zeta) = \exp(-\cos^{-2}(\zeta\pi/2))\mathbb{1}(|\zeta| \leq 1)$. The prefactor ensures that $\kappa(0) = 1$ and therefore that $\int K(z)\,dz = 1$, as should be the case for a valid kernel. It is the fact that $\kappa(\zeta)$ is constant over $[-0.1, 0.1]$ that makes $K(z)$ an infinite order kernel. The function $\kappa(\zeta)$ inherits the smoothness of the function $\sigma(\zeta)$, thus ensuring that $K(z)$ is rapidly decaying.

The "optimal" bandwidth parameter h and trimming parameter τ are chosen so as to minimize the generalized method of moments (GMM) objective function associated with the proposed estimator evaluated at θ^*. In our simulation study, this is achieved by scanning values of h from 0.5 to 1.5 in multiplicative increments of 1.1 and values of τ from 0.005 to 0.05 in multiplicative increments of 1.5. The GMM objective function for the given level of smoothing and trimming is then evaluated for 50 replicated samples of 1,000 observations and averaged. The optimal bandwidth and trimming parameters are found to be $h = 0.585$ and $\tau = 0.026$. The optimal values obtained for all three models considered are the same, within the accuracy implied by the spacings between the consecutive values of h or τ scanned. This is perhaps not surprising because the distribution of Z to be nonparametrically estimated is common across all the models. Although the bandwidth and trimming parameters were optimized using knowledge of the experimental setup (i.e., the true value θ^*), the simulation results should not be overly optimistic. Semiparametric estimators tend to be less sensitive to the exact choice of the bandwidth than fully nonparametric estimators are. Also, keeping the trimming parameter fixed over all replications demands more aggressive trimming to ensure that all replications give reasonable estimates. Empirical researchers would typically fine-tune the trimming parameter for each given sample and could probably do better, on average, than the current simulations show.

The finite-sample properties of the proposed estimator (for the given values of h and τ) are studied by drawing 5,000 samples of 1,000 independent observations. As a point of comparison, we also calculate the standard instrumental variable estimator using $\partial g(Z,\theta)/\partial\theta$ as a vector of instruments and X as the regressor in addition to a standard (nonlinear) least squares estimator using X as the regressor, although both of these estimators are clearly biased in the presence of measurement error.

Let $\hat{\theta}_k$ denote any element of $\hat{\theta}$, the parameter vector estimated by any one of the three estimators, and let θ_k^* denote any element of θ^*, the true value of the parameter vector. The three estimators are compared on the basis of their bias, standard deviation, root mean squared error (RMSE), and overall root mean squared error, given, respectively, by

$$\begin{aligned}
&\text{bias} = E[\hat{\theta}_k] - \theta_k^*,\\
&\text{std. dev.} = \big(E\big[(\hat{\theta}_k - E[\hat{\theta}_k])^2\big]\big)^{1/2},\\
&\text{RMSE} = \big(E[(\hat{\theta}_k - \theta_k^*)^2]\big)^{1/2},\\
&\text{RMSE}_{\text{all}} = \big(\operatorname{tr} E[(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)']\big)^{1/2}.
\end{aligned}$$

Note that the last quantity is a convenient summary measure of the overall performance of an estimator.

Although our estimator is based on moment conditions that have zero expectation at the true value of the parameter vector, it is perfectly normal that it could be biased in a finite sample. First, the moment conditions used for estimation are nonlinear in θ and it is well known that, in this context, just identified GMM exhibits a bias of order n^{-1}, where n is sample size (see, for instance, Newey and Smith (2004)). Second, the implementation of the estimator relies on kernel smoothing and trimming, two techniques that introduce their own bias. Simulations prove to be a helpful tool to verify that the potential presence of such biases does not overcome the benefits of the elimination of the measurement error-induced bias. We now describe the specifics of each simulation.

S.6.1. *Polynomial Model*

This model is defined by

$$g(x^*, \theta) = \theta_1 + \theta_2 x^* + \theta_3 (x^*)^2 + \theta_4 (x^*)^3,$$
$$\Delta Y \sim N(0, 1/4),$$

where $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 0$, and $\theta_4 = -0.5$. The Fourier transform of this polynomial contains no ordinary function component, but only a linear combination of delta function derivatives, and therefore the weighting functions $\omega(\zeta)$ and $\varpi(\zeta)$ do not need to be introduced. Following the discussion in Section 3.2.2, the weighting functions $\nu_{y,j}(\zeta, \theta)$ and $\nu_{xy,j}(\zeta, \theta)$ for $j = 0, \dots, \bar{k}$ (where $\bar{k} = 3$) are chosen to be of the form[7]

$$\nu_{y,j}(\zeta, \theta) = (\mathbf{i}\zeta)^j \exp\left(-\frac{1}{2}\left(\frac{\zeta}{(1.1)\pi/2}\right)^2\right), \tag{S.16}$$

$$\nu_{xy,j}(\zeta, \theta) = (\mathbf{i}\zeta)^j \exp\left(-\frac{1}{2}\left(\frac{\zeta}{(1.1)\pi/2}\right)^2\right). \tag{S.17}$$

Table I compares the performance of the proposed estimator relative to instrumental variables (IV) and ordinary least squares (OLS). Although the bias of the proposed estimator is slightly larger that of IV for three of the coefficients (θ_1, θ_3, and θ_4), the bias of IV for the remaining coefficient (θ_2) is overwhelmingly large, making the overall performance of IV poor. This is best illustrated by substituting the expected values[8] of the coefficients obtained from each estimator into the polynomial specification and by overlapping the graph of each resulting polynomial over the "true" model specification. As seen in Figure S.1(a), the proposed estimator is much closer to the true specification

[7] Because the ordinary part of the Fourier transform of $g(x^*, \theta)$ is zero ($\gamma_o(\zeta, \theta) = 0$), there is no need to ensure that the $\nu_{y,j}(\zeta, \theta)$ and $\nu_{xy,j}(\zeta, \theta)$ are orthogonal to the ordinary part. Hence we can specify $\nu_{y,j}(\zeta, \theta)$ and $\nu_{xy,j}(\zeta, \theta)$ directly without first introducing the functions $\mu_{y,j}, \mu_{xy,j} \in \mathcal{S}_0 \cap \mathcal{C}$, as done in Section 3.2.2.

[8] That is, their average over the replications.

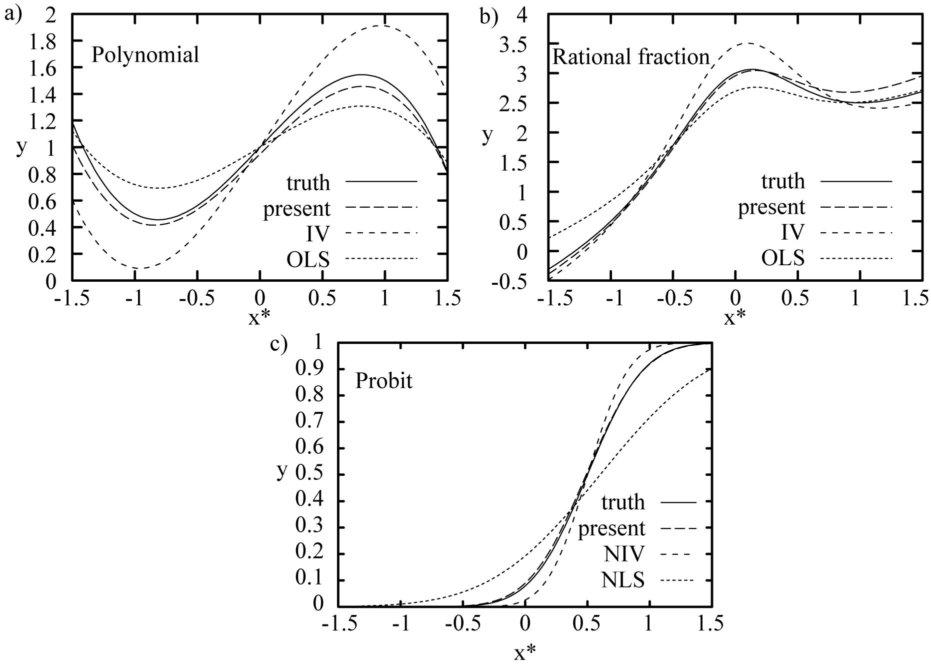

FIGURE S.1.—Graphical representation of the bias of each estimator studied. Note that for the probit model in (c), the curve for the standard nonlinear instrumental variable (NIV) estimator excludes the 50% of the replications that do not yield a finite estimate of θ_2. The actual performance of NIV is therefore far worse than indicated by the graph.

than any of the other estimators. Although the reduction in bias achieved with our estimator comes at the expense of increased standard errors for some coefficients, the overall RMSE (the column labeled “All” in Table I) is still lower for the proposed estimator than for the other two estimators.

S.6.2. *Rational Fraction*

The second example is a specification of the form

$$g(x^*, \theta) = \theta_1 + \theta_2 x^* + \frac{\theta_3}{(1+(x^*)^2)^2},$$
$$\Delta Y \sim N(0, 1/4),$$

where $\theta_1 = 1$, $\theta_2 = 1$, and $\theta_3 = 2$. The Fourier transform of $g(x^*, \theta)$ in this case contains both an ordinary and a singular component:

$$\gamma(\zeta, \theta) = \theta_1 2\pi\delta(\zeta) - \theta_2 2\pi\mathbf{i}\delta^{(1)}(\zeta) + \theta_3 \frac{\pi}{2}(1+|\zeta|)e^{-|\zeta|}. \tag{S.18}$$

As discussed in Section 3.2.2, to determine the singular component, we need to construct some functions $\mu_{y,j}$, $\mu_{xy,j} \in \mathcal{S}_0 \cap \mathcal{C}$. In the case of $\mu_{y,j}$, this is accomplished using Definition 2 with

$$\text{(S.19)} \qquad \lambda(\zeta) = (\mathbf{i}\zeta)^j \exp\left(-\frac{1}{2}\left(\frac{\zeta}{(2.1)\pi/2}\right)^2\right) - 2(\mathbf{i}2\zeta)^j \exp\left(-\frac{1}{2}\left(\frac{2\zeta}{(2.1)\pi/2}\right)^2\right)$$

for $j = 1, \dots, 2$. The function $\mu_{xy,j}$ is obtained similarly, with $j = 1, \dots, 3$.

The ordinary part in Equation (S.18) depends on a single parameter and, consequently, only the scale of the ordinary part needs to be determined. As explained in Section 3.2.1, the vector of weighting function ω associated with the "shape" of the regression function is therefore not needed; only the weighting function ϖ, associated with the "scale" is. Definition 2 is then used to obtain $\varpi \in \mathcal{S}_1 \cap \mathcal{C}$ with

$$\lambda(\zeta) = (\mathbf{i}\zeta)^2 \exp\left(-\frac{1}{2}\left(\frac{\zeta}{(1.6)\pi/2}\right)^2\right) \times \left(\int (\mathbf{i}\xi)^2 \exp\left(-\frac{1}{2}\left(\frac{\xi}{(1.6)\pi/2}\right)^2\right) d\xi\right)^{-1}.$$

The prefactor $(\mathbf{i}\zeta)^2$ ensures that the singular parts do not affect the estimation of the ordinary part.

Table II summarizes the results of the simulations for the rational fraction model and clearly illustrates the bias-correcting power of the proposed estimator. Although the IV estimator exhibits a fortuitously low bias on the θ_2 parameter, it clearly fails to produce unbiased estimates of the coefficient on the nonlinear term (θ_3). As is seen in Figure S.1(b), the proposed estimator provides a nearly unbiased estimate of the height of the nonlinear component of the specification, unlike IV, which overestimates it, and OLS, which underestimates it. The proposed estimator has, overall, a bias of only about 10% for this model. Because our estimator typically exhibits larger standard error than both IV and OLS, it is instructive to verify whether it still comes out ahead when both bias and variance are taken into account. Indeed, the overall RMSE clearly points toward the proposed estimator as the best alternative.

S.6.3. *Probit*

The probit model can be written as a regression model with the specification

$$\text{(S.20)} \qquad g(x^*, \theta) = \frac{1}{2}(1 + \operatorname{erf}(\theta_1 + \theta_2 x^*)),$$

where we set $\theta_1 = -1$ and $\theta_2 = 2$ and where the distribution of ΔY conditional on $X^* = x^*$ is given by

$$\Delta Y = \begin{cases} 1 - g(x^*, \theta) & \text{with probability } g(x^*, \theta), \\ -g(x^*, \theta) & \text{with probability } 1 - g(x^*, \theta). \end{cases}$$

The Fourier transform of $g(x^*, \theta)$ given in Equation (S.20) is

$$\text{(S.21)} \qquad \gamma(\zeta, \theta) = \pi\delta(\zeta) - \frac{1}{\mathbf{i}\zeta}\exp\left(-\mathbf{i}\zeta\frac{\theta_1}{\theta_2} - \frac{\zeta^2}{4\theta_2^2}\right).$$

Because the singular component of Equation (S.21) does not depend on θ, it provides no information to estimate the model and we therefore need to consider only the ordinary part. In addition, the scale of the regression function is entirely determined by the constraint that it must tend to 1 as $x^* \to \infty$ and tend to 0 as $x^* \to -\infty$ (for $\theta_2 > 0$), so there is no need to estimate the scale. As a result, probit falls into the class of models where the only weighting function needed is $\omega(\zeta)$. As prescribed in Section 3.2.1, the two elements of $\omega(\zeta)$ are chosen to be

$$\text{(S.22)} \qquad \omega_j(\zeta) = (\mathbf{i}\zeta)^{j+2}\exp\left(-\frac{1}{2}\left(\frac{\zeta}{(1.5)\pi/2}\right)^2\right)e^{\mathbf{i}\zeta/2}$$

for $j = 1, 2$. Note that the prefactor $(\mathbf{i}\zeta)^{j+2}$ in Equation (S.22) is chosen to ensure that $\gamma_o(\zeta, \theta)\omega(\zeta)$ and $\dot{\gamma}_o(\zeta, \theta)\omega(\zeta)$ are well behaved. Indeed, the ordinary part $\gamma_o(\zeta, \theta)$ behaves like ζ^{-1} as $\zeta \to 0$ (and thus $\dot{\gamma}_o(\zeta, \theta)$ behaves like ζ^{-2}) and the foregoing choice of $\omega(\zeta)$ guarantees that its product with $\gamma_o(\zeta, \theta)$ or $\dot{\gamma}_o(\zeta, \theta)$ is bounded. Finally, the factor $e^{\mathbf{i}\zeta/2}$ simply introduces a shift in $r_y(z, \theta)$ and $r_{xy}(z, \theta)$ so that their respective modes fall within the regions where $E[Y|Z = z]$ and $E[XY|Z = z]$ vary the most rapidly.

The results shown in Table III and the graph of Figure S.1(c) clearly indicate that the proposed estimator is nearly unbiased, unlike nonlinear instrumental variables (NIV) and nonlinear least squares (NLS). Once again, despite its relatively large standard errors, our estimator still outperforms both NIV and NLS in terms of overall RMSE (see last column). It should also be noted that, for the probit model, the NIV estimator using $\partial g(z, \theta)/\partial\theta$ as instruments exhibits the undesirable tendency to give a $\hat{\theta}_2$ that diverges to infinity about 50% of the time. The results for the NIV estimator given in Table III and Figure S.1(c) are averages over only the replications that did converge to a finite value. The actual performance of NIV is therefore far worse than reported in the table and in the figure.

S.7. APPLICATION

S.7.1. *Introduction*

The estimation of the wage differential between workers of a different race offers an opportunity to assess the presence of discrimination in the labor market and has received considerable attention in the economics literature (Neal and Johnson (1996), Bollinger (2003), Carneiro, Heckman, and Masterov (2003), Card and Lemieux (1994), and many others). A crucial aspect of this estimation problem is the necessity to control for other factors that affect income so as to separate actual labor market discrimination from premarket factors, such as family socioeconomic background or schooling quality. Following Neal and Johnson (1996), we use the score on a standardized test taken by virtually all respondents prior to job market entry as an explanatory variable that controls for all premarket factors. Neal and Johnson argue that such an approach offers the advantage that the control variable does not suffer from endogeneity, because it is not affected by the respondent's own decisions, unlike other frequently used controls such as years of schooling, occupation, marital status, or geographical location. Neal and Johnson's findings indicate that the apparent black–white male wage gap of 24% is reduced to only 7% when premarket skills are taken into account.

Although Neal and Johnson's argument supports the assumption of the exogeneity of skills, it does not rule out that skills may be measured with error, thus making OLS estimates potentially inconsistent. This issue was investigated by Bollinger (2003), who provided bounds on the wage gap that account for measurement error. Although his widely applicable bounding technique does not depend on the availability of instruments, it is only able to estimate consistently an interval that contains the true wage gap, instead of providing a consistent point estimate. As a result, the method gives rather wide bounds on the black–white wage gap (with values ranging from 7% to −126%). Surprisingly, this interval mostly contains negative values of the black–white wage gap, apparently suggesting that discrimination is more likely to be against whites. Also, while Neal and Johnson's study allows for a nonlinear relationship between skill and wages, Bollinger's analysis focuses on a linear specification, because bounding techniques are not available for nonlinear specifications.[9]

Our proposed estimation strategy offers the opportunity to combine the strengths of both studies, allowing for the presence of both nonlinearity and measurement error. Moreover, our instrumental variable approach makes it possible to obtain consistent point estimates and consequently improves the accuracy of the estimated black–white wage gap relative to a bounding approach. Note that this investigation is mainly intended to briefly describe a

[9] Unless bounds on the magnitude of the measurement error are available (see, e.g., Stoker, Berndt, Ellerman, and Schennach (2005)).

relevant example of an application of our proposed estimator and is necessarily less detailed than a thorough study that focuses exclusively on the wage gap issue.

S.7.2. *Data and Methodology*

The data we use originate from the "young men" survey group obtained from the National Longitudinal Survey (NatLS). Our analysis focuses on a subsample that contains all individuals for whom appropriate measures of income, skills, and parental characteristics are available. This subsample consists of 2,133 white and 333 black respondents.[10] This data set is different from that used in the studies of Neal and Johnson (1996) and Bollinger (2003), because we found that the measure of ability used in these studies, the Armed Forces Qualification Test (AFQT), has one main limitation. The AFQT score suffers from a significant censoring bias (see Figure S.2), which causes two problems. First, this score is less able to distinguish relative abilities among highly skilled people and, second, censoring introduces a measurement error that is negatively correlated with true ability, thus violating our conditional mean assumption regarding the measurement error.[11] We rely instead on Intellectual Quotient (IQ), as reported in the NatLS study.[12] Although IQ is, technically, also a bounded quantity, the probability density of IQ quickly decays away from its mean, so that the upper and lower bounds on IQ are never reached at the sample sizes available in this study (see Figure S.3). As a result, the distribution of IQ is virtually indistinguishable from a continuous distribution supported on $\mathbb{R}$, so that the assumption of classical measurement error is plausible.

Our analysis also requires a measure of permanent income, which we calculate from the NatLS's record of the respondents' yearly wage income over the whole the period of the study. We average log wage income over all years for which the respondent was at least 25 years old.[13] Admittedly, this is far from a

[10] The 53 respondents who did not belong to either racial groups were omitted.

[11] The same caveat applies to the work of Bollinger (2003). Note that, so far, it is not known if it is possible to correct for this type of nonclassical measurement error in the absence of validation data.

[12] The IQ reported in the NatLS study actually comes from a variety of different types of IQ tests taken while the respondent was attending high school. Even if each type of IQ test had a different systematic bias, this would not invalidate our analysis, because the heterogeneity in the IQ tests can simply be considered as a form of measurement error, provided it satisfies the appropriate conditional mean restriction. To investigate this potential problem, we have compared summary statistics for IQ score and income within each subgroup that shares the same IQ test type. We repeated our analysis after excluding each of the groups that exhibit the largest deviations from the overall sample average and obtained similar results, which also supports our conclusions. The omitted groups were those for which the IQ scores were calculated from the Scholastic Aptitude Test (SAT) or the grade point average (GPA).

[13] All amounts were previously deflated using the consumer price index. All incomes reported to be below $3,000 (in 1984 dollars) were also excluded from the average, because those values are probably the result of temporary loss of work or return to school.

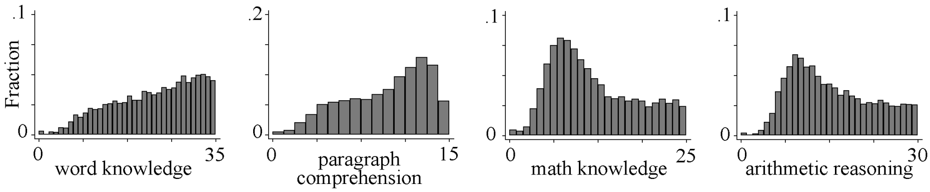

FIGURE S.2.—Distribution of scores for each portion of the AFQT test. The thickness of the upper tail of the distributions is indicative that many respondents may have "true" abilities that exceed the maximum possible test score. The absence of a spike at the highest score value can be explained by the fact that even very able respondents can make random mistakes.

perfect measure of log permanent income. However, our setup trivially allows for log permanent income to be error-contaminated, because it is the dependent variable. This error is heteroscedastic, given that income is available for a different number of years, depending on the respondent, but our setup allows for that possibility as well. Our measure of permanent income may also be biased without invalidating our analysis, as long as the bias does not depend on race or IQ. Although it is common to control for age in wage gap analyses, we favor an approach that avoids an explicit modeling of life cycle effects. We simply consider the effect of age as a random disturbance, because it is reasonable to assume that the age of the respondent at the beginning of the study is independent of race and IQ measured at a fixed age.

Our model is defined by

$$Y_i = \theta_1 + \frac{\theta_2}{2}\operatorname{erf}\left(\frac{X_i^* - \theta_4}{\theta_3}\right) + \Delta Y_i, \tag{S.23}$$

where Y_i is individual's i log annualized permanent income, X_i^* is his true intellectual quotient, and the disturbance ΔY_i is assumed to satisfy the assumptions

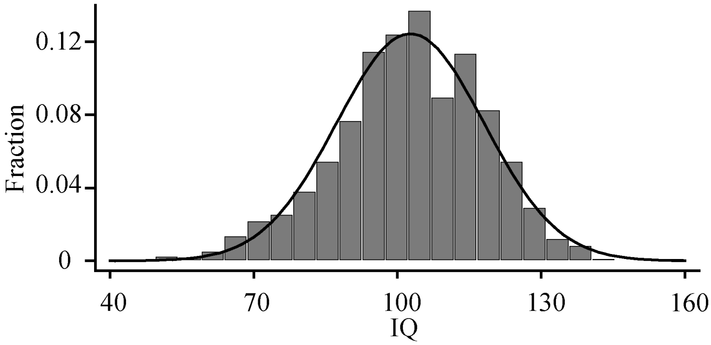

FIGURE S.3.—Distribution of IQ.

of model (2). A separate model is used for each racial group to allow for a completely general coupling between race and skill response. The parameters of this S-shaped specification have the following meaning. The parameter θ_1 is the overall level of log income, θ_2 is the income change from the low end to the high end of the IQ range, θ_3 determines the width of the "S", and θ_4 indicates the IQ level where the income varies most rapidly. The choice of specification is guided by a preliminary analysis based on a nonparametric regression of Y_i on X_i, which revealed that the response of income as a function of measured IQ saturates at low and high IQ. Note that the chosen specification reduces to a simple linear specification as $\theta_3 \to \infty$ and $\theta_2/\theta_3 \to c \in \mathbb{R}$. Hence, nonlinearity is not imposed in our model—if the response were actually linear, this would be reflected by a very large estimated value of θ_3.

Because IQ is an error-contaminated measure of true ability, we model observed IQ as $X_i = X_i^* + \Delta X_i$, where the measurement error ΔX_i is assumed to fulfill the requirements of model (2). Our vector of instruments, W_i, is constructed from (i) the respondent's mother's highest completed grade[14] and (ii) its square, (iii) the number of siblings the respondent has, (iv) a measure of availability of reading material during the respondent's childhood, and (v) the respondent's race.[15] This selection of instruments[16] is guided by the predictors of skills identified by Neal and Johnson (1996).

A Gaussian kernel[17] was used with a bandwidth (as measured by standard deviation) of 1.8 IQ points for whites and 2.5 IQ points for blacks. Trimming was activated whenever the density of predicted IQ fell below 0.002 $(\mathrm{IQ})^{-1}$ for whites and 0.006 $(\mathrm{IQ})^{-1}$ for blacks. These settings were determined by gradually varying the bandwidth and trimming parameters in search for the values where the point estimates were the least sensitive to changes in bandwidth and trimming parameters. Our preferred bandwidth and trimming parameters for the two racial groups differ because of their different sample sizes. However, it was verified that our results are robust to setting parameters for the white subsample equal to those of the black subsample. The weighting functions used are given in Section S.7.5. The same weighting functions were used for the two racial groups.

[14] For 73 respondents out of 2,466, the mother's highest completed grade was not available and the father's highest completed grade was used instead.

[15] The instrumental equation $X_i = W_i'\alpha + \Delta X_i^* + \Delta X_i$ is estimated jointly for both racial groups.

[16] Although each of these instruments is, strictly speaking, discretely distributed, a linear combination of them exhibits a distribution whose support consists of such a large number of points that it is virtually indistinguishable from a continuously distributed variable.

[17] Bias-reducing kernels were also tried, but were found to require substantial trimming to avoid the "vanishing denominator" problem. A positive second-order kernel was found to provide results that were less sensitive to bandwidth and trimming parameter selection.

S.7.3. *Diagnostic Tests*

Before considering the issue of the wage gap, we perform a few diagnostic tests to verify our assumptions, assess the presence of measurement error, and verify that the estimator performs as intended.

Although most of our assumptions take the form of relatively weak conditional mean restrictions, our assumption of independence between the predicted value$Z \equiv E[X|W]$ and the prediction error U in the instrument equation of model (5) warrants verification. Unfortunately, U is not directly observable (because X^* is not observed), so our test of instrument validity will instead rely on the more stringent constraint that $(\Delta X - U)$ is independent from Z. This test can be considered more stringent because, even if it failed, it could be the result of a dependence between ΔX and Z, which would not violate the assumptions of our estimation procedure.[18] This test is also feasible because $(\Delta X - U)$ can be obtained from the residuals of the regression of X on the instrument vector W. We rely on a Spearman rank correlation test for independence (see, for instance, van deer Vaar (1998, Example 13.22)). The test statistic is simply the sample correlation (scaled by $\sqrt{n}$) between the respective ranks[19] of the two variables of interest (here $(\Delta X - U)^2$ and Z). We use $(\Delta X - U)^2$ instead of simply $(\Delta X - U)$ to improve the power of the test, because the conditional mean restriction on $(\Delta X - U)$ would tend to make the rank correlation between $(\Delta X - U)$ and Z small by construction. (Other powers of $(\Delta X - U)$ and of the instruments yield similar conclusions.) To account for the presence of preliminary estimated parameters in the calculations of both $(\Delta X - U)$ and Z, we rely on 1,000 bootstrap replications to calibrate the asymptotic variance of this asymptotically normal test statistic. Our conclusion is that, in our application, the null hypothesis of independence is not rejected, because the rank correlation test statistic is 0.63, corresponding to a p-value of 0.53.

We next turn to the issue of testing for the actual presence of measurement error in our data. Table S.I reports the point estimates obtained with our method (labeled “Fourier”) and with conventional nonlinear least squares (NLS) estimates. A consistent test for the presence of measurement error can be constructed by verifying the statistical significance of the difference $(\hat{\theta}_F - \hat{\theta}_{LS})$, where $\hat{\theta}_F$ and $\hat{\theta}_{LS}$ denote the 8×1 vectors of all coefficients for the Fourier-based estimator and for nonlinear least squares, respectively. We employ the test statistic

$$\chi^2_{\mathrm{rank}(S)} = n(\hat{\theta}_F - \hat{\theta}_{LS})' S' \bigl(S' \hat{E}[(\psi_F - \psi_{LS})(\psi_F - \psi_{LS})'] S\bigr)^{-1} \times S(\hat{\theta}_F - \hat{\theta}_{LS}), \tag{S.24}$$

[18] Of course, a dependence between U and ΔX could fortuitously yield a $(\Delta X - U)$ that is independent of Z although U is dependent on Z, but this appears highly unlikely.

[19] The rank of a variable is its position in the sample when the sample is sorted according to the value of that variable.

TABLE S.I

POINT ESTIMATES AND HETEROSCEDASTICITY-ROBUST STANDARD ERRORS (IN PARENTHESES)

	Fourier (θ_F)				NLS (θ_{LS})			
	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
White	9.97 (0.04)	0.38 (0.14)	2.60 (3.58)	105.1 (2.9)	9.88 (0.08)	0.41 (0.18)	21.18 (10.45)	86.4 (8.2)
Black	9.74 (0.05)	0.59 (0.09)	4.27 (2.26)	102.8 (2.4)	9.78 (0.04)	0.38 (0.09)	7.12 (5.61)	98.3 (2.6)

where ψ_F and ψ_{LS} denote the influence functions of the corresponding estimator, the $\hat{E}$ operator denotes a sample average operation, and S is a rectangular selection matrix that extracts the degrees of freedom we wish to test. This type of test statistic reduces to a Hausman test if nonlinear least squares happens to be efficient and has a covariance matrix estimate that is positive definite by construction. As reported in Table S.II, our tests clearly reject the null hypothesis of the absence of measurement error for both racial groups.

We now verify that the estimator is effective at capturing the essential features of the data. Figure S.4 graphs the returns to IQ implied by specification (S.23) and our point estimates, both for the Fourier-based and the NLS estimators. Also shown in Figure S.4 are the isodensity contours of a nonparametric estimate of the joint density of Y_i and X_i. Our analysis centers on white respondents only, because it is the only subsample that is large enough to obtain a reliable nonparametric bivariate density estimate. The fact that our estimator closely follows the noticeable ridge in the joint density of Y_i and X_i is strongly indicative that the estimator properly identifies the presence of errors in both variables. Its ability to resolve the very sharp marginal returns to IQ in the region of highest density is especially striking. In contrast, the least squares estimator simply tracks the conditional mean of Y_i given X_i and does not detect the sharp increase in income that is clearly noticeable in the nonparametric density plot. The presence of a region with very sharp marginal returns to IQ has a very plausible explanation. The marginal density of IQ is largest in this region and it follows that a small change in IQ there leads to the

TABLE S.II

TESTING FOR THE PRESENCE OF MEASUREMENT ERROR

Null Hypothesis	Test Statistic	Degrees of Freedom	p-Value
No measurement error (white subsample)	15.4	4	0.0039
No measurement error (black subsample)	15.0	4	0.0047
No measurement error (whole sample)	29.9	8	0.0002

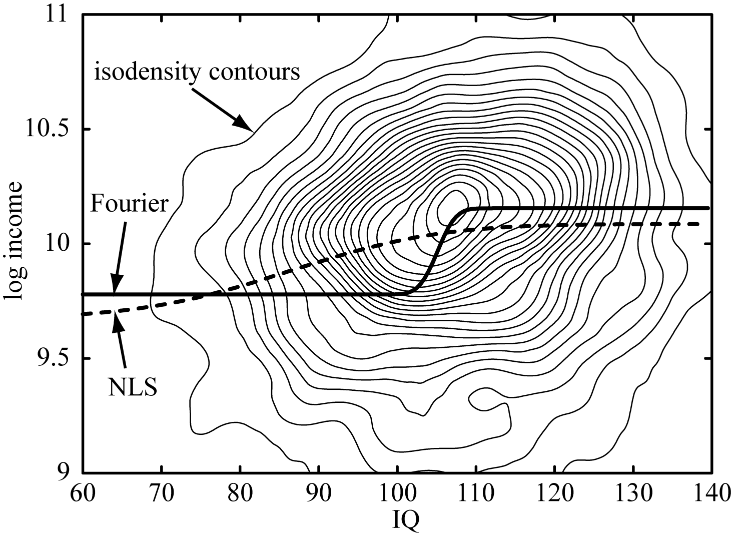

FIGURE S.4.—Comparison between IQ response curve obtained with the Fourier-based IV estimator (Fourier) and nonlinear least squares (NLS). Also shown are the isodensity contours of a nonparametric estimate of the joint density of Y_i and X_i.

largest changes in ranking in the overall population. Assuming that job market outcomes mostly depend on an individual's ranking, a large change in average income would then be expected.

It is interesting to note that, based solely on a conventional least squares analysis, a linear specification would have appeared to be adequate because the width of the "S" curve obtained with NLS is so large. This application thus provides a clear example where measurement error actually masks the extent of the nonlinearity of the specification and only a nonlinear approach that is robust to measurement error can reliably detect this situation.

The "S" shape of the response also has an unintended advantage in terms of the robustness of our analysis. It has been argued (Neal and Johnson (1996)) that measures of skills, such as IQ, may be a racially biased. However, the only effect of such a bias would be to shift the response horizontally (i.e., bias the θ_4 parameter). Hence, as will become evident in the next section, our estimates of the wage gap would be essentially robust to such biases over the relatively wide range of IQ where the income response is flat.

S.7.4. *Results*

We now return to the determination of the black–white male wage gap. Because we have allowed the response to IQ to differ between the two racial groups, we are able to determine the wage gap as a function of measurement error-free IQ (see Figure S.5), which provides new insight into the issue. The

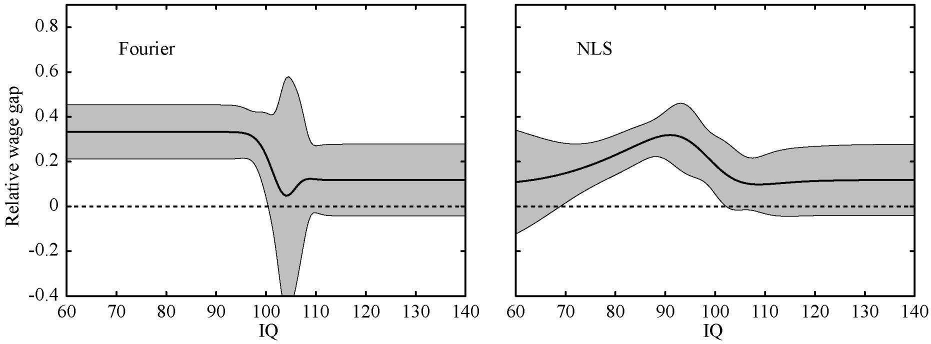

FIGURE S.5.—Estimated black–white wage gap as a function of IQ, estimated with the proposed Fourier-based estimator and with NLS. The plotted 95% confidence bands were determined with the delta method from the estimated covariance matrix of the coefficients.

confidence bands in Figure S.5 were obtained via the delta method using the estimated covariance matrix of each estimator. The relatively wide confidence bands are mainly attributable to the relatively small size of the black subsample.[20]

Although we have already established that the Fourier-based and NLS results are statistically significantly different, Figure S.5 illustrates that the results of the two procedures are also qualitatively very different. Our main findings, based on the measurement error-robust Fourier-based estimates, are twofold:

1. Below an IQ of about 100, the wage gap is on the order of 33% and is statistically significant at the 95% level.
2. Above 110 of IQ, the gap shrinks to about 12%, a value which is not statistically significantly different from 0. However, the fact that the wage gap decreases is statistically significant: The χ^2_1 statistic that tests that the wage gap is the same below an IQ of 100 and above an IQ of 110 is equal to 5.27, which rejects the null at the 95% level.

It is instructive to compare our findings with those of Neal and Johnson (1996) and Bollinger (2003). First, it should be noted that differences between our results and these earlier studies can at least in part be traced back to differences in the data used. Repeating Neal and Johnson's main least squares analysis (using a quadratic dependence on skills and a dummy for race) with our sample yields a wage gap of 33% without controlling for skills, which is reduced to 21% after controlling for skills via IQ (Neal and Johnson found 24% and 7%, respectively). Hence, it should not be surprising that our results more

[20] The spikes in the confidence bands around the elbow of the curve are due to the large IQ dependence of income in this region, which magnifies the noise in the estimated θ_4 coefficient. A similar feature is not clearly visible in the NLS estimates because the estimated IQ dependence happens to be far weaker for NLS. Fortunately, the very large standard errors in the Fourier-based approach only affect a small portion of the curve and will thus not affect our main findings.

strongly indicate the presence of discrimination. One possible source of the difference is that Neal and Johnson used hourly wages, while we use yearly income.[21] If there is discrimination in the hiring process, black respondents may remain unemployed for longer periods, an effect that would be visible in the reported yearly income but not in the reported hourly wage. Of course, sample selection bias issues may still play a role in our study (Chandra (2003)).

Our results confirm Neal and Johnson's observation that when the skill dependence of income is allowed to differ across racial groups, the gap appears to narrow at the higher end of the skill distribution. This trend was not statistically significant in their study, but is clear in Figure S.5. The well-known measurement error-induced attenuation phenomenon is a possible source of the lack of significance Neal and Johnson observed, although differences in the data used could also be a factor. Figure S.5 also shows that it would be inappropriate to use Bollinger's bounds on the wage gap to conclude that the wage gap is inexistent or negative. The relatively narrow width of our confidence bands enabled by the use of instruments permits us to pin down the magnitude of the wage gap more precisely and to show that it is still statistically significant at least over a portion of the skill distribution when measurement error is taken into account. Perhaps our most striking finding is the sharpness of the drop in the wage gap as a function of IQ, a feature that simply cannot be detected in our data set without properly accounting for both measurement error and nonlinearity.

The results of our analysis are consistent with a number of interpretations. For instance, applicants for jobs that requiring low skill levels are typically recruited locally, while more skill-intensive positions are often advertised over a larger geographical area, through newspapers, specialized magazines, or recruiting services. Hence, the low-skill wage gap mainly reflects a gap in the prevailing wages in different, segregated, neighborhoods. The gap is smaller among highly skilled individuals, who do not necessarily work in their native neighborhood. An alternative, but related, explanation is that, beyond a certain level of ability, undertaking a college education becomes more likely, which often brings young black men out of their native neighborhoods and into other communities where the prevailing wages may be higher. Finally, it is possible that, beyond discrimination in wages, there exists discrimination in the hiring/firing process, which would cause black workers to be employed for a smaller fraction of the year on average than equally qualified white workers, thus resulting in a black–white gap in yearly income. If the turnover rate is higher in occupations that demand lower skills, this would result in a larger income gap between racial groups for lower skilled workers.

[21] Our use of yearly income was guided by the fact that hourly wages (calculated or reported) were missing for a large fraction of the respondents in our sample.

S.7.5. *Weighting Functions Used in the Application*

We first observe that the Fourier transform of the S-shaped specification given in Equation (S.23) is

$$\gamma(\zeta,\theta)=\theta_1 2\pi\delta(\zeta)+\theta_2(-\mathbf{i}\zeta)^{-1}\exp\left(\mathbf{i}\zeta\theta_4-\frac{\zeta^2}{4\theta_3^2}\right). \tag{S.25}$$

Consequently, a weighting function of the form $\nu_{y,0}(\zeta,\theta)$ is needed to extract the magnitude of the singularity θ_1, a weighting function of the form $\varpi(\zeta)$ is required to determine θ_2 and a two-dimensional vector of weighting functions of the form $\omega(\zeta)$ is needed to obtain θ_3 and θ_4. (Refer to Section 3.2 for a description of these different types of weighting functions.) The functional forms of these weighting functions, given in Table S.III, were derived as follows.

The starting point of the construction of $\omega_j(\zeta)$ (in Section 3.2.1) is a Gaussian function (of z) with a center of mass and a width such that the Gaussian takes a negligible value outside of the range of values of Z actually observed in the sample. After a Fourier transform operation, this yields another Gaussian function (of ζ) multiplied by a phase factor $e^{\mathbf{i}\zeta c}$, where c depends on the center of mass of the original Gaussian. Each element of the vector $\omega(\zeta)$ is obtained from a Gaussian with a slightly different center of mass. Next, this expression is multiplied by a positive power of $(\mathbf{i}\zeta)$ that is (i) sufficiently large to cancel the $(-\mathbf{i}\zeta)^{-1}$ divergence in Equation (S.25) or the $(-\mathbf{i}\zeta)^{-2}$ divergence in its derivative $\dot\gamma(\zeta,\theta)$ and (ii) such that the inner products of $\omega_j(\zeta)\dot\gamma(\zeta,\theta^*)$ with $\delta(\zeta)$ and $\omega_j(\zeta)\gamma(\zeta,\theta^*)$ with $\delta^{(1)}(\zeta)$ vanish, thus achieving orthogonality to the singular part.

As described in Section 3.2, the weighting function $\varpi\in\mathcal{S}_1\cap\mathcal{C}$ is derived from some function $\lambda\in\mathcal{G}'$ (which is the extensions of $\mathcal{G}$ provided in Section S.5). The functional form of $\lambda(\zeta)$ (given in Table S.III) is obtained by first noting that the weighting function used to identify the height of the S-shaped function (the θ_2 parameter) should essentially sample the difference between the value of $E[Y|Z=z]$ for values of z before and after the "jump." Hence, a natural starting point is the difference between two Gaussian functions (of z)

TABLE S.III

WEIGHTING FUNCTIONS USED FOR THE FOURIER-BASED ESTIMATOR

Function[a]	Expression
$\omega_j(\zeta)$	$\omega_j(\zeta)=(\mathbf{i}\zeta)^3\exp(-(1.402)\zeta^2)e^{-101\mathbf{i}\zeta}e^{-\mathbf{i}\zeta(2.75)(j-1)}$ for $j=1,2$
$\varpi(\zeta)$	$\lambda(\zeta)=C\frac{\mathbf{i}\sin(27.5\zeta)}{\mathbf{i}\zeta}\exp(-(20.264)\zeta^2)e^{100\mathbf{i}\zeta}$ where C: $\int\lambda(\zeta)\,d\zeta=1$
$\nu_{y,0}(\zeta,\theta)$	$\lambda(\zeta)=\frac{\exp(-(0.72)\zeta^2)}{\mathbf{i}\zeta}e^{100\mathbf{i}\zeta}$

[a] The functions $\omega_j(\zeta)$, $\varpi(\zeta)$, and $\nu_{y,0}(\zeta,\theta)$ refer to the functions used in Section 3.2 to construct the moment conditions, using the function $\lambda\in\mathcal{G}$ as a starting point.

centered somewhere before and after the jump. After a Fourier transform operation, we again obtain a Gaussian, but the phase factor now includes a multiplicative factor of the form $(e^{\mathbf{i}\zeta c} - e^{-\mathbf{i}\zeta c})/2 = \mathbf{i}\sin(\zeta c)$ due to the presence of two shifted Gaussians with opposite signs. Because our procedure (in Section 3.2.1) requires the resulting function (of ζ) to be divided by $\gamma_o(\zeta, \theta)$, we insert a multiplicative factor of the form $(\mathbf{i}\zeta)^{-1}$ that is designed to cancel a similar divergence in the expression of $\gamma_o(\zeta, \theta)$. No additional step is required to achieve orthogonality of the singular part, because the behavior of the resulting function at the origin already guarantees a vanishing inner product with a delta function. However, we need to introduce a C numerically determined multiplicative constant to ensure that our function is properly normalized to integrate to 1, as required by the constraint $\varpi \in \mathcal{S}_1 \cap \mathcal{C}$.

As described in Section 3.2, the weighting function $\nu_{y,0}(\zeta, \theta)$ is derived from some function $\mu_{y,0} \in \mathcal{S}_0 \cap \mathcal{C}$, which, in turn, is derived from some $\lambda \in \mathcal{G}'$. The expression for $\lambda(\zeta)$ is again based on the Fourier transform of a shifted Gaussian. For the same reason as in the case of the $\varpi(\zeta)$ function, we introduce an $(\mathbf{i}\zeta)^{-1}$ factor to cancel a similar divergence in the expression of $\gamma_o(\zeta, \theta)$ when constructing $\mu_{y,0}(\zeta)$. The resulting expression already integrates to 0 (in the Cauchy principal value sense) and directly satisfies the constraint $\mu_{y,0} \in \mathcal{S}_0 \cap \mathcal{C}$, which ensures orthogonality of the ordinary part.

These steps provide us with a family of weighting functions with up to two adjustable parameters, typically one for the width of the Gaussian and one for its location (on the z axis). These numerical coefficients were selected by using the estimated asymptotic variance as an informal guide. The point estimates are not very sensitive to the exact values of these coefficients, as long as they are such that the general region where the functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$, and $r_{1y}(z, \theta)$ are the largest in magnitude corresponds to the range of values of Z found in the actual sample.

S.8. COMPUTATIONAL ASPECTS

The implementation of the estimator is considerably simplified by the fact that all the relatively abstract operations that require Fourier transforms involve nonrandom quantities. The end result of these operations is a vector of nonlinear functions whose expectations are to be evaluated from the observed data.

The first step in the implementation of the estimator is calculation of the Fourier transform $\gamma(\zeta, \theta)$ of $g(x^*, \theta)$. Symbolic mathematical packages such as Maple and Mathematica are often able to carry out such transforms automatically, even when the answers involve delta function derivatives. When an analytic expression for $\gamma(\zeta, \theta)$ is not available, the following hybrid analytical and numerical approach can be used. The idea is to write

$$g(x^*, \theta) = (g(x^*, \theta) - T(x^*, \theta)) + T(x^*, \theta),$$

where $T(x^*,\theta)$ represents the asymptotic behavior of $g(x^*,\theta)$ for large $|x^*|$ and where $(g(x^*,\theta)-T(x^*,\theta))$ is absolutely integrable (with respect to x^*). If the tail $T(x^*,\theta)$ follows a simple behavior such as a linear combination of functions of the form $(x^*)^{k_1}(\ln(x^*))^{k_2}$, then its Fourier transforms $\Theta(\zeta,\theta)$ can be found in standard Fourier transform tables (such as Table I in Lighthill (1962)). Typically, $\Theta(\zeta,\theta)$ will contain both a sum of delta function derivatives, which will provide the values of $\gamma_j(\theta)$ in Equations (32) and (33), and an ordinary function part $\Theta_o(\zeta,\theta)$. The Fourier transform of the remaining absolutely integrable contribution $(g(x^*,\theta)-T(x^*,\theta))$ can then be obtained numerically via

$$\gamma(\zeta,\theta)-\Theta(\zeta,\theta)=\lim_{\substack{t^*\to\infty\\ b\to 0}}\sum_{t=-t^*}^{t^*}(g(tb,\theta)-T(tb,\theta))e^{i\zeta tb}.$$

All the ordinary function contributions, $\gamma_o(\zeta,\theta)=\Theta_o(\zeta,\theta)+\gamma(\zeta,\theta)-\Theta(\zeta,\theta)$, are then added and their value over a grid $\mathbb{G}=\{\zeta\in\mathbb{R}:\zeta=tb,t=-t^*,\dots,0,\dots,t^*\}$ is stored, while making sure that the grid is sufficiently fine ($b\to 0$) and extended ($t^*\to\infty$) to provide an accurate numerical approximation to $\gamma_o(\zeta,\theta)$.

Dept. of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, U.S.A.; smschenn@uchicago.edu.

Manuscript received December, 2005; final revision received July, 2006.

REFERENCES

ANDREWS, D. W. K. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596. [9]

BOLLINGER, C. R. (2003): "Measurement Error in Human Capital and the Black–White Wage Gap," *Review of Economics and Statistics*, 85, 578–585. [32,33,39]

CARD, D., AND T. LEMIEUX (1994): "Changing Wage Structure and Black–White Wage Differentials among Men and Women: A Longitudinal Analysis," Working Paper 4755, National Bureau of Economic Research. [32]

CARNEIRO, P., J. J. HECKMAN, AND D. V. MASTEROV (2003): "Labor Market Discrimination and Racial Differences in Premarket Factors," Working Paper 10068, National Bureau of Economic Research. [32]

CHANDRA, A. (2003): "Is the Convergence of the Racial Wage Gap Illusory," Working Paper 9476, National Bureau of Economic Research. [40]

GEL'FAND, I. M., AND G. E. SHILOV (1964): *Generalized Functions*. New York: Academic Press. [1, 3]

HAUSMAN, J., W. NEWEY, H. ICHIMURA, AND J. POWELL (1991): "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 273–295. [22,23]

LIGHTHILL, M. J. (1962): *Introduction to Fourier Analysis and Generalized Function*. London: Cambridge University Press. [1,43]

NEAL, D. A., AND W. R. JOHNSON (1996): "The Role of Premarket Factors in Black–White Wage Differences," *Journal of Political Economy*, 104, 869–895. [32,33,35,38,39]

NEWEY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. IV, ed. by R. F. Engel and D. L. McFadden. New York: Elsevier Science, 2111–2245. [11-13,22]

NEWEY, W., AND R. J. SMITH (2004): "Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255. [28]

PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge, U.K: Cambridge University Press. [9]

PHILLIPS, P. C. B. (1991): "A Shortcut to LAD Estimator Asymptotics," *Econometric Theory*, 7, 450–463. [1]

POWELL, J., J. STOCK, AND T. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430. [18]

SCHWARTZ, L. (1966): *Théorie des Distributions*. Paris: Hermann. [1]

STOKER, T., E. BERNDT, D. ELLERMAN, AND S. M. SCHENNACH (2005): "Panel Data Analysis of U.S. Coal Productivity," *Journal of Econometrics*, 127, 131–164. [32]

TEMPLE, G. (1963): "The Theory of Weak Functions. I," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 276, 149–167. [3]

VAN DER VAAR, A. W. (1998): *Asymptotic Statistics*. Cambridge, U.K.: Cambridge University Press. [36]

ZINDE-WALSH, V., AND P. C. B. PHILLIPS (2003): "Fractional Brownian Motion as a Differentiable Generalized Gaussian Process," Working Paper 1391, Cowles Foundation, Yale University. [1]