Cowles Foundation Paper 863

Econometrica, Vol. 62, No. 1 (January, 1994), 43–72

ASYMPTOTICS FOR SEMIPARAMETRIC ECONOMETRIC MODELS VIA STOCHASTIC EQUICONTINUITY

BY DONALD W. K. ANDREWS[1]

This paper provides a general framework for proving the $\sqrt{T}$-consistency and asymptotic normality of a wide variety of semiparametric estimators. The class of estimators considered consists of estimators that can be defined as the solution to a minimization problem based on a criterion function that may depend on a preliminary infinite dimensional nuisance parameter estimator. The method of proof exploits results concerning the stochastic equicontinuity of stochastic processes. The results are applied to the problem of semiparametric weighted least squares estimation of partially parametric regression models. Primitive conditions are given for $\sqrt{T}$-consistency and asymptotic normality of this estimator.

KEYWORDS: Asymptotic normality, empirical process, infinite dimensional nuisance parameter, nonparametric estimation, semiparametric estimation, semiparametric model, stochastic equicontinuity, weak convergence.

1. INTRODUCTION

SEMIPARAMETRIC MODELS AND ESTIMATION PROCEDURES have become increasingly popular in econometrics in recent years. A large number of semiparametric estimators have been introduced and many have been shown to be $\sqrt{T}$-consistent and asymptotically normal. The proofs of such results are given in the literature on a case by case basis. No general results are available. The purpose of this paper is to provide a general framework for establishing the $\sqrt{T}$-consistency and asymptotic normality of a wide class of semiparametric estimators for time-series, cross-section, and panel data models. The general results are applied in the paper to establish the consistency and asymptotic normality of semiparametric weighted least squares (WLS) estimators of partially parametric regression (PPR) models.

The estimators considered in this paper are called MINPIN estimators. They are estimators that MINimize a criterion function that may depend on a Preliminary Infinite dimensional Nuisance parameter estimator. The criterion function need not be differentiable. As it happens, many of the semiparametric (and parametric) estimators in the literature are MINPIN estimators. Examples are given below.

The method of proof used here employs a condition called stochastic equicontinuity. This condition can be verified using empirical process results. An important feature of the method used is its generality. The same method can be used with a wide variety of estimators in different semiparametric

[1] I thank Whitney K. Newey, Adrian Pagan, David Pollard, B. M. Pötscher, James L. Powell, Yoon-Jae Whang, two referees, and a co-editor for helpful comments. I gratefully acknowledge research support from the Alfred P. Sloan Foundation and the National Science Foundation through a Research Fellowship and Grant Nos. SES-8618617, SES-8821021, and SES-9121914 respectively. This paper is a revision of a working paper whose first draft appeared in August, 1988.

models. The same method can be applied with independent identically distributed (iid), independent nonidentically distributed (inid), and dependent nonidentically distributed (dnid) random variables (rv's).[2]

A second feature of the method used is that the assumptions on the infinite dimensional nuisance parameter estimator and on the random criterion function are separated. Thus, there is no need to use sample splitting procedures and the results given below are flexible regarding the choice of estimator of the infinite dimensional nuisance parameter.

A third feature of the method used here is the simplicity of the structure of the proof. With the method used here, the key steps of the proof are highlighted and compartmentalized. The results given here, however, do not provide a complete proof of asymptotic normality except in the WLS/PPR application. The main results of this paper are proved under a set of "high-level" assumptions. In particular, we take as basic assumptions certain properties, such as consistency, of the infinite dimensional nuisance parameter estimator and the fulfillment of a uniform weak law of large numbers (WLLN), a CLT, and a stochastic equicontinuity condition for certain random variables. Verification of the uniform WLLN and CLT conditions is relatively easy, because there are numerous uniform WLLN and CLT results in the literature that are suitable without alteration. In addition, this paper and Andrews (1991a, 1994a) provide primitive conditions under which the stochastic equicontinuity condition holds. The remaining "high-level" assumptions that require verification concern the properties of the infinite dimensional nuisance parameter estimator. For kernel regression and density estimators, Andrews (1994b) provides results that establish the requisite properties. When the nuisance parameter is estimated by some method other than kernel estimation, the literature on nonparametric regression and density estimation can be exploited, although special tailoring of existing nonparametric results may be required.

A fourth feature of the method used here is the flexibility it affords with respect to the type of estimator considered. Many results in the semiparametric literature apply only to one-step estimators because of their technical tractability, among other reasons. The results of this paper apply to one-step versions of estimators as well as to the pure minimization versions of the estimators. One consequence of this is that LM and LR tests of parametric restrictions can be constructed in semiparametric contexts; see Andrews (1989a).

On the other hand, a drawback of the method used here is that in some examples it requires more smoothness conditions on certain underlying unknown functions than are necessary for $\sqrt{T}$-consistency and asymptotic normality of the estimator in question.

A second drawback of the method used here arises in those examples where trimming of nonparametric function estimators is required. In such examples, the method used here places more restrictions on the form of trimming that can be used than is necessary. This drawback and the previous one are conse-

[2]This is not to suggest that existing methods cannot be extended in such directions.

quences of the stochastic equicontinuity results that are currently available. It is possible that future developments of stochastic equicontinuity results will ameliorate these drawbacks.

A third drawback of the method used here is that, while the method is quite general, it is not applicable to all semiparametric estimators that are $\sqrt{T}$-consistent and asymptotically normal. Examples are given below.

The remainder of this paper is organized as follows: Section 2 introduces the stochastic equicontinuity property, outlines how it can be used to establish the asymptotic normality of semiparametric estimators, and provides a set of primitive sufficient conditions for it. Section 3 defines the class of MINPIN estimators, introduces the WLS/PPR application, and discusses which estimators fall in the MINPIN class. Section 4 gives $\sqrt{T}$-consistency and asymptotic normality results for MINPIN estimators. Section 5 uses the results of Section 4 and the Appendix to provide primitive conditions for the consistency and asymptotic normality of the WLS estimator of the PPR model. The Appendix provides consistency results for MINPIN estimators and proofs of the results given in the paper.

This paper does not cover tests of nonlinear parametric restrictions nor tests of model specification. See Andrews (1989a) and Whang and Andrews (1993), respectively, for treatments of these testing problems.

Throughout the paper all limits are taken as the sample size, T, goes to infinity. We let "wp $\to$ 1" abbreviate "with probability that goes to one as $T \to \infty$." We let $\|A\|$ denote the Euclidean norm of a vector or matrix A, i.e., $\|A\| = (\text{trace}(A'A))^{1/2}$. For notational simplicity, we let Σ_a^b denote $\Sigma_{t=a}^b$ and $E\|X\|^a$ denote $E(\|X\|^a)$.

2. STOCHASTIC EQUICONTINUITY

In this section, we introduce the concept of stochastic equicontinuity, show how it can be used in establishing the asymptotic normality of semiparametric estimators, provide a set of sufficient conditions for it, and sketch a proof of the sufficiency of the latter conditions.

2.1. *Some Basics Regarding Stochastic Equicontinuity*

Let $\{W_t : t = 1, 2, \dots\}$ be a sequence of $\mathscr{W}$-valued rv's defined on a probability space $(\Omega, \mathscr{B}, \mathbf{P})$, where $\mathscr{W} \subset R^k$. Let $\mathscr{T}$ be a pseudo-metric space with pseudo-metric $\rho_{\mathscr{T}}(\cdot, \cdot)$. (I.e., $\mathscr{T}$ is a metric space except that $\rho_{\mathscr{T}}(\tau_1, \tau_2) = 0$ does not necessarily imply that $\tau_1 = \tau_2$.) Let

$$(2.1) \qquad \mathscr{M} = \{m(\cdot, \tau) : \tau \in \mathscr{T}\}$$

be a class of R^v-valued functions defined on $\mathscr{W}$ and indexed by $\tau \in \mathscr{T}$. Define an empirical process $\nu_T(\cdot)$ by

$$(2.2) \qquad \nu_T(\tau) = \frac{1}{\sqrt{T}} \sum_1^T \left(m(W_t, \tau) - Em(W_t, \tau)\right) \quad \text{for} \quad \tau \in \mathscr{T},$$

where Σ_1^T abbreviates $\Sigma_{t=1}^T$. $m(\cdot, \tau)$ could depend on t in (2.2) if need be.

For the applications considered in this paper, τ is a vector-valued function (defined on some Euclidean space) and $\mathscr{T}$ is an infinite dimensional set of such functions. In particular, $\mathscr{T}$ often contains possible realizations of nonparametric regression or density estimators (viewed as estimators of entire functions, not just of functions at a single point). The nonparametric estimators are preliminary nuisance parameter estimators that appear in the definition of a semiparametric estimator. In the applications considered here, the summand $(1/\sqrt{T})\Sigma_1^T m(W_t, \tau)$ of the empirical process $\nu_T(\tau)$ equals the normalized first-order conditions for a semiparametric optimization estimator evaluated at the true value of the parameter $\theta_0 \in R^p$ of interest and evaluated at the value τ of the nonparametric regression or density function. Examples of the pseudo-metric $\rho_{\mathscr{T}}$ are given below.

Definition: $\{\nu_T(\cdot): T \geqslant 1\}$ is *stochastically equicontinuous* at τ_0 if for all $\varepsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ such that

$$(2.3) \qquad \overline{\lim_{T \to \infty}} \mathbf{P}^* \left(\sup_{\tau \in \mathscr{T}, \rho_{\mathscr{T}}(\tau, \tau_0) < \delta} |\nu_T(\tau) - \nu_T(\tau_0)| > \eta \right) < \varepsilon$$

where $\mathbf{P}^*$ denotes outer probability.

(If the rv in parentheses is measurable, then $\mathbf{P}^*$ can be replaced by $\mathbf{P}$.)

As can be seen from its definition, stochastic equicontinuity is a stochastic and asymptotic version of the concept of the continuity of a function. Essentially, it requires that $\nu_T(\cdot)$ is continuous at τ_0 at least with probability close to one for T large.

An equivalent definition of stochastic equicontinuity at τ_0 is: $\{\nu_T(\cdot): T \geqslant 1\}$ is stochastically equicontinuous at τ_0 if for all sequences of random elements $\{\hat{\tau}_T: T \geqslant 1\}$ that satisfy $\rho_{\mathscr{T}}(\hat{\tau}_T, \tau_0) \xrightarrow{p} 0$, we have $\nu_T(\hat{\tau}_T) - \nu_T(\tau_0) \xrightarrow{p} 0$.

The concept of stochastic equicontinuity is not new. It has appeared in the literature under various guises. For example, it appears in Theorem 8.2 of Billingsley (1968, p. 55), which is attributed to Prohorov (1956), for the case of $C[0,1]$ random elements.

The plausibility of the stochastic equicontinuity property can be demonstrated as follows. Suppose $\{m(w, \tau): \tau \in \mathscr{T}\}$ is a class of linear functions, i.e., $m(w, \tau) = w'\tau$ for some $\tau \in R^k$, and $\rho_{\mathscr{T}}(\cdot, \cdot)$ is the Euclidean metric. In this case, the left-hand side of (2.3) equals

$$(2.4) \qquad \overline{\lim_{T \to \infty}} \mathbf{P}^* \left[\sup_{\rho_{\mathscr{T}}(\tau, \tau_0) < \delta} \left| \frac{1}{\sqrt{T}} \sum_1^T (W_t - EW_t)'(\tau - \tau_0) \right| > \eta \right]$$
$$\leqslant \overline{\lim_{T \to \infty}} \mathbf{P} \left[\left\| \frac{1}{\sqrt{T}} \sum_1^T (W_t - EW_t) \right\| > \eta/\delta \right] < \varepsilon$$

provided $(1/\sqrt{T})\Sigma_1^T (W_t - EW_t) = O_p(1)$ and δ is sufficiently small. In consequence, stochastic equicontinuity holds in this case if the rv's $\{W_t - EW_t\}$ satisfy an ordinary CLT.

The stochastic equicontinuity property is substantially more difficult to verify for classes of nonlinear functions than for classes of linear functions. Indeed, it does not hold for all classes of functions. Some restrictions on the functions are necessary. The following example illustrates why. Suppose $\{W_t: t \geqslant 1\}$ are iid with distribution P_1 that is absolutely continuous with respect to Lesbesgue measure and $\mathscr{M}$ is the class of indicator functions of all Borel sets in $\mathscr{W}$. Let τ denote a Borel set in $\mathscr{W}$ and let $\mathscr{T}$ denote the collection of all such sets. Then, $m(w,\tau) = 1(w \in \tau)$. Take $\rho_{\mathscr{T}}(\tau_1, \tau_2) = [\int (m(w,\tau_1) - m(w,\tau_2))^2 \, dP_1(w)]^{1/2}$. For any two sets τ_1, τ_2 in $\mathscr{T}$ with finite numbers of elements, $\nu_T(\tau_j) = (1/\sqrt{T})\Sigma_1^T 1(W_t \in \tau_j)$ and $\rho_{\mathscr{T}}(\tau_1, \tau_2) = 0$ since $P_1(W_t \in \tau_j) = 0$ for $j = 1,2$. Given any $T \geqslant 1$ and any realization $\omega \in \Omega$, there exist finite sets $\tau_{1T\omega}$ and $\tau_{2T\omega}$ in $\mathscr{T}$ such that $W_t(\omega) \in \tau_{1T\omega}$ and $W_t(\omega) \notin \tau_{2T\omega}$ $\forall t \leqslant T$, where $W_t(\omega)$ denotes the value of W_t when ω is realized. This yields $\nu_T(\tau_{1T\omega}) = \sqrt{T}$, $\nu_T(\tau_{2T\omega}) = 0$, and $\sup_{\rho_{\mathscr{T}}(\tau_1,\tau_2)<\delta} |\nu_T(\tau_1) - \nu_T(\tau_2)| \geqslant \sqrt{T}$. Thus, $\{\nu_T(\cdot): T \geqslant 1\}$ is not stochastically equicontinuous. The class of functions considered is too large. Below we provide one set of conditions that restricts the class of functions sufficiently such that stochastic equicontinuity holds.

Although the stochastic equicontinuity property is used here in establishing the asymptotic distribution of semiparametric estimators, its primary use in the probability literature is in the proof of weak convergence results, including abstract functional central limit theorems (CLTs). The following result indicates how it is used to establish weak convergence: If $(\mathscr{T}, \rho_{\mathscr{T}})$ is a totally bounded pseudo-metric space, $\{\nu_T(\cdot): T \geqslant 1\}$ is stochastically equicontinuous, and $\{(\nu_T(\tau_1), \ldots, \nu_T(\tau_L))': T \geqslant 1\}$ converges in distribution for all finite dimensional vectors $(\tau_1, \ldots, \tau_L)'$ of elements of $\mathscr{T}$, then $\{\nu_T(\cdot): T \geqslant 1\}$ converges weakly to a stochastic process on $\mathscr{T}$ that has uniformly $\rho_{\mathscr{T}}$-continuous sample paths. Conversely, if $(\mathscr{T}, \rho_{\mathscr{T}})$ is a totally bounded pseudo-metric space and $\{\nu_T(\cdot): T \geqslant 1\}$ converges weakly to such a process on $\mathscr{T}$, then $\{\nu_T(\cdot): T \geqslant 1\}$ is stochastically equicontinuous. See Pollard (1990, Thm. 10.2) for details. (As the above result indicates, stochastic equicontinuity and *tightness* are very closely related properties.)

2.2. *Sketch of the Proof of Asymptotic Normality of Semiparametric Estimators via Stochastic Equicontinuity*

We now give a heuristic description of how stochastic equicontinuity can be used to establish the asymptotic normality of semiparametric estimators. For the time being, suppose $\hat{\theta}$ is a consistent estimator of a parameter $\theta_0 \in R^p$ that solves for θ the first order conditions

$$\sqrt{T}\,\overline{m}_T(\theta, \hat{\tau}) = \mathbf{0} \qquad \text{wp} \to 1, \tag{2.5}$$

where

$$\overline{m}_T(\theta, \hat{\tau}) = \frac{1}{T}\sum_1^T m(W_t, \theta, \hat{\tau}) \in R^p.$$

Here, $\hat{\tau}$ typically is some preliminary nonparametric estimator of regression or density functions. The summand $m(W_t, \theta, \hat{\tau})$ might depend on $\hat{\tau}$ through the value of the regression or density function estimator at the rv X_t, i.e., $\hat{\tau}(X_t)$, where X_t is a subvector of W_t. Alternatively, $m(W_t, \theta, \hat{\tau})$ might depend on $\hat{\tau}$ in a more complicated way, such as through integrals or derivatives of $\hat{\tau}$.

Suppose $\hat{\tau}$ lies in a pseudo-metric space $\mathscr{T}$ wp $\to 1$ and is consistent for $\tau_0 \in \mathscr{T}$ with respect to the pseudo-metric $\rho_{\mathscr{T}}$. (I.e., $\rho_{\mathscr{T}}(\hat{\tau}, \tau_0) \xrightarrow{p} 0$.) Assume the population first order conditions, $\overline{m}_T^*(\theta_0, \tau_0) = E\overline{m}_T(\theta_0, \tau_0) = \mathbf{0}$, hold.

We consider the case where $m(W_t, \theta, \tau)$ is differentiable in θ. If τ was finite dimensional one could establish the asymptotic normality of $\hat{\theta}$ by expanding $\sqrt{T}\,\overline{m}_T(\hat{\theta}, \hat{\tau})$ about (θ_0, τ_0) using element by element mean value expansions. Since τ is infinite dimensional, however, mean value expansions in (θ, τ) are not available. In consequence, we expand $\sqrt{T}\,\overline{m}_T(\hat{\theta}, \hat{\tau})$ about θ_0 only (using element by element mean value expansions) and use stochastic equicontinuity and an asymptotic orthogonality condition to handle $\hat{\tau}$:

$$(2.6)\qquad o_p(1) = \sqrt{T}\,\overline{m}_T(\hat{\theta}, \hat{\tau}) = \sqrt{T}\,\overline{m}_T(\theta_0, \hat{\tau}) + \frac{\partial}{\partial \theta'}\overline{m}_T(\theta^*, \hat{\tau})\sqrt{T}\left(\hat{\theta} - \theta_0\right),$$

where θ^* lies on the line segment joining $\hat{\theta}$ and θ_0 (and takes different values in each row of $(\partial/\partial\theta')\overline{m}_T(\theta^*, \hat{\tau})$). Under suitable assumptions on $\{m(W_t, \theta, \tau)\colon t \geqslant 1\}$, one can show that

$$(\partial/\partial\theta')\overline{m}_T(\theta^*, \hat{\tau}) \xrightarrow{p} M = \lim_{T\to\infty} (1/T) \sum_1^T E(\partial/\partial\theta') m_t(\theta_0, \tau_0).$$

Thus, provided M is nonsingular,

$$(2.7)\qquad \sqrt{T}\left(\hat{\theta} - \theta_0\right) = \left(M^{-1} + o_p(1)\right)\sqrt{T}\,\overline{m}_T(\theta_0, \hat{\tau}).$$

If $\hat{\tau}$ is replaced by τ_0 in (2.7), the right-hand side of (2.7) is asymptotically normal, say $N(\mathbf{0}, S)$, under general conditions by a CLT, since $\sqrt{T}\,\overline{m}_T(\theta_0, \tau_0)$ is a mean zero sample average normalized by $\sqrt{T}$. Hence, if we can show that

$$(2.8)\qquad \sqrt{T}\,\overline{m}_T(\theta_0, \hat{\tau}) - \sqrt{T}\,\overline{m}_T(\theta_0, \tau_0) \xrightarrow{p} \mathbf{0},$$

then we will have established that $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, M^{-1}S(M^{-1})')$. Note that in this case the estimation of τ_0 by $\hat{\tau}$ does not affect the asymptotic distribution of $\hat{\theta}$.

Stochastic equicontinuity is useful in establishing (2.8). In particular, stochastic equicontinuity of $\nu_T(\tau) = \sqrt{T}(\overline{m}_T(\theta_0, \tau) - \overline{m}_T^*(\theta_0, \tau))$ (indexed by $\tau \in \mathscr{T}$) at τ_0, consistency of $\hat{\tau}$ for τ_0 with respect to the pseudo-metric $\rho_{\mathscr{T}}$, and $P(\hat{\tau} \in \mathscr{T}) \to 1$ yield

$$(2.9)\qquad \nu_T(\hat{\tau}) - \nu_T(\tau_0) \xrightarrow{p} \mathbf{0}.$$

This follows because given any η, $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$(2.10)\quad \overline{\lim_{T\to\infty}} P(|\nu_T(\hat{\tau}) - \nu_T(\tau_0)| > \eta)$$

$$\leqslant \overline{\lim_{T\to\infty}} P(|\nu_T(\hat{\tau}) - \nu_T(\tau_0)| > \eta, \hat{\tau} \in \mathscr{T}, \rho_{\mathscr{T}}(\hat{\tau}, \tau_0) \leqslant \delta)$$

$$+ \overline{\lim_{T\to\infty}} P(\hat{\tau} \notin \mathscr{T} \text{ or } \rho_{\mathscr{T}}(\hat{\tau}, \tau_0) > \delta)$$

$$\leqslant \overline{\lim_{T\to\infty}} P^*\Big(\sup_{\tau \in \mathscr{T}:\, \rho_{\mathscr{T}}(\tau, \tau_0) \leqslant \delta} |\nu_T(\tau) - \nu_T(\tau_0)| > \eta \Big)$$

$$< \varepsilon.$$

Since

$$(2.11)\quad \sqrt{T}\overline{m}_T(\theta_0, \hat{\tau}) - \sqrt{T}\overline{m}_T(\theta_0, \tau_0) = \nu_T(\hat{\tau}) - \nu_T(\tau_0) - \sqrt{T}\overline{m}_T^*(\theta_0, \hat{\tau}),$$

equation (2.8) now holds if and only if

$$(2.12)\quad \sqrt{T}\overline{m}_T^*(\theta_0, \hat{\tau}) \xrightarrow{p} \mathbf{0}.$$

The latter is an asymptotic orthogonality condition between $\hat{\theta}$ and $\hat{\tau}$ that is analogous to the block diagonality of the information matrix between θ_0 and τ_0 in the case of ML estimation with finite dimensional τ. This condition is usually satisfied by adaptive estimators of adaptive models, but is also satisfied by numerous semiparametric estimators of nonadaptive models (for suitable estimators $\hat{\tau}$), as discussed below.

2.3. *Sufficient Conditions for Stochastic Equicontinuity*

In this subsection, we provide a set of sufficient conditions for stochastic equicontinuity developed in Andrews (1991a). These sufficient conditions are employed in the WLS/PPR application in Section 5 below. Andrews (1994a) gives several alternative sets of primitive sufficient conditions for stochastic equicontinuity based on results in the probability literature.

The conditions considered here apply to classes of smooth functions and underlying dnid rv's that are strong mixing. First we define strong mixing. Let $\mathscr{F}_s^t$ denote the σ-field generated by $\{W_s, \ldots, W_t\}$.

Definition: The rv's $\{W_t: t \geqslant 1$ are *strong mixing* if $\alpha(s) \downarrow 0$ as $s \to \infty$, where

$$(2.13)\quad \alpha(s) = \sup_{t \geqslant 1} \sup_{A \in \mathscr{F}_1^t,\, B \in \mathscr{F}_{t+s}^\infty} |P(A \cap B) - P(A)P(B)| \quad \text{for } s \geqslant 1.$$

The functions we consider are smooth on an open bounded set $\mathscr{W}^* \subset \mathscr{W}$ and constant elsewhere. If $\mathscr{W}^* = \mathscr{W}$, then the functions are smooth on their entire domain. $\mathscr{W}^*$ is required to be a set with minimally smooth boundary as defined, e.g., in Stein (1970). Examples of such sets include convex sets and finite unions of convex sets with disjoint closures.

Smoothness of the functions on $\mathscr{W}^*$ is defined in terms of the L^2-Sobolev norm. For an R^v-valued function f on $\mathscr{W} \subset R^k$, a k-vector $\mu = (\mu_1, \dots, \mu_k)'$ of nonnegative integers, and a nonnegative integer q, define

$$(2.14) \quad D^\mu f(w) = \frac{\partial^{|\mu|}}{\partial w_1^{\mu_1} \times \cdots \times \partial w_k^{\mu_k}} f(w), \quad \text{where} \quad |\mu| = \sum_{j=1}^{k} \mu_j, \quad \text{and}$$

$$\| f(\cdot) \|_{q, \mathscr{W}^*} = \left(\sum_{|\mu| \leqslant q} \int_{\mathscr{W}^*} (D^\mu f(w))^2 \, dw \right)^{1/2}.$$

$\| \cdot \|_{q, \mathscr{W}^*}$ is the L^2-Sobolev norm of order q over $\mathscr{W}^*$.

By restricting the class of functions $\mathscr{M}$ to contain functions whose L^2-Sobolev norm (of some order $q > k/2$) is bounded by some constant, one can obtain stochastic equicontinuity. The pseudo-metric that is employed is the L^2-pseudo-metric:

$$(2.15) \quad \rho_{\mathscr{T}}(\tau_1, \tau_2) = \left[\int_{\mathscr{W}^*} (m(w, \tau_1) - m(w, \tau_2))^2 dw \right]^{1/2}.$$

The requisite assumptions are summarized as follows:

Assumption SE: (a) $\sup_{\tau \in \mathscr{T}} \|m(\cdot, \tau)\|_{q, \mathscr{W}^*} < \infty$ *for some* $q > k/2$.

(b) *For some constant* K, $m(w, \tau) = K \ \forall w \in \mathscr{W} - \mathscr{W}^* \ \forall \tau \in \mathscr{T}$.

(c) $\mathscr{W}^* \subset \mathscr{W}$ *and* $\mathscr{W}^*$ *is an open bounded subset of* R^k *with minimally smooth boundary.*

(d) $\{W_t : t \geqslant 1\}$ *is a strong mixing sequence of rv's with* $\sum_{s=1}^{\infty} \alpha(s) < \infty$.

Proposition: *Assumption SE implies that* $\{\nu_T(\cdot): T \geqslant 1\}$ *is stochastically equicontinuous at each* $\tau_0 \in \mathscr{T}$ *and* $\mathscr{T}$ *is totally bounded under* $\rho_{\mathscr{T}}$ *(for* $\rho_{\mathscr{T}}$ *given in* (2.15)).

Comments: 1. The Proposition is a slight variant of a special case of Theorem 4 of Andrews (1991a) (discussed in Comment 1 to that Theorem). The proof is given in Andrews (1991a).

2. The Proposition is employed in semiparametric applications by showing that the nonparametric estimators $\hat{\tau}$ are such that $m(w, \theta_0, \hat{\tau})$ satisfies the conditions of Assumption SE for all realizations of $\hat{\tau}$, at least wp $\to 1$.

3. One undesirable feature of the Proposition is that the functions are nonconstant only on a bounded set. This restriction is relaxed to a certain extent in an extension given in Andrews (1989b, Thm. 7).

3. MINPIN ESTIMATORS

3.1. *Definition of MINPIN Estimators*

The data are given by a sequence of random vectors (rv's) $\{W_t: t = 1, 2, \dots\}$ defined on some probability space $(\Omega, \mathscr{B}, \mathbf{P})$. The observed sample is $\{W_t: t = 1, \dots, T\}$. A MINPIN estimator $\hat{\theta}$ of an unknown p-vector θ_0 is defined to minimize a criterion function $d(\overline{m}_T(\theta, \hat{\tau}), \hat{\gamma})$ over a parameter space Θ. Here, $\overline{m}_T(\theta, \hat{\tau})$ is a sample average of terms of the form $m_t(W_t, \theta, \hat{\tau})$, $\hat{\tau}$ is a preliminary infinite dimensional nuisance parameter estimator, $\hat{\gamma}$ is a preliminary nuisance parameter estimator (often an estimated weight matrix), and $d(m, \gamma)$ is a fixed discrepancy function (such as a quadratic form $m'\gamma m/2$).

DEFINITION: A *sequence of MINPIN estimators* $\{\hat{\theta}\} = \{\hat{\theta}: T \geqslant 1\}$ is any sequence of rv's such that

$$(3.1) \qquad d\left(\overline{m}_T(\hat{\theta}, \hat{\tau}), \hat{\gamma}\right) = \inf_{\theta \in \Theta} d\left(\overline{m}_T(\theta, \hat{\tau}), \hat{\gamma}\right) \quad \text{wp} \to 1,$$

where $\overline{m}_T(\theta, \tau) = (1/T)\Sigma_1^T m_t(\theta, \tau)$, $m_t(\theta, \tau)$ denotes $m_t(W_t, \theta, \tau)$, $m_t(\cdot, \cdot, \cdot)$ is a function from $R^k \times \Theta \times \mathscr{T}$ to R^v, $\Theta \subset R^p$, $\hat{\tau}$ is a random element of $\mathscr{T}$ wp $\to 1$, $\hat{\gamma}$ is a random element of Γ (and $\hat{\tau}$ and $\hat{\gamma}$ depend on T in general), $\mathscr{T}$ and Γ are pseudo-metric spaces, and $d(\cdot, \cdot)$ is a nonrandom, real-valued function (which does not depend on T).[3]

Note that $\hat{\tau}$ and $\hat{\gamma}$ are preliminary, possibly infinite dimensional, estimators used in the definition of $\hat{\theta}$. Usually, however, $\hat{\gamma}$ is finite dimensional or does not appear. Convergence in probability of $\hat{\tau}$ and $\hat{\gamma}$ means convergence in probability with respect to the pseudo-metrics on $\mathscr{T}$ and Γ respectively. Throughout this paper, all functions that are introduced (such as $\hat{\theta}$, $\hat{\tau}$, $\hat{\gamma}$, $m_t(\cdot, \cdot, \cdot)$, and $d(\cdot, \cdot)$) are assumed to be $\mathscr{B}$/Borel or Borel/Borel measurable. The only exception is the stochastic process $\nu_T(\cdot)$ defined below, which need not be measurable. We note that one set of sufficient conditions for the existence of a measurable sequence $\{\hat{\theta}\}$ is that $d(\overline{m}_T(\theta, \hat{\tau}), \hat{\gamma})$ viewed as a function from $\Omega \times \Theta$ to R is continuous in θ for each $\omega \in \Omega$ and is measurable for each fixed $\theta \in \Theta$ and Θ is a compact subset of some Euclidean space (see Jennrich (1969, Lemma 2)).

[3]The criterion function $d(\overline{m}_T(\theta, \hat{\tau}), \hat{\gamma})$ is allowed to depend on *two* preliminary estimators, $\hat{\tau}$ and $\hat{\gamma}$. We do not merge $\hat{\tau}$ and $\hat{\gamma}$ into one estimator, because the use of two preliminary estimators allows one to simplify and weaken the assumptions. In particular, with two preliminary estimators, $\nu_T(\cdot)$ (defined below) only has to be indexed by τ in Assumption N(e) rather than by (τ, γ).

The estimator $\hat{\theta}$ is required to solve (3.1) only wp $\to 1$ to enable one to define the same estimator using different $m_t(\theta, \tau)$ and $d(m, \gamma)$ functions for the purposes of (i) consistency and (ii) asymptotic normality.

The infinite dimensional estimator $\hat{\tau}$ is only required to lie in $\mathscr{T}$ wp $\to 1$, because $\mathscr{T}$ is taken below to contain elements that satisfy certain properties, i.e., smoothness properties. In many cases, not all realizations of $\hat{\tau}$ satisfy these properties, but the realizations in a set whose probability $\to 1$ do satisfy them.

3.2. *WLS/PPR Application*

In this subsection, we discuss a semiparametric weighted least squares (WLS) estimator in terms of the above definition of MINPIN estimators. The model considered is a partially parametric regression (PPR) model. This model is a generalization of the partially linear regression (PLR) model that has been considered in numerous papers, e.g., see Robinson (1988). The PPR model allows the parametric part of the regression function to be nonlinear. The model is

$$\begin{aligned} (3.2) \qquad & Y_t = h(Z_t, \theta_0) + g(X_t) + U_t \quad \text{when} \quad X_t \in \mathscr{X}^*, \\ & E(U_t \mid Z_t, X_t) = 0 \quad \text{a.s., and} \\ & E\left(U_t^2 \mid Z_t, X_t\right) = \tau_{30}(X_t) \quad \text{a.s. for} \quad t = 1, \ldots, T, \end{aligned}$$

where the real function $h(\cdot, \cdot)$ is known, the real functions $g(\cdot)$ and $\tau_{30}(\cdot)$ are unknown, $Y_t, U_t \in R$, $\theta_0 \in R^p$, $Z_t \in R^l$, $X_t \in R^k$, $W_t = (Y_t, Z_t', X_t')'$, and $\mathscr{X}^*$ is some bounded subset of R^k. ($\mathscr{X}^*$ is a trimming set.) Unlike most analyses in the literature of the PLR model, we consider PPR models that may exhibit conditional heteroskedasticity, nonidentical distributions, and temporal dependence of the data. An interesting example of the latter is a dynamic sample selection model in which the selection equation for period t depends on Y_{t-1} and, in consequence, Y_{t-1} is an element of X_t. For example, a model of this sort might be postulated in the context of time series observations on regulated utility rate of return grants (which exhibit a selectivity feature because a grant occurs only if a firm requests it); see Roberts, Maddala, and Enholm (1978) for a parametric analysis of such data. The treatment of nonidentical distributions allows one to cover data generated via stratified sampling schemes.

We consider a nonlinear WLS estimator of θ_0 that is an analogue of the linear LS estimator of Robinson (1988). The WLS estimator is designed for the case where the conditional variance of U_t given (Z_t, X_t) depends on X_t. To motivate this estimator, we note that the PPR model with heteroskedasticity of this form is generated by the following sample selection model:

$$\begin{aligned} (3.3) \qquad & \tilde{Y}_t = h\left(\tilde{Z}_t, \theta_0\right) + v_1\left(\tilde{X}_t, \phi_0\right) + \tilde{U}_t, \qquad D_t = 1\left(v_2\left(\tilde{X}_t, \tilde{\varepsilon}_t\right) > 0\right), \quad \text{and} \\ & (Y_t, D_t, Z_t, X_t) = \left(\tilde{Y}_t D_t, D_t, \tilde{Z}_t D_t, \tilde{X}_t D_t\right) \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{are observed for } t = 1, \ldots, T, \end{aligned}$$

where $h(\cdot, \cdot)$ is known, $v_1(\cdot, \cdot)$ may or may not be known, $v_2(\cdot, \cdot)$ is unknown, $(\tilde{U}_t, \tilde{\varepsilon}_t, \tilde{Z}_t, \tilde{X}_t)$ is identically distributed for $t \geqslant 1$, and $(\tilde{U}_t, \tilde{\varepsilon}_t)$ is independent of $(\tilde{Z}_t, \tilde{X}_t)$ and has unknown distribution. By multiplying the first equation of (3.3) by D_t, one sees that the sample selection model (3.3) generates the PPR model (3.2) with the unknown function $g(\cdot)$ of (3.2) given by $g(x) = v_1(x, \phi_0) + E(D_t \tilde{U}_t \mid X_t = x)$ and with the error of (3.2) given by $U_t = D_t \tilde{U}_t - E(D_t \tilde{U}_t \mid X_t)$. Note that U_t has conditional variance given (Z_t, X_t) that depends only on X_t as in (3.2). The sample selection-generated PPR model is particularly useful when there is incomplete observation of the selection equation regressor variables for nonselected observations, as often occurs. In such cases, some other semiparametric estimators, such as Powell's (1987) and Newey's (1988) two-step estimators, are not applicable.

For the PPR model of (3.2), define $\tau_0 = (\tau_{10}, \tau_{20}, \tau_{30})$, where

$$\tau_{10}(x) = E(Y_t \mid X_t = x) \quad \text{and} \quad \tau_{20}(x, \theta_0) = E(h(Z_t, \theta_0) \mid X_t = x). \tag{3.4}$$

The WLS estimator of θ_0 for the PPR model minimizes

$$\frac{1}{T}\sum_1^T \xi(X_t)(Y_t - \hat{\tau}_1(X_t) - h(Z_t, \theta) + \hat{\tau}_2(X_t, \theta))^2/\hat{\tau}_3(X_t) \tag{3.5}$$

over $\Theta \subset R^p$ wp $\to 1$, where $\xi(X_t) = 1(X_t \in \mathscr{X}^*)$ and $\hat{\tau} = (\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3)$ is a nonparametric estimator of τ_0 (defined in Section 5 below). The WLS estimator is a MINPIN estimator with

$$d(m, \gamma) = m \quad \text{and} \tag{3.6}$$

$$m_t(\theta, \tau) = \xi(X_t)(Y_t - \tau_1(X_t) - h(Z_t, \theta) + \tau_2(X_t, \theta))^2/\tau_3(X_t) \quad \text{or}$$

$$d(m, \gamma) = m'm/2 \quad \text{and} \tag{3.7}$$

$$m_t(\theta, \tau) = \xi(X_t)(Y_t - \tau_1(X_t) - h(Z_t, \theta) + \tau_2(X_t, \theta)) \times \left[\frac{\partial}{\partial\theta}h(Z_t, \theta) - \frac{\partial}{\partial\theta}\tau_2(X_t, \theta)\right] \Big/ \tau_3(X_t).$$

We use (3.6) when establishing consistency of $\hat{\theta}$ and (3.7) when establishing asymptotic normality of $\hat{\theta}$.

Note that the WLS estimator $\hat{\theta}$ is defined using a trimming function $\xi(X_t)$, because the regression model (3.2) is only assumed to hold on the bounded set $\mathscr{X}^*$. Even if the model holds for x in a larger set than $\mathscr{X}^*$, there are two reasons why we still define $\hat{\theta}$ using the trimming function $\xi(X_t)$. First, trimming can eliminate observations from the computation of $\hat{\theta}$ for which the nuisance parameter estimator $\hat{\tau}(X_t)$ is estimated with relatively large error in comparison to the nontrimmed observations. Second, trimming makes it much simpler to establish the consistency and asymptotic normality of $\hat{\theta}$, because one can obtain uniform consistency of $\hat{\tau}(x)$ for $\tau_0(x)$ over a bounded set $\mathscr{X}^*$ under suitable conditions, but not over unbounded sets in general. On the other hand, trimming using a single fixed set $\mathscr{X}^*$ affects the asymptotic distribution of $\hat{\theta}$ and usually sacrifices some asymptotic efficiency if, in fact, the model holds for x in a larger set than $\mathscr{X}^*$.

3.3. *Further Applications*

In this section, we specify additional examples of estimators that fall within the MINPIN class. Those marked with an asterisk are discussed briefly in a supplement to the paper that is available from the author.

The examples include: (1)* Generalized method of moments estimators of models defined by conditional moment restrictions; see Newey (1990a) and Robinson (1987). (2)* Semiparametric instrumental variable estimators for regression models with unobserved risk variables; see Pagan and Ullah (1988). (3)* Estimators of index regression models, such as Klein and Spady's (1993) efficient semiparametric estimator of the binary choice model and Ichimura and Lee's (1990) LS estimator. (4)* Two- and three-step estimators of the sample selection model. For the former, see Powell (1987) and Newey (1988). (5)* Adaptive estimators for regression models with errors of unknown distribution; see Bickel (1982) and Manski (1984) for one-step estimators for independent error models. (6) Adaptive estimators of autoregressive moving average models with innovations with unknown distribution; see Kreiss (1987) for one-step estimators. (7) Profile likelihood estimators for semiparametric models; see Severini and Wong (1987) and Lee (1989).

4. ASYMPTOTIC NORMALITY

In this section, sufficient conditions for the asymptotic normality of sequences of MINPIN estimators $\{\hat{\theta}\}$ are given. Before stating these conditions, we define the asymptotic covariance matrix of $\{\hat{\theta}\}$ and introduce some notation and definitions used in the assumptions.

4.1. *The Asymptotic Covariance Matrix of MINPIN Estimators*

Let $d(m,\gamma)$ and $m_t(\theta,\tau)$ be defined such that the dimension v of $\overline{m}_T(\theta,\tau)$ is at least as large as the dimension p of θ. For example, for the WLS/PPR application, we use (3.7), not (3.6). The asymptotic covariance matrix V of $\{\hat{\theta}\}$ is then defined by

$$(4.1)\qquad S=\lim_{T\to\infty}\operatorname{Var}_{\mathbf{P}}\left(\sqrt{T}\,\overline{m}_T(\theta_0,\tau_0)\right),\qquad M=\lim_{T\to\infty}\frac{1}{T}\sum_1^T\frac{\partial}{\partial\theta'}Em_t(\theta_0,\tau_0),$$

$$D=\frac{\partial^2}{\partial m\,\partial m'}d\big(m(\theta_0,\tau_0),\gamma_0\big),\qquad \mathscr{J}=M'DM,\qquad \mathscr{I}=M'DSDM,$$

and

$$V=\mathscr{J}^{-1}\mathscr{I}\mathscr{J}^{-1},$$

where $(\partial^2/\partial m\,\partial m')d(\cdot,\cdot)$ denotes the matrix of second partial derivatives of $d(\cdot,\cdot)$ with respect to its first argument and $m(\theta_0,\tau_0)=\lim_{T\to\infty}(1/T)\Sigma_1^TEm_t(\theta_0,\tau_0)$. In the common case where $p=v$, the covariance matrix V simplifies to

$$(4.2)\qquad V=M^{-1}S(M^{-1})'.$$

For example, for the WLS estimator of the PPR model, we have $p = v$ and

(4.3) $V = M^{-1}SM^{-1}$, where

$$M = -\lim_{T\to\infty}\frac{1}{T}\sum_{1}^{T}E\xi(X_t)\left[\frac{\partial}{\partial\theta}h(Z_t,\theta_0) - \frac{\partial}{\partial\theta}\tau_{20}(X_t,\theta_0)\right]$$
$$\times\left[\frac{\partial}{\partial\theta}h(Z_t,\theta_0) - \frac{\partial}{\partial\theta}\tau_{20}(X_t,\theta_0)\right]'\Big/\tau_{30}(X_t), \qquad D = M^{-1}, \text{ and}$$
$$S = \lim_{T\to\infty}\operatorname{Var}\left[\frac{1}{\sqrt{T}}\sum_{1}^{T}E\xi(X_t)U_t \times\left[\frac{\partial}{\partial\theta}h(Z_t,\theta_0) - \frac{\partial}{\partial\theta}\tau_{20}(X_t,\theta_0)\Big/\tau_{30}(X_t)\right]\right].$$

If W_t is independent across t, S simplifies to M and $V = M^{-1}$.

4.2. *Asymptotic Normality of MINPIN Estimators*

We now state Assumption N that is sufficient for asymptotic normality of the MINPIN estimator $\hat{\theta}$. Let Θ_0 be a subset of $\Theta(\subset R^p)$ that contains a neighborhood of θ_0. Define

(4.4) $$B_t = \sup_{\theta\in\Theta_0,\,\tau\in\mathscr{T}}\left\|\frac{\partial^2}{\partial\theta\partial\theta'}m_t(\theta,\tau)\right\| \quad\text{and}$$
$$\nu_T(\tau) = \sqrt{T}\left(\overline{m}_T(\theta_0,\tau) - \overline{m}_T^*(\theta_0,\tau)\right),$$

where

$$\overline{m}_T^*(\theta,\tau) = \frac{1}{T}\sum_{1}^{T}Em_t(\theta,\tau).$$

ASSUMPTION N (Normality): (a) *$\hat{\theta}\xrightarrow{p}\theta_0\in\Theta\subset R^p$ and θ_0 is in the interior of Θ.*

(b) *$P(\hat{\tau}\in\mathscr{T})\to 1$, $\hat{\tau}\xrightarrow{p}\tau_0$, and $\hat{\gamma}\xrightarrow{p}\gamma_0$ for some $\tau_0\in\mathscr{T}$ and $\gamma_0\in\Gamma$.*

(c) *$\sqrt{T}(\partial/\partial m)d(\overline{m}_T^*(\theta_0,\hat{\tau}),\hat{\gamma})\xrightarrow{p}\mathbf{0}$.*

(d) *$\nu_T(\tau_0)\xrightarrow{d}N(\mathbf{0},S)$.*

(e) *$\{\nu_T(\cdot)\}$ is stochastically equicontinuous at τ_0.*

(f) *$(\partial/\partial m)d(m,\gamma)$ and $(\partial^2/\partial m\,\partial m')d(m,\gamma)$ exist for $(m,\gamma)\in\mathscr{M}_0\times\Gamma_0$ and are continuous at $(m,\gamma) = (m(\theta_0,\tau_0),\gamma_0)$, where $\mathscr{M}_0$ and Γ_0 are subsets of R^v and Γ that contain neighborhoods of $m(\theta_0,\tau_0)$ and γ_0 respectively (using the Euclidean norm on R^v and the pseudo-metric on Γ).*

(g) *$m_t(\theta,\tau)$ is twice continuously differentiable in θ on Θ_0, $\forall\tau\in\mathscr{T}$, $\forall t\geqslant 1$, $\forall\omega\in\Omega$. $\{m_t(\theta,\tau)\}$ and $\{(\partial/\partial\theta')m_t(\theta,\tau)\}$ satisfy uniform WLLNs over $\Theta_0\times\mathscr{T}$. $m(\theta,\tau) = \lim_{T\to\infty}(1/T)\sum_1^T Em_t(\theta,\tau)$ and*

$$M(\theta,\tau) = \lim_{T\to\infty}(1/T)\sum_{1}^{T}E(\partial/\partial\theta')m_t(\theta,\tau)$$

each exist uniformly over $\Theta_0 \times \mathscr{T}$ *and are continuous at* (θ_0, τ_0) *with respect to some pseudo-metric on* $\Theta_0 \times \mathscr{T}$ *for which* $(\hat{\theta}, \hat{\tau}) \xrightarrow{p} (\theta_0, \tau_0)$. $\overline{\lim}_{T\to\infty} (1/T)\Sigma_1^T EB_t < \infty$.

(h) $M'DM$ *is nonsingular.*

THEOREM 1: *Under Assumption N, every sequence of MINPIN estimators* $\{\hat{\theta}\}$ *satisfies*

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(\mathbf{0}, V).$$

The proofs of Theorem 1 and other results stated below are given in the Appendix.

COMMENT: If $d(m, \gamma)$ is of the form $m'\gamma m/2$, then Assumption N(g) can be relaxed. In particular, $m_t(\theta, \tau)$ only needs to be once, not twice, continuously differentiable in θ on Θ_0 and the condition on EB_t can be dropped in Assumption N(g). The proof of Theorem 1 in this case is altered along the lines of the proof of Theorem 1 in Andrews (1993). That is, one does mean value expansions of $\sqrt{T}\,\overline{m}_T(\hat{\theta}, \hat{\tau})$ about θ_0 rather than of $\sqrt{T}(\partial/\partial m)d(\overline{m}_T(\hat{\theta}, \hat{\tau}), \hat{\gamma})$ about θ_0.

4.3. *Discussion of Assumption N*

Assumption N(a) can be established by Theorem A-1 given in the Appendix or some other consistency proof. Assumptions N(b), (c), and (e) are key assumptions—they are discussed below. Assumption N(d) can be verified using a CLT for a sequence of rv's, such as the Lindeberg-Lévy CLT or one of the CLTs given by McLeish (1975b, 1977), Hall and Heyde (1980, Chs. 3–5), Herrndorf (1984), Gallant (1987), or Wooldridge and White (1988a, b). Assumption N(f) usually is not restrictive and is easy to verify.

Assumption N(g) requires that $\{m_t(\theta, \tau)\}$ and $\{(\partial/\partial\theta')m_t(\theta, \tau)\}$ satisfy uniform WLLNs over $\Theta_0 \times \mathscr{T}$ (i.e., $\sup_{\theta\in\Theta_0, \tau\in\mathscr{T}} \|\overline{m}_T(\theta, \tau) - \overline{m}_T^*(\theta, \tau)\| \xrightarrow{p} 0$ etc.). These conditions can be verified using stochastic equicontinuity results, such as those given in Andrews (1989b, 1991a, 1994a). Alternatively, it can be verified using the generic uniform WLLN results of Andrews (1987, 1992), Pötscher and Prucha (1989), or Newey (1991) combined with a pointwise WLLN, such as that of Andrews (1988) or McLeish (1975a) for dnid rv's. As a third alternative, it can be verified using empirical process WLLN results; see Pollard (1984).

Assumption N(g) also requires continuity of $m(\theta, \tau)$ and $M(\theta, \tau)$ with respect to some pseudo-metrics on $\Theta_0 \times \mathscr{T}$ for which $(\hat{\theta}, \hat{\tau}) \xrightarrow{p} (\theta_0, \tau_0)$. The most convenient choices of pseudo-metrics are

$$(4.5) \qquad \rho((\theta_1, \tau_1), (\theta_2, \tau_2)) = \overline{\lim_{N\to\infty}} \frac{1}{N}\sum_1^N E\left\|m_t(\theta_1, \tau_1) - m_t(\theta_2, \tau_2)\right\| \quad \text{and}$$

$$(4.6) \qquad \rho((\theta_1, \tau_1), (\theta_2, \tau_2)) = \overline{\lim_{N\to\infty}} \frac{1}{N}\sum_1^N E\left\|\frac{\partial}{\partial\theta'}m_t(\theta_1, \tau_1) - \frac{\partial}{\partial\theta'}m_t(\theta_2, \tau_2)\right\|$$

for establishing continuity of $m(\theta, \tau)$ and $M(\theta, \tau)$ respectively.[4] With these choices, continuity of $m(\theta, \tau)$ and $M(\theta, \tau)$ automatically holds, so it suffices to verify that $\rho((\hat{\theta}, \hat{\tau}), (\theta_0, \tau_0)) \xrightarrow{p} 0$ for each choice of $\rho(\cdot, \cdot)$. The latter usually holds under similar assumptions to those used to verify the condition of Assumption N(b) that $\rho_{\mathscr{T}}(\hat{\tau}, \tau_0) \xrightarrow{p} 0$. The requirement of Assumption N(g) that $m_t(\theta, \tau)$ is twice differentiable in θ can be relaxed, if necessary, and replaced by the assumption that $Em_t(\theta, \tau)$ is twice differentiable in θ; see Andrews (1989a) for details. Assumption N(h) is a standard condition that is often closely related to the identification of θ_0. It reduces to nonsingularity of the information matrix in iid ML contexts.

The stochastic equicontinuity assumption, Assumption N(e), can be verified using the Proposition stated in Section 2, using results given in Andrews (1991a, 1994a), or by using other results in the literature. To obtain stochastic equicontinuity, the index set $\mathscr{T}$ needs to satisfy some conditions. This creates a tension between Assumption N(e) and the first part of Assumption N(b), since the more restricted is $\mathscr{T}$, the more difficult it is to show that $P(\hat{\tau} \in \mathscr{T}) \to 1$. For example, if $\mathscr{T}$ is an infinite dimensional class of functions, the Proposition and the stochastic equicontinuity results of Andrews (1991a, 1994a) require the functions in $\mathscr{T}$ and in consequence the functions $m(w, \theta_0, \tau)$ to satisfy smoothness conditions (as functions of w). When $\mathscr{T}$ is defined as such, one has to show that $\hat{\tau}$ also satisfies these smoothness conditions wp $\to 1$ to verify the first part of Assumption N(b). It will if τ_0 is a function of x for $x \in \mathscr{X}$, τ_0 satisfies the smoothness conditions of the Proposition or of Andrews (1991a, 1994a), and $\hat{\tau}$ and a suitable number of its derivatives converge in probability uniformly over $x \in \mathscr{X}$ to τ_0 and its corresponding derivatives. Note that uniform convergence of nonparametric regression estimators and their derivatives generally requires the domain $\mathscr{X}$ of the functions to be bounded and the absolutely continuous components of the distributions of the regressor variables $\{X_t\}$ to have densities bounded away from zero on $\mathscr{X}$.

Next we discuss the pseudo-metric $\rho_{\mathscr{T}}$ on $\mathscr{T}$. There is a tradeoff between Assumptions N(b) and (e) with regard to the choice of $\rho_{\mathscr{T}}$. The stronger is the $\rho_{\mathscr{T}}$, the easier it is to verify stochastic equicontinuity, but the more difficult it is to verify that $\rho_{\mathscr{T}}(\hat{\tau}, \tau_0) \xrightarrow{p} 0$. For the stochastic equicontinuity results of Andrews (1994a), the following pseudo-metrics are considered:

$$(4.7) \qquad \rho_{\mathscr{T}}(\tau_1, \tau_2) = \sup_{N \geqslant 1} \left[\frac{1}{N} \sum_1^N E \left\| m_t(\theta_0, \tau_1) - m_t(\theta_0, \tau_2) \right\|^2 \right]^{1/2} \quad \text{and}$$

$$(4.8) \qquad \rho_{\mathscr{T}}(\tau_1, \tau_2) = \left[\int_{\mathscr{W}} \left\| m(w, \theta_0, \tau_1) - m(w, \theta_0, \tau_2) \right\|^2 dw \right]^{1/2}.$$

The latter applies when $m_t(\cdot, \cdot, \cdot)$ does not depend on T or t and W_t takes values in a bounded set $\mathscr{W}$. Consistency of $\hat{\tau}$ for τ_0 with respect to a pseudo-

[4]Here and below, pseudo-metrics $\rho(\cdot, \cdot)$ are defined using a dummy variable N (rather than T) to avoid confusion when we consider objects such as $\text{plim}_{T \to \infty} \rho(\hat{\tau}, \tau_0)$. Note that the pseudo-metrics are assumed to be independent of the sample size T.

metric such as that of (4.7) or (4.8) (as required by Assumption N(b)) can usually be reduced to L^Q-consistency of $\hat{\tau}$ for some $2 \leqslant Q \leqslant \infty$ when $\mathscr{T}$ is an infinite dimensional class of functions, plus some moment conditions on certain functions of W_t and sometimes some uniform boundedness condition on τ_0 and $\hat{\tau}$ that must hold wp $\to 1$.

Assumption N(c) is a key assumption. It is an asymptotic orthogonality condition between the estimators $\hat{\theta}$ and $\hat{\tau}$. It is needed to show that preliminary estimation of τ_0 does not affect the asymptotic distribution of $\hat{\theta}$. If $d(m,\gamma) = m'm/2$ or $m'\gamma m/2$ and $\hat{\gamma} = O_p(1)$, then Assumption N(c) reduces to

$$(4.9) \qquad \sqrt{T}\,\overline{m}_T^*(\theta_0, \hat{\tau}) \xrightarrow{p} \mathbf{0}.$$

Note that $\overline{m}_T^*(\theta_0, \tau_0) = \mathbf{0}$ because these are the population first order conditions for the estimator $\hat{\theta}$ and θ_0 is an interior point. Thus, condition (4.9) requires that the replacement of τ_0 by $\hat{\tau}$ in $\overline{m}_T^*(\theta_0, \tau)$ have an effect that is at most $o_p(T^{-1/2})$.

Condition (4.9) (and hence, Assumption N(c)) is trivial to verify whenever

$$(4.10) \qquad \overline{m}_T^*(\theta_0, \tau) = \mathbf{0} \quad \forall \tau \text{ in some neighborhood of } \tau_0$$

for all T sufficiently large. The reason is, in this case,

$$(4.11) \qquad \begin{aligned} \sqrt{T}\,\overline{m}_T^*(\theta_0, \hat{\tau}) &= \sqrt{T}\,\overline{m}_T^*(\theta_0, \tau) 1(\rho_{\mathscr{T}}(\tau, \tau_0) < \varepsilon)\big|_{\tau = \hat{\tau}} \\ &\quad + \sqrt{T}\,\overline{m}_T^*(\theta_0, \hat{\tau}) 1(\rho_{\mathscr{T}}(\hat{\tau}, \tau_0) \geqslant \varepsilon) \\ &= \mathbf{0} + o_p(1) \end{aligned}$$

for some $\varepsilon > 0$, using (4.10) and Assumption N(b). It is easy to see that (4.10) holds for the GMM estimator of the conditional moments restriction model, which is not an adaptive model. Condition (4.10) also holds for most adaptive estimators of adaptive models, such as Bickel's (1982) and Manski's (1984) adaptive estimators of linear and nonlinear regression models with errors of unknown distribution.

On the other hand, Assumption N(c) and (4.9) do not require (4.10) to hold. For example, if $m_t(\theta, \tau)$ is of the form $m_t(\theta, \tau(X_t))$ for $\tau(X_t) \in R^s$, then (by a Taylor expansion) (4.9) holds under the following conditions:

$$(4.12) \qquad \begin{aligned} &Em_t(\theta_0, \tau_0(X_t)) = 0 \quad \forall t, \\ &E\left[\frac{\partial}{\partial \tau} m_t(\theta_0, \tau_0(X_t)) \bigg| X_t = x\right] = 0 \quad \forall x \; \forall t, \quad \text{and} \end{aligned}$$

$$(4.13) \qquad \begin{aligned} &\frac{1}{T}\sum_1^T \int T^{1/4}(\hat{\tau}(x) - \tau_0(x))' \frac{\partial^2}{\partial \tau \partial \tau'} m_t(\theta_0, \tau_t^*(x)) T^{1/4} \\ &\qquad \times (\hat{\tau}(x) - \tau_0(x))\, dP_t(x) \xrightarrow{p} 0, \end{aligned}$$

where $\tau_t^*(x)$ lies between $\hat{\tau}(x)$ and $\tau_0(x)$ and P_t denotes the distribution of X_t. The WLS estimator of the PPR model satisfies (4.13), as do the estimators of Klein and Spady (1993), Ichimura and Lee (1990), and Powell (1987) among others. Thus, if the estimators $\hat{\tau}(x)$ of $\tau_0(x)$ satisfy the $T^{1/4}$-consistency condition (4.13) in these examples, then Assumption N(c) holds.

When Assumption N(c) holds, a MINPIN estimator $\hat{\theta}$ has the same asymptotic distribution as the estimator that minimizes the same criterion function as $\hat{\theta}$ but with $\hat{\tau}$ replaced by τ_0. If, in addition, the latter estimator is an asymptotically efficient estimator of θ_0 for the case where θ_0 is the only unknown parameter in the problem, then $\hat{\theta}$ is an adaptive estimator. The latter condition only holds in special cases and, in consequence, Assumption N(c) is a much weaker requirement than is adaptability. In addition, as noted above, adaptability is not a necessary condition for Assumption N(c) to hold trivially via (4.10).

We now consider the case where Assumption N(c) fails and, hence, the estimation of τ_0 has an effect on the asymptotic distribution of the MINPIN estimator; also see the discussion of Newey (1990b, Sec. 4.4). Some examples include Cox's (1975) partial likelihood estimator of the proportional hazards model, Han's (1987) maximum rank correlation estimator of the generalized regression model, Horowitz's (1988) M-estimator of the censored regression model, and Powell, Stock, and Stoker's (1989) estimator of weighted average derivatives and index regression models. To fit such estimators into the framework developed here, one needs to present conditions under which $\sqrt{T}\,\overline{m}_T^*(\theta_0, \hat{\tau})$ is asymptotically normal jointly with $\nu_T(\tau_0)$ rather than $o_p(1)$, when $d(m, \gamma) = m'\gamma m/2$. This is done in Andrews (1991c) for the case where $\hat{\tau}$ consists of nonparametric kernel density and/or regression function estimators and/or their derivatives.

Lastly, we note that the principle difference between Assumption N and assumptions commonly used to establish asymptotic normality of nonlinear *parametric* estimators is the appearance of Assumptions N(c) and (e).

4.3. *Covariance Matrix Estimation*

We now consider estimation of the covariance matrix V of the MINPIN estimators $\{\hat{\theta}\}$. We estimate V by plugging estimates of D, M, and S into the expression for V given in (4.1). Define

$$(4.14) \qquad \hat{D} = \frac{\partial^2}{\partial m\, \partial m'} d\big(\overline{m}_T(\hat{\theta}, \hat{\tau}), \hat{\gamma}\big) \quad \text{and} \quad \hat{M} = \frac{1}{T}\sum_1^T \frac{\partial}{\partial \theta'} m_t(\hat{\theta}, \hat{\tau}).$$

Under Assumption N, $\hat{D}$ and $\hat{M}$ are consistent for D and M respectively.

Next, we discuss estimation of the matrix S. Let $\hat{S}$ be an estimator of S. If $\{m_t(\theta_0, \tau_0)\}$ is a sequence of independent or orthogonal rv's, then we can take

$$(4.15) \qquad \hat{S} = \frac{1}{T}\sum_1^T m_t(\hat{\theta}, \hat{\tau}) m_t(\hat{\theta}, \hat{\tau})'.$$

In this case, $\hat{S}$ is consistent for S if Assumption N holds with $(\partial/\partial\theta')m_t(\theta, \tau)$ replaced by $m_t(\theta, \tau)m_t(\theta, \tau)'$ in Assumption N(g).

If $\{m_t(\theta_0, \tau_0)\}$ is m-dependent, then the following estimator can be used:

$$(4.16) \qquad \hat{S} = \frac{1}{T}\sum_1^T \hat{m}_t \hat{m}_t' + \sum_{v=1}^m \frac{1}{T}\sum_{1+v}^T \left[\hat{m}_t \hat{m}_{t-v}' + \hat{m}_{t-v}\hat{m}_t'\right],$$

where $\hat{m}_t = m_t(\hat{\theta}, \hat{\tau})$. If $\{m_t(\theta_0, \tau_0)\}$ is neither orthogonal nor m-dependent, then a more complicated estimator of S is required. In particular, heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimators that have been defined for parametric models can be used; see White (1984, pp. 147–161), Newey and West (1987), Gallant (1987, pp. 552, 553, 573), Andrews (1991b), and Andrews and Monahan (1992). For semiparametric models these estimators can be defined in exactly the same way as for parametric models, using $\{m_t(\hat{\theta}, \hat{\tau})\}$ as the underlying rv's. The consistency of such HAC estimators when τ is infinite dimensional does not follow from the results given in any of the papers above, however, because these results make use of mean value or Taylor expansions in the estimated parameters, which rely on the finite dimensional character of the parameters. Nevertheless, if $\hat{\tau}$ affects the summand $m_t(\hat{\theta}, \hat{\tau})$ only through the value of a finite dimensional random vector $\hat{\tau}(X_t)$, say, then the same mean value expansion method can be used to establish consistency of the HAC estimator $\hat{S}$.

5. WLS / PARTIALLY PARAMETRIC REGRESSION APPLICATION

In this section, we apply Theorem 1 of Section 4 and Theorem A-1 of the Appendix to yield conditions under which the WLS estimator of the PPR model is consistent and asymptotically normal. We consider strong mixing nonidentically distributed bounded rv's. Kernel estimators are used to estimate $\tau_{10}(x)$, $\tau_{20}(x, \theta)$, and $\tau_{30}(x)$:

$$(5.1) \qquad \hat{\tau}_1(x) = \left(\frac{1}{T} \sum_1^T Y_t \hat{K}_1 \left[\frac{x - X_t}{\hat{\sigma}_{1T}} \right] \Big/ \hat{\sigma}_{1T} \right) \Big/ \hat{f}_1(x),$$

$$\hat{\tau}_2(x, \theta) = \left(\frac{1}{T} \sum_1^T h(Z_t, \theta) \hat{K}_2 \left[\frac{x - X_t}{\hat{\sigma}_{2T}} \right] \Big/ \hat{\sigma}_{2T}^k \right) \Big/ \hat{f}_2(x),$$

$$\hat{\tau}_3(x) = \left(\frac{1}{T} \sum_1^T \hat{U}_t^2 \hat{K}_3 \left[\frac{x - X_t}{\hat{\sigma}_{3T}} \right] \Big/ \hat{\sigma}_{3T} \right) \Big/ \hat{f}_3(x),$$

$$\hat{f}_j(x) = \frac{1}{T} \sum_1^T \hat{K}_j \left[\frac{x - X_t}{\hat{\sigma}_{jT}} \right] \Big/ \hat{\sigma}_{jT},$$

$$\hat{U}_t = Y_t - \hat{\tau}_1(X_t) - h(Z_t, \theta^*) + \hat{\tau}_2(X_t, \theta^*),$$

$$\hat{K}_j(x) = \det{}^{-1/2}\left(\hat{\Sigma}_j\right) K_j\left(\hat{\Sigma}_j^{-1/2} x\right),$$

$$\hat{\Sigma}_j = \frac{1}{T} \sum_1^T (X_t - \bar{X}_T)(X_t - \bar{X}_T)', \quad \text{and}$$

$$\bar{X}_T = \frac{1}{T} \sum_1^T X_t \quad \text{for} \quad j = 1, 2, 3,$$

where $K_j(\cdot)$ and $\hat{\sigma}_{jT}$ for $j = 1,2,3$ are kernels and bandwidth parameters that satisfy the conditions specified below and θ^* is the semiparametric LS estimator of the PPR model that is defined in the same way as the WLS estimator except that $\hat{\tau}_3(x) \equiv 1$.

We make the following definitions. Let Θ_0 be some neighborhood of θ_0. Let Θ^* be some open set that contains Θ. Let $\mathscr{X}_\varepsilon^*$ be an ε-neighborhood of $\mathscr{X}^*$ for some $\varepsilon > 0$. For large $B < \infty$ and small $\delta > 0$, let

$$(5.2)\qquad \mathscr{T} = \Big\{(\tau_1, \tau_2, \tau_3) \colon \|\tau_1(\cdot)\|_{q_1, \mathscr{X}_\varepsilon^*} \leqslant B, \|\tau_2(\cdot,\cdot)\|_{q_2, \mathscr{X}_\varepsilon^* \times \Theta^*} \leqslant B,$$
$$\|\tau_{30}(\cdot)\|_{q_3, \mathscr{X}^*} \leqslant B, \inf_{x \in \mathscr{X}^*} |\tau_3(x)| \geqslant \delta \Big\},$$

where q_1, q_2, and q_3 are positive integers that satisfy

$$(5.3)\qquad q_1 > (k + l)/2, \qquad q_2 > (k + \max\{p, l, 2\})/2, \quad \text{and} \quad q_3 > k/2$$

and k, l, and p are the dimensions of X_t, Z_t, and θ_0 respectively. (The constants B and δ above are chosen sufficiently large and small, respectively, that $(\tau_{10}, \tau_{20}, \tau_{30}) \in \mathscr{T}$.) Let $\tau_0^* = (\tau_{10}, \tau_{20}, \iota)$, where $\iota(x) = 1$ for all $x \in R^k$. τ_0^* is the limit of the nonparametric estimator $\hat{\tau}$ when $\hat{\tau}_3(\cdot)$ is replaced by $\iota(\cdot)$, which occurs in the definition of the nonweighted LS estimator θ^*.

Next, define

$$(5.4)\qquad m_a(\theta, \tau) = \lim_{T \to \infty} \frac{1}{T} \sum_1^T E\xi(X_t) \times (Y_t - \tau_1(X_t) - h(Z_t, \theta) + \tau_2(X_t, \theta))^2 / \tau_3(X_t),$$

$$m_b(\theta, \tau) = \lim_{T \to \infty} \frac{1}{T} \sum_1^T E m_t(\theta, \tau), \quad \text{and}$$

$$M(\theta, \tau) = \lim_{T \to \infty} \frac{1}{T} \sum_1^T E \frac{\partial}{\partial \theta'} m_t(\theta, \tau),$$

where $m_t(\theta, \tau)$ is as in (3.7).

Next, let μ denote a k-vector of nonnegative integers and define $x^\mu = \prod_{l=1}^k x_l^{\mu_l}$ for $x \in R^k$. For nonnegative integers k, q, and ω with $\omega \geqslant q$, we define the following class of kernels:

$$(5.5)\qquad \mathscr{K}_{k,q,\omega} = \Big\{K(\cdot) \colon R^k \to R \,\Big|\, \int K(x)\, dx = 1,$$

$$\int x^\mu K(x)\, dx = 0 \ \forall 1 \leqslant |\mu| \leqslant \omega - q - 1,$$

$$\int |x^\mu K(x)|\, dx < \infty \ \forall \mu \text{ with } |\mu| = \omega - q,$$

$$D^\mu K(x) \to 0 \text{ as } \|x\| \to \infty \ \forall \mu \text{ with } |\mu| < q,$$

$D^\mu K(x)$ is absolutely integrable and has Fourier transform

$$\Psi_\mu(r) = (2\pi)^k \int \exp(ir'x) D^\mu K(x)\, dx \text{ that satisfies}$$

$$\int (1+\|r\|) \sup_{b \geqslant 1} \left| \Psi_\mu(br) \right| dr < \infty \; \forall \mu \text{ with } |\mu| \leqslant q,$$

$$\text{and } \sup_{x \in R^k} \left| D^{\mu+e_j} K(x) \right| (\|x\| \vee 1) < \infty \; \forall \mu \text{ with } |\mu| \leqslant q,$$

$\forall j = 1, \ldots, k$, where e_j is the jth elementary

k-vector and $i = \sqrt{-1}\Big\}$.

For example, the following kernels considered by Bierens (1987) are in $\mathscr{K}_{k,q,\omega}$:

$$(5.6) \qquad K(x) = (2\pi)^{-k/2} \sum_{r=1}^{R} a_r b_r^{-k} \exp\left[-\tfrac{1}{2} x'x / b_r^2\right],$$

where R is a positive integer greater than or equal to $(\omega - q)/2$ and $\{(a_r, b_r)$: $r \leqslant R\}$ are constants that satisfy

$$(5.7) \qquad b_r > 0, \quad \sum_{r=1}^{R} a_r = 1, \quad \text{and} \quad \sum_{r=1}^{R} a_r b_r^{2l} = 0 \quad \text{for} \quad l = 1, \ldots, R-1.$$

For $\omega = q$ and $R = 1$, this gives the standard normal kernel of dimension k.

Let $\mathscr{K}^*_{k,q,\omega}$ be the class of kernels $K(\cdot)$ that (i) satisfy the conditions of $\mathscr{K}_{k,q,\omega}$ except (possibly) the condition on the Fourier transform of $D^\mu K(x)$ and (ii) are zero outside a bounded set in R^k. For example, the kernel $K(x) = \exp(1/(\|x\|^2 - 1))$ for $\|x\| < 1$ and $K(x) = 0$ for $\|x\| > 1$ is in $\mathscr{K}^*_{k,q,q+1}$ for all positive integers q.

When applied to a function $g(x, \theta)$, we let D^{μ_x} and D^{μ_θ} denote vectors of derivatives with respect to x and θ respectively. We say mixing numbers $\{\alpha(s)$: $s \geqslant 1\}$ are of size $-\beta$ if $\alpha(s) = O(s^{-\beta-\epsilon})$ for some $\varepsilon > 0$.

The following assumption is sufficient for asymptotic normality of $\hat{\theta}$.

ASSUMPTION WLS / PPR: (a) *Θ is bounded, θ_0 lies in the interior of Θ, and $\mathscr{X}^*$ is an open bounded subset of R^k with minimally smooth boundary.*

(b) *$\{(U_t, X_t, Z_t)$: $t \geqslant 1\}$ is a sequence of strong mixing rv's with mixing numbers of size -2. (Y_t, X_t) lies in an open bounded set with minimally smooth boundary $\forall t \geqslant 1$. (Y_t, Z_t) and (X_t, Z_t) do likewise.*

(c) *$m_a(\theta, \tau)$, $m_b(\theta, \tau)$, and $M(\theta, \tau)$ exist uniformly over (θ, τ) in $\Theta \times \mathscr{T}$, $\Theta_0 \times \mathscr{T}$, and $\Theta_0 \times \mathscr{T}$, respectively. The matrices M and S exist.*

(d) $\sup_{\theta \in \Theta^*} \sup_{z \in \mathscr{Z}} \|D^{\mu_z} D^{\mu_\theta} h(z, \theta)\| < \infty \; \forall \mu_z, \mu_\theta$ *with* $|\mu_z| \leqslant q_1$ *and* $|\mu_\theta| \leqslant q_2 + 1$, *where $\mathscr{Z}$ contains the z-values in the open bounded set of part* (b) *that contains (X_t, Z_t) $\forall t$.*

(e) *M is nonsingular and $m_a(\theta, \tau_0)$ and $m_a(\theta, \tau_0^*)$ are bounded away from zero for all θ outside any given neighborhood of θ_0.*

(f) *The distribution of* X_t *is absolutely continuous with respect to Lebesgue measure with density* $f_t(x)$ $\forall t \geqslant 1$, $\inf_{T \geqslant 1} \inf_{x \in \mathscr{X}_\varepsilon^*} (1/T)\Sigma_1^T f_t(x) > 0$, $D^\mu f_t(x)$ *exists and is continuous on* R^k $\forall t \geqslant 1$, *and* $\sup_{T \geqslant 1} \sup_{x \in R^k} |(1/T)\Sigma_1^T D^\mu f_t(x)| < \infty$ $\forall \mu$ *with* $|\mu| \leqslant \max\{\omega_1, \omega_2, q_2\}$, *where* ω_1 *and* ω_2 *are positive integers that also appear in parts* (g)–(i) *below.*

(g) $\tau_{10}(x) = E(Y_t | X_t = x)$, $D^{\mu_\theta}\tau_{20}(x,\theta) = D^{\mu_\theta}E(h(Z_t,\theta)|X_t = x)$, $\tau_{30}(x) = E(U_t^2 | X_t = x)$, *and* $\tau_{40}(x,\theta) = E([h(Z_t,\theta) - h(Z_t,\theta_0)]^2 | X_t = x)$ *do not depend on* t $\forall t \geqslant 1$, $\forall \theta \in \Theta^*$, $\forall \mu_\theta$ *with* $|\mu_\theta| \leqslant q_2$. $D^\mu[\tau_{10}(x)f_t(x)]$ *exists and is continuous on* R^k $\forall t \geqslant 1$, $\{(1/T)\Sigma_1^T D^\mu[\tau_{10}(x)f_t(x)]: T \geqslant 1\}$ *is uniformly equicontinuous on* R^k, $\sup_{x \in \mathscr{X}_\varepsilon^*} |(1/T)\Sigma_1^T D^\mu[\tau_{10}(x)f_t(x)]| < \infty$, *and* $\sup_{x \in \mathscr{X}_\varepsilon^*} |D^\mu \tau_{10}(x)| < \infty$ $\forall \mu$ *with* $|\mu| \leqslant \max\{q_1, \omega_1\}$. $D^{\mu_x}D^{\mu_\theta}[\tau_{20}(x,\theta)f_t(x)]$ *exists and is continuous in* x *on* R^k $\forall \theta \in \Theta^*$ $\forall t > 1$, $\sup_{T \geqslant 1} \sup_{\theta \in \Theta^*} \sup_{x \in R^k} |(1/T)\Sigma_1^T D^{\mu_x}D^{\mu_\theta}[\tau_{20}(x,\theta)f_t(x)]| < \infty$, *and* $\sup_{\theta \in \Theta^*} \sup_{x \in \mathscr{X}_\varepsilon^*} |D^{\mu_x}D^{\mu_\theta}\tau_{20}(x,\theta)| < \infty$ $\forall \mu_x$, μ_θ *with* $|\mu_x| \leqslant q_2 + 1$ *and* $|\mu_\theta| \leqslant q_2$. $\sup_{x \in \mathscr{X}^*} |D^{\mu_x}\tau_{20}(x,\theta_0)| < \infty$ $\forall \mu_x$ *with* $|\mu_x| \leqslant \omega_2$. $D^\mu \tau_{20}(x,\theta)$ *is continuous in* θ *at* θ_0 *uniformly over* $x \in \mathscr{X}^*$ $\forall \mu$ *with* $|\mu| \leqslant q_3$. $D^\mu[\tau_{30}(x)f_t(x)]$ *exists and is continuous on* R^k $\forall t \geqslant 1$, $\sup_{T \geqslant 1} \sup_{x \in R^k} |(1/T)\Sigma_1^T D^\mu[\tau_{30}(x)f_t(x)]| < \infty$, *and* $\sup_{x \in \mathscr{X}_\varepsilon^*} |D^\mu \tau_{30}(x)| < \infty$ $\forall \mu$ *with* $|\mu| \leqslant q_3 + 1$. $D^\mu[\tau_{40}(x,\theta)f_t(x)]$ *exists and is continuous on* R^k $\forall \theta \in \Theta_0$ $\forall t \geqslant 1$,

$$\sup_{T \geqslant 1} \sup_{\theta \in \Theta_0} \sup_{x \in \mathscr{X}_\varepsilon^*} \left|(1/T)\sum_1^T D^\mu[\tau_{40}(x,\theta)f_t(x)]\right| < \infty,$$

$D^\mu \tau_{40}(x,\theta)$ *is continuous at* θ_0 *uniformly over* $x \in \mathscr{X}^*$, *and*

$$\sup_{\theta \in \Theta_0} \sup_{x \in \mathscr{X}_\varepsilon^*} |D^\mu \tau_{40}(x,\theta)| < \infty \;\forall \mu \quad \textit{with } |\mu| \leqslant q_3 + 1.$$

(h) $K_1 \in \mathscr{K}_{k,q_1,q_1} \cap \mathscr{K}_{k,0,\omega_1}$. $K_2 \in \mathscr{K}_{k,q_2,q_2+1} \cap \mathscr{K}_{k,0,\omega_2}$ *and*

$$\sup_{x \in R^k} |D^\mu K_2(x)| < \infty \;\forall \mu \quad \textit{with } |\mu| \leqslant q_1 + q_2.$$

$K_3 \in \mathscr{K}_{k,q_3,q_3+1}^*$ *and* $\sup_{x \in R^k} |D^\mu K_3(x)| < \infty$ $\forall \mu$ *with* $|\mu| \leqslant q_3$.

(i) *For* $j = 1,2,3$, $\{\hat{\sigma}_{jT}: T \geqslant 1\}$ *satisfies* $C_{j1}\sigma_{j1T} \leqslant \hat{\sigma}_{jT} \leqslant C_{j2}\sigma_{j2T}$ *wp* $\to 1$ *for some positive bounded constants* $\{(\sigma_{j1T}, \sigma_{j2T}): T \geqslant 1\}$, C_{j1}, *and* C_{j2} *that satisfy*

(1) $$\sigma_{12T} = o(T^{-1/(4\omega_1)}) \quad \textit{and} \quad T^{\min\{1/(2k+2q_1),\,1/(4k)\}}\sigma_{11T} \to \infty,$$

(2) $$\sigma_{22T} = o(T^{-1/(4\omega_2)}) \quad \textit{and} \quad T^{\min\{1/(3k+2q_2+l),\,1/(4k)\}}\sigma_{21T} \to \infty, \quad \textit{and}$$

(3) $$\sigma_{32T} = o(1) \quad \textit{and} \quad T^{1/(2k+2q_3)}\sigma_{31T} \to \infty.$$

(j) $\lim_{T \to \infty}(1/T)\Sigma_1^T EX_t$ *exists*, $\lim_{T \to \infty}(1/T)\Sigma_1^T EX_t X_t'$ *exists, and*

$$\lim_{T \to \infty} \lambda_{\min}\left(\frac{1}{T}\sum_1^T E(X_t - E\bar{X}_T)(X_t - E\bar{X}_T)'\right) > 0.$$

We now comment on Assumption WLS/PLR. Regarding WLS/PPR(a), see Stein (1970, pp. 181, 189) for the definition of minimally smooth boundary. Examples of sets in R^k with minimally smooth boundaries include open bounded sets that are convex or whose boundaries are C^1-embedded in R^k. Finite unions of sets of the aforementioned type with disjoint closures also have minimally smooth boundaries. The boundedness assumption on $\mathscr{X}^*$ in WLS/PPR(a) is convenient due to three aspects of the problem that cause complications. First, nonlinearity of $h(Z_t, \theta)$ in θ means that we need to establish convergence of the nonparametric estimator $\hat{\tau}_2(x, \theta)$ and its derivatives with respect to x and θ *uniformly over* θ. Second, establishing consistency of the estimator $\hat{\tau}_3(x)$ of the weight function $\tau_{30}(x)$ involves establishing uniform consistency of nonparametric regression estimators over a nonparametric class of functions (τ_1, τ_2), since the residuals $\hat{U}_t^2$ depend on the nonparametric estimates $(\hat{\tau}_1, \hat{\tau}_2)$. Third, the observations are allowed to possess strong mixing temporal dependence.

Assumption WLS/PPR(c) holds automatically in the identically distributed case given the moment assumptions. WLS/PPR(e) is an identification condition that is used to obtain consistency of $\hat{\theta}$.

Assumption WLS/PPR(f) assumes that the regressors are continuous. This assumption is made to obtain the desired uniform consistency properties of the kernel estimators in a simple way. It can be relaxed along the lines given in Andrews (1994b, Comment 5 to Thm. 1) if X_{jt} contains a subvector of discrete rv's with finite support and a subvector of continuous rv's that satisfies the conditions of WLS/PPR(f). For brevity, we do not treat this case explicitly here. WLS/PPR(f) also requires the average density of $\{X_{jt}\colon t \leqslant T\}$ to be bounded away from zero on the set $\mathscr{X}_\varepsilon^*$. This condition could be relaxed, but not if one wants to verify Assumptions N and C (given in the Appendix) using uniform convergence of the nonparametric estimators, as is done here.

The use of bias-reducing kernels $K_1(\cdot)$ and $K_2(\cdot)$ (Assumption WLS/PPR(h)) is due to the need to establish $T^{1/4}$-convergence of $\hat{\tau}_2(x, \theta_0)$ and $(\partial/\partial\theta)\hat{\tau}_2(x, \theta_0)$ to $\tau_{20}(x, \theta_0)$ and $(\partial/\partial\theta)\tau_{20}(x, \theta_0)$ respectively. The latter results are needed in order to verify the asymptotic orthogonality condition Assumption N(c).

For the case where $h(Z_t, \theta_0)$ is linear in θ_0, the errors U_t are conditionally homoskedastic, and the observations are independent, Assumption WLS/PPR is much more restrictive than the assumptions given by Robinson (1988) for asymptotic normality of the nonweighted linear LS estimator of θ_0. This is due to two factors. First, Assumption WLS/PPR is designed for the more general model for which the above three conditions are relaxed. Second, Assumption WLS/PPR is formulated as a special case of the more general results given in Theorems 1 and A-1, whereas Robinson's assumptions are specially tailored to the partially linear regression model. As often occurs, the results designed on a case by case basis are sharper than those derived from a more general set of results.

Asymptotic normality of the WLS estimator $\hat{\theta}$ is given in the following result.

THEOREM 2: *Under Assumption WLS/PPR, the WLS estimator* $\hat{\theta}$ *satisfies* $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, M^{-1}SM^{-1})$ *for* M *and* S *as in* (4.3).

COMMENTS: 1. The nonweighted LS estimator θ^* is consistent under Assumption WLS/PPR. If Assumption WLS/PPR holds with M and S defined with $\tau_{30}(X_t)$ replaced by 1, then $\sqrt{T}(\theta^* - \theta_0) \xrightarrow{d} N(0, M^{-1}SM^{-1})$ with M and S so defined.

2. When the observations are iid and $E(U_t^2|Z_t, X_t) = E(U_t^2|X_t)$ a.s. (but the latter is not part of the prior restrictions on the model), then the WLS estimator $\hat{\theta}$ attains the semiparametric asymptotic efficiency bound given by Chamberlain (1992) (provided the model is assumed to hold only for $X_t \in \mathscr{X}^*$).

3. The asymptotic covariance matrix $M^{-1}SM^{-1}$ of $\hat{\theta}$ can be estimated by $\hat{M}^{-1}\hat{S}\hat{M}^{-1}$, where

$$\hat{M} = \frac{1}{T}\sum_1^T \xi(X_t)\left(\frac{\partial}{\partial\theta}h(Z_t,\hat{\theta}) - \frac{\partial}{\partial\theta}\hat{\tau}_2(X_t,\hat{\theta})\right) \times \left(\frac{\partial}{\partial\theta}h(Z_t,\hat{\theta}) - \frac{\partial}{\partial\theta}\hat{\tau}_2(X_t,\hat{\theta})\right)' \Big/ \hat{\tau}_3(X_t)$$

and $\hat{S}$ is as defined in Section 4.3 depending on the dependence and heterogeneity properties assumed of

$$\{\xi(X_t)U_t((\partial/\partial\theta)h(Z_t,\theta_0) - (\partial/\partial\theta)\tau_{20}(X_t,\theta_0))/\tau_{30}(X_t) : t \geqslant 1\}.$$

$\hat{M}$ is consistent for M and $\hat{S}$ is consistent for S under Assumption WLS/PPR when the definition of $\hat{S}$ given in (4.16) or (4.17) is appropriate. When $\hat{S}$ is a HAC estimator, consistency needs to be verified.

Cowles Foundation for Research in Economics, P.O. Box 208281 Yale Station, New Haven, CT 06520-208281, U.S.A.

Manuscript received May, 1989; final revision received May, 1993.

APPENDIX

A.1. *Consistency of MINPIN Estimators*

We now provide sufficient conditions for the consistency of the MINPIN estimator $\hat{\theta}$. Let Θ/Θ_0 denote the set of points θ that are in Θ, but are not in Θ_0.

ASSUMPTION C (Consistency): (a) *There exists a function* $m(\cdot,\cdot)$: $\Theta \times \mathscr{T} \to R^v$ *such that* $\overline{m}_T(\theta,\tau) \xrightarrow{p} m(\theta,\tau)$ *uniformly over* $(\theta,\tau) \in \Theta \times \mathscr{T}$.

(b) $\sup_{\theta\in\Theta}\|m(\theta,\hat{\tau}) - m(\theta,\tau_0)\| \xrightarrow{p} 0$ *for some* $\tau_0 \in \mathscr{T}$, $P(\hat{\tau} \in \mathscr{T}) \to 1$, *and* $\hat{\gamma} \xrightarrow{p} \gamma_0$ *for some* $\gamma_0 \in \Gamma$.

(c) $d(m,\gamma)$ *is uniformly continuous over* $\mathscr{M} \times \Gamma_0$, *where* $\mathscr{M} = \{m \in R^v: m = m(\theta,\tau)$ *for some* $\theta \in \Theta$, $\tau \in \mathscr{T}\}$ *and* $\Gamma_0(\subset \Gamma)$ *contains a neighborhood of* γ_0.

(d) *For every neighborhood* $\Theta_0(\subset\Theta)$ *of* θ_0, $\inf_{\theta\in\Theta/\Theta_0} d(m(\theta,\tau_0),\gamma_0) > d(m(\theta_0,\tau_0),\gamma_0)$.

THEOREM A-1: *Under Assumption C, every sequence of MINPIN estimators* $\{\hat{\theta}\}$ *satisfies* $\hat{\theta} \xrightarrow{p} \theta_0$ *under* **P**.

The proof of Theorem A-1 given below is similar to many other consistency proofs in the literature.

We now discuss Assumption C. The function $m(\theta,\tau)$ of Assumption $C(a)$ usually is given by $\lim_{T\to\infty}(1/T)\Sigma_1^T Em_t(\theta,\tau)$. Thus, Assumption C(a) holds if $m(\theta,\tau) = \lim_{T\to\infty}(1/T)\Sigma_1^T Em_t(\theta,\tau)$ exists uniformly over $\Theta \times \mathscr{T}$ (i.e., $\sup_{(\theta,\tau)\in\Theta\times\mathscr{T}}\|(1/T)\Sigma_1^T Em_t(\theta,\tau) - m(\theta,\tau)\| \to 0$) and $\{m_t(\theta,\tau): t \leqslant T, T \geqslant 1\}$ satisfies a uniform WLLN over $\Theta \times \mathscr{T}$. The latter can be verified in the same manner as for Assumption N(g). The first part of Assumption C(b) specifies the manner in which $\hat{\tau}$ must converge to τ_0. The condition shows that $\hat{\tau}$ must estimate τ_0 well only in so far as $m(\theta,\hat{\tau})$ estimates $m(\theta,\tau_0)$ well uniformly over $\theta \in \Theta$ for large T. When τ_0 is a function, the latter usually requires L^Q consistency of $\hat{\tau}(\cdot)$ for $\tau_0(\cdot)$ for some $1 \leqslant Q \leqslant \infty$. If $\hat{\tau}$ is a nonparametric regression or density estimator, then consistency results in the literature can be exploited when verifying the first part of Assumption C(b). In particular, the nonparametric kernel density and regression function estimation results of Andrews (1994b) are designed especially for use in semiparametric models with dnid observations.

Note that the establishment of the first part of Assumption C(b) may involve a step that is not treated in the nonparametric literature (although it is treated in Andrews (1994b) for kernel estimators). This step arises when $\hat{\tau}$ is based on estimated variables, such as residuals, rather than the true variables themselves. In such cases, one has to show that the error introduced by preliminary parameter estimation is $o_p(1)$. This can be done directly on a case by case basis, or by using the results of Andrews (1994b) when kernel estimators are employed, or by using a discretization/contiguity argument as in Bickel (1982, p. 657) and Manski (1984, pp. 173–178).

The second part of Assumption C(b) requires that wp $\to 1$ $\hat{\tau}$ lies in the set $\mathscr{T}$ over which $\overline{m}_T(\theta,\tau)$ converges uniformly to $m(\theta,\tau)$. There is a tension between this condition and Assumption C(a), since the larger $\mathscr{T}$ is the easier it is to verify this condition, but the more difficult it is to verify Assumption C(a) and vice versa. If Assumption C(a) is verified using a smoothness condition on all $\tau \in \mathscr{T}$, as is the case when the Proposition of Section 2 or the stochastic equicontinuity results of Andrews (1989b, 1991a, 1994a) are used, then Assumption C(b) requires that $\hat{\tau}$ satisfy this smoothness condition wp $\to 1$. See the discussion of Assumption N(b) for further details. Again, consistency results in the nonparametric literature or Andrews (1994b) can be used to verify such a condition. The third part of Assumption C(b) often holds trivially since no estimator $\hat{\gamma}$ arises. When $\hat{\gamma}$ does arise, it is almost always finite dimensional.

In almost all examples, $d(m,\gamma) = m$, $m'm/2$, or $m'\gamma m/2$. In these cases, a sufficient condition for Assumption C(c) is $\sup_{\theta \in \Theta, \tau \in \mathscr{T}} \|m(\theta,\tau)\| < \infty$, which is not overly restrictive. Assumption C(d) is the uniqueness/identification assumption that ensures that $\{\hat{\theta}\colon T \geqslant 1\}$ neither converges to a multi-element subset of Θ nor diverges to "∞." The same condition is often used for nonlinear parametric models. Sufficient conditions for Assumption C(d) are: Θ is compact and $d(m(\theta,\tau_0),\gamma_0)$ is continuous in θ on Θ and is uniquely minimized at θ_0.

A.2. Proofs

For notational simplicity, we let $\overline{m}_T(\theta)$ abbreviate $\overline{m}_T(\theta,\hat{\tau})$ and $m(\theta)$ abbreviate $m(\theta,\tau_0)$ in the remainder of the Appendix, except in those places where the dependence on $\hat{\tau}$ or τ_0 must be made explicit for reasons of clarity.

The proof of Theorem A-1 uses the following lemma, which is similar to results in the literature. The lemma appears in this form in Pötscher and Prucha (1989, Lemma 3.1) (with a different proof than that given below) and perhaps elsewhere in the literature.

LEMMA A-1: *Suppose $\hat{\theta}$ minimizes a random real function $Q_T(\theta)$ over $\theta \in \Theta$* wp $\to 1$, *where Θ is a pseudo-metric space. If (a) $\sup_{\theta \in \Theta} |Q_T(\theta) - Q(\theta)| \xrightarrow{p} 0$ for some real function Q on Θ and (b) for every neighborhood Θ_0 of θ_0, $\inf_{\theta \in \Theta/\Theta_0} Q(\theta) > Q(\theta_0)$, then $\hat{\theta} \xrightarrow{p} \theta_0$.*

PROOF OF LEMMA A-1: By Assumption (b), given any neighborhood Θ_0 of θ_0, there exists a constant $\delta > 0$ such that $\inf_{\theta \in \Theta/\Theta_0} Q(\theta) > Q(\theta_0) + \delta$. Thus,

$$P\big(\hat{\theta} \in \Theta/\Theta_0\big) \leqslant P\big(Q(\hat{\theta}) - Q(\theta_0) > \delta\big) \to 0, \tag{A.1}$$

where $\to 0$ holds provided $Q(\hat{\theta}) \xrightarrow{p} Q(\theta_0)$. The latter follows from

$$\begin{aligned} 0 &\leqslant Q(\hat{\theta}) - Q(\theta_0) = Q(\hat{\theta}) - Q_T(\hat{\theta}) + Q_T(\hat{\theta}) - Q(\theta_0) \\ &\leqslant Q(\hat{\theta}) - Q_T(\hat{\theta}) + Q_T(\theta_0) - Q(\theta_0) + o_p(1) \leqslant 2 \sup_{\theta \in \Theta} |Q_T(\theta) - Q(\theta)| + o_p(1) \xrightarrow{p} 0. \end{aligned} \tag{A.2}$$

Q.E.D.

PROOF OF THEOREM A-1: We show that Assumption C implies that conditions (a) and (b) of Lemma A-1 hold with $Q_T(\theta) = d(\overline{m}_T(\theta,\hat{\tau}),\hat{\gamma})$ and $Q(\theta) = d(m(\theta,\tau_0),\gamma_0)$. Condition (b) holds by

Assumption C(d). Condition (a) follows from

$$\text{(A.3)}\qquad \begin{aligned}&\sup_{\theta\in\Theta}\left|d(\overline{m}_T(\theta,\hat\tau),\hat\gamma)-d(m(\theta,\tau_0),\gamma_0)\right|\\
&\quad\leqslant \sup_{\theta\in\Theta}\left|d(\overline{m}_T(\theta,\hat\tau),\hat\gamma)-d(m(\theta,\hat\tau),\hat\gamma)\right|\\
&\qquad+\sup_{\theta\in\Theta}\left|d(m(\theta,\hat\tau),\hat\gamma)-d(m(\theta,\tau_0),\gamma_0)\right|\\
&\quad\leqslant \sup_{\substack{\theta\in\Theta,\tau\in\mathscr{T},\\ \gamma\in\Gamma_0}}\left|d(\overline{m}_T(\theta,\tau),\gamma)-d(m(\theta,\tau),\gamma)\right|+o_p(1)\\
&\qquad+\sup_{\theta\in\Theta}\left|d(m(\theta,\hat\tau),\hat\gamma)-d(m(\theta,\tau_0),\gamma_0)\right|\\
&\quad\overset{p}{\to}0,\end{aligned}$$

where "$\overset{p}{\to} 0$" holds using Assumptions C(a)–(c). *Q.E.D.*

Proof of Theorem 1: Element by element mean value expansions of $\sqrt{T}(\partial/\partial\theta)d(\overline{m}_T(\hat\theta),\hat\gamma)$ about θ_0 give: $\forall j=1,\dots,p$,

$$\text{(A.4)}\qquad \begin{aligned}o_p(1)&=\sqrt{T}\frac{\partial}{\partial\theta_j}d(\overline{m}_T(\hat\theta),\hat\gamma)\\
&=\sqrt{T}\frac{\partial}{\partial\theta_j}d(\overline{m}_T(\theta_0),\hat\gamma)+\frac{\partial^2}{\partial\theta'\partial\theta_j}d(\overline{m}_T(\theta^*),\hat\gamma)\sqrt{T}\left(\hat\theta-\theta_0\right),\end{aligned}$$

where θ^* is a rv that depends on j and lies on the line segment joining $\hat\theta$ and θ_0, and hence, $\theta^*\overset{p}{\to}\theta_0$. (See Jennrich (1969, Lemma 3) for the mean value theorem for random functions.) The first equality holds because $\hat\theta$ minimizes $d(\overline{m}_T(\theta),\hat\gamma)$ and $\hat\theta$ is in the interior of Θ wp$\to 1$ by Assumption N(a). The second equality actually only holds wp$\to 1$, since the mean value expansions require $\hat\theta\in\Theta_0$.

Below we show that

$$\text{(A.5)}\qquad \frac{\partial^2}{\partial\theta'\partial\theta_j}d(\overline{m}_T(\theta^*),\hat\gamma)=\frac{\partial^2}{\partial\theta'\partial\theta_j}d(m(\theta_0),\gamma_0)+o_p(1),$$

where $(\partial^2/\partial\theta\partial\theta')d(m(\theta_0),\gamma_0)=M'DM$, and

$$\text{(A.6)}\qquad \sqrt{T}\frac{\partial}{\partial\theta}d(\overline{m}_T(\theta_0),\hat\gamma)\overset{d}{\to}N(\mathbf{0},M'DSDM).$$

These results, equation (A.4), and the nonsingularity of $M'DM$ give

$$\text{(A.7)}\qquad \sqrt{T}\left(\hat\theta-\theta_0\right)=-(M'DM)^{-1}\frac{\partial}{\partial\theta}d(\overline{m}_T(\theta_0),\hat\gamma)+o_p(1)\overset{d}{\to}N(\mathbf{0},V).$$

To show (A.5), we proceed as follows:

$$\text{(A.8)}\qquad \begin{aligned}\frac{\partial^2}{\partial\theta_j\partial\theta_l}d(\overline{m}_T(\theta^*),\hat\gamma)&=\frac{\partial^2}{\partial\theta_j\partial\theta_l}\overline{m}_T(\theta^*)'\frac{\partial}{\partial m}d(\overline{m}_T(\theta^*),\hat\gamma)\\
&\quad+\frac{\partial}{\partial\theta_j}\overline{m}_T(\theta^*)'\frac{\partial^2}{\partial m\,\partial m'}d(\overline{m}_T(\theta^*),\hat\gamma)\frac{\partial}{\partial\theta_l}\overline{m}_T(\theta^*).\end{aligned}$$

By Assumptions N(a), (b), and (g),

$$\text{(A.9)}\qquad \begin{aligned}\left\|\overline{m}_T(\theta^*)-m(\theta_0)\right\|&\leqslant\left\|\overline{m}_T(\theta^*,\hat\tau)-\overline{m}_T^*(\theta^*,\hat\tau)\right\|\\
&\quad+\left\|\overline{m}_T^*(\theta^*,\hat\tau)-m(\theta^*,\hat\tau)\right\|+\left\|m(\theta^*,\hat\tau)-m(\theta_0,\tau_0)\right\|\overset{p}{\to}0.\end{aligned}$$

Using this result, the continuity of $(\partial/\partial m)d(m,\gamma)$ at $(m(\theta_0),\gamma_0)$ (Assumption N(f)), the Assumption N(b) that $\hat{\gamma} \xrightarrow{p} \gamma_0$, and the continuous mapping theorem, we get

$$\text{(A.10)} \qquad \frac{\partial}{\partial m} d(\overline{m}_T(\theta^*),\hat{\gamma}) \xrightarrow{p} \frac{\partial}{\partial m} d(m(\theta_0),\gamma_0) = \mathbf{0},$$

where the equality holds by Assumptions N(b), (c), (f), and (g). Using Assumption N(g) and Markov's inequality, it is straightforward to show that $(\partial^2/\partial\theta_j\partial\theta_l)\overline{m}_T(\theta^*) = O_p(1)$. This result and (A.10) imply that the first term of (A.8) is $o_p(1)$.

Similarly, the continuity of $(\partial^2/\partial m\,\partial m')d(m,\gamma)$ at $(m(\theta_0),\gamma_0)$ (Assumption N(f)), equation (A.9), $\hat{\gamma} \xrightarrow{p} \gamma_0$, and the continuous mapping theorem give

$$\text{(A.11)} \qquad \frac{\partial^2}{\partial m\,\partial m'} d(\overline{m}_T(\theta^*),\hat{\gamma}) \xrightarrow{p} \frac{\partial^2}{\partial m\,\partial m'} d(m(\theta_0),\gamma_0) \equiv D.$$

It follows from Assumptions N(a), (b), and (g) that $M = M(\theta_0,\tau_0)$ and

$$\text{(A.12)} \qquad \left\| \frac{\partial}{\partial\theta'}\overline{m}_T(\theta^*) - M \right\| \leqslant \left\| \frac{\partial}{\partial\theta'}\overline{m}_T(\theta^*) - M(\theta^*,\hat{\tau}) \right\| + \| M(\theta^*,\hat{\tau}) - M(\theta_0,\tau_0) \| \xrightarrow{p} 0.$$

Equations (A.11) and (A.12) imply that the second term of (A.8) equals $[M'DM]_{jl} + o_p(1)$, and hence, (A.5) is established.

To establish equation (A.6), we write

$$\text{(A.13)} \qquad \sqrt{T}\frac{\partial}{\partial\theta} d(\overline{m}_T(\theta_0),\hat{\gamma}) = \sqrt{T}\left[\frac{\partial}{\partial\theta'}\overline{m}_T(\theta_0)\right]' \frac{\partial}{\partial m} d(\overline{m}_T(\theta_0),\hat{\gamma})$$
$$= M'\sqrt{T}\frac{\partial}{\partial m} d(\overline{m}_T(\theta_0),\hat{\gamma}) + o_p(1)$$

using N(g) provided $\sqrt{T}(\partial/\partial m)d(\overline{m}_T(\theta_0),\hat{\gamma}) = O_p(1)$, as we now demonstrate.

By the mean value theorem, the jth element of $\sqrt{T}(\partial/\partial m)d(\overline{m}_T(\theta_0,\hat{\tau}),\hat{\gamma})$ can be expanded about $\overline{m}_T^*(\theta_0,\hat{\tau})$ to get:

$$\text{(A.14)} \qquad \sqrt{T}\frac{\partial}{\partial m_j} d(\overline{m}_T(\theta_0,\hat{\tau}),\hat{\gamma}) = \sqrt{T}\frac{\partial}{\partial m_j} d(\overline{m}_T^*(\theta_0,\hat{\tau}),\hat{\gamma})$$
$$+ \frac{\partial^2}{\partial m'\,\partial m_j} d(m^*,\hat{\gamma})\sqrt{T}\left(\overline{m}_T(\theta_0,\hat{\tau}) - \overline{m}_T^*(\theta_0,\hat{\tau})\right),$$

where m^* is on the line segment joining $\overline{m}_T(\theta_0,\hat{\tau})$ and $\overline{m}_T^*(\theta_0,\hat{\tau})$, and hence, $m^* \xrightarrow{p} m(\theta_0)$. (More precisely, (A.14) holds wp $\to$ 1.)

The first term of the right-hand side of (A.14) is $o_p(1)$ by Assumption N(c). Also, using Assumption N(f), $(\partial^2/\partial m'\,\partial m_j)d(m^*,\hat{\gamma}) = [D]_j + o_p(1)$, where $[D]_j$ denotes the jth row of D. Hence, if $\sqrt{T}(\overline{m}_T(\theta_0,\hat{\tau}) - \overline{m}_T^*(\theta_0,\hat{\tau})) = O_p(1)$, the above results and (A.14) yield

$$\text{(A.15)} \qquad \sqrt{T}\frac{\partial}{\partial m} d(\overline{m}_T(\theta_0,\hat{\tau}),\hat{\gamma}) = D\sqrt{T}\left(\overline{m}_T(\theta_0,\hat{\tau}) - \overline{m}_T^*(0_0,\hat{\tau})\right) + o_p(1).$$

The proof is complete once we show that

$$\text{(A.16)} \qquad \nu_T(\hat{\tau}) = \sqrt{T}\left(\overline{m}_T(\theta_0,\hat{\tau}) - \overline{m}_T^*(\theta_0,\hat{\tau})\right) \xrightarrow{d} N(\mathbf{0},S),$$

since this implies that (A.15) and (A.13) hold, which establishes (A.6).

Using Assumption N(d), one sees that (A.16) holds if $\nu_T(\hat{\tau}) - \nu_T(\tau_0) \xrightarrow{p} \mathbf{0}$. The latter follows from Assumptions N(b) and (e) by (2.10). *Q.E.D.*

PROOF OF THEOREM 2: The proof uses the following properties, which hold under Assumption WLS/PPR, of the nonparametric estimators $\hat{\tau}_1(x)$, $\hat{\tau}_2(x,\theta)$, and $\hat{\tau}_3(x)$:

$$\text{(A.17)} \qquad T^{1/4} \sup_{x \in \mathscr{X}^*} |\hat{\tau}_1(x) - \tau_{10}(x)| \xrightarrow{p} 0,$$

$$\text{(A.18)} \qquad T^{1/4} \sup_{x \in \mathscr{X}^*} |\hat{\tau}_2(x,\theta_0) - \tau_{20}(x,\theta_0)| \xrightarrow{p} 0,$$

$$\text{(A.19)} \qquad T^{1/4} \sup_{x \in \mathscr{X}^*} \left\| \frac{\partial}{\partial\theta}\hat{\tau}_2(x,\theta_0) - \frac{\partial}{\partial\theta}\tau_{20}(x,\theta_0) \right\| \xrightarrow{p} 0,$$

$$\text{(A.20)} \qquad \sup_{x \in \mathscr{X}_\varepsilon^*} |D^\mu \hat{\tau}_1(x) - D^\mu \tau_{10}(x)| \xrightarrow{p} 0 \quad \forall \mu \quad \text{with} \quad |\mu| < q_1,$$

$$\text{(A.21)} \qquad \sup_{\theta \in \Theta^*} \sup_{x \in \mathscr{X}_\varepsilon^*} |D^{\mu_x} D^{\mu_\theta} \hat{\tau}_2(x,\theta) - D^{\mu_x} D^{\mu_\theta} \tau_{20}(x,\theta)| \xrightarrow{p} 0 \quad \forall \mu_x, \mu_\theta$$

$$\text{with} \quad |\mu_x| + |\mu_\theta| \leq q_2, \quad \text{and}$$

$$\text{(A.22)} \qquad \sup_{x \in \mathscr{X}^*} |D^\mu \hat{\tau}_3(x) - D^\mu \tau_{30}(x)| \xrightarrow{p} 0 \quad \forall \mu \quad \text{with} \quad |\mu| < q_3.$$

Equations (A.17)–(A.19) are established using Theorem 1(b) of Andrews (1994b) by verifying its Assumptions NP1–NP5. We note that $(\partial/\partial\theta)\hat{\tau}_2(x,\theta_0)$ is the kernel estimate of $(\partial/\partial\theta)h(Z_t,\theta_0)$ on X_t, so for (A.19) we verify NP1–NP5 with $Y_t = (\partial/\partial\theta)h(Z_t,\theta_0)$. Under Assumption WLS/PPR, $(\partial/\partial\theta)\tau_{20}(x,\theta_0) = (\partial/\partial\theta)E(h(Z_t,\theta_0)|X_t = x) = E((\partial/\partial\theta)h(Z_t,\theta_0)|X_t = x)$, so the result of Theorem 1(b) is the desired result for (A.19). (A.20) is established using Comment 4 following Theorem 1 of Andrews (1994b) by verifying its Assumptions NP1, NP2′, NP3′, NP4, and NP5′. (A.21) is established using Corollary 2 of Andrews (1994b) by verifying its Assumptions NP1*-D and NP2*-NP5*.

Equation (A.22) is established by writing $\sup_{x \in \mathscr{X}^*}|\hat{\tau}_3(x) - \tau_{30}(x)| \leq \hat{A}_{1T} + \hat{A}_{2T}$ provided $\hat{\tau}_1 \in \mathscr{T}_1$ and $\hat{\tau}_2 \in \mathscr{T}_2$ wp $\to 1$, where $\hat{A}_{1T} = \sup_{\theta \in \Theta_0, \tau_1 \in \mathscr{T}_1, \tau_2 \in \mathscr{T}_2} \sup_{x \in \mathscr{X}^*}|D^\mu \hat{\tau}_3(x,\theta,\tau_1,\tau_2) - D^\mu \tau_{30}(x,\theta,\tau_1,\tau_2)|$ and $\hat{A}_{2T} = \sup_{x \in \mathscr{X}^*}|\tau_{30}(x,\theta^*,\hat{\tau}_1,\hat{\tau}_2) - \tau_{30}(x,\theta_0,\tau_{10},\tau_{20})|$. $\hat{\tau}_3(x,\theta,\tau_1,\tau_2)$ denotes the kernel estimate $\hat{\tau}_3(x)$ of $\hat{U}_t^2$ on X_t when the residuals $\hat{U}_t$ are based on "estimates" (θ,τ_1,τ_2) of $(\theta_0,\tau_{10},\tau_{20})$. $\tau_{30}(x,\theta,\tau_1,\tau_2)$ is the corresponding expected value of these residuals given $X_t = x$. $\mathscr{T}_1$ and $\mathscr{T}_2$ are neighborhoods of τ_{10} and τ_{20} that are sufficiently restricted in terms of smoothness on $\mathscr{X}_\varepsilon^*$ of the functions they contain that Corollary 2 of Andrews (1994b) can be used to prove that $\hat{A}_{1T} \xrightarrow{p} 0$ by verifying its Assumptions NP1*-D and NP2*-NP5*. Given that $K_3(\cdot)$ is assumed to be zero outside a bounded set, uniform convergence of $D^\mu \hat{\tau}_3(x,\theta,\tau_1,\tau_2)$ to $D^\mu \tau_{30}(x,\theta,\tau_1,\tau_2)$ over $x \in \mathscr{X}^*$ only requires smoothness of $\tau_{30}(x,\theta,\tau_1,\tau_2)f_t(x)$ on an ε-neighborhood $\mathscr{X}_\varepsilon^*$ of $\mathscr{X}^*$ and not on all of R^k; see Comment 4 to Theorem 3 of Andrews (1994b). In consequence, $\mathscr{T}_1$ and $\mathscr{T}_2$ are defined to include functions that are suitably smooth on $\mathscr{X}_\varepsilon^*$ and, in order for $\hat{\tau}_1$ and $\hat{\tau}_2$ to lie in $\mathscr{T}_1$ and $\mathscr{T}_2$ wp $\to 1$, (A.20) and (A.21) must hold for $x \in \mathscr{X}_\varepsilon^*$ and $\forall \mu$ with $|\mu| \leq q$ for some integer $q \geq (k+l)/2$. The result $\hat{A}_{2T} \xrightarrow{p} 0$ is obtained using (A.20) and (A.21), the conditions on $\tau_{40}(x,\theta)$, and the continuity of $D^\mu \tau_{20}(x,\theta)$ at θ_0 uniformly over $x \in \mathscr{X}^*$.

Theorem 2 is proved by verifying Assumptions C and N with $d(m,\gamma)$ and $m_t(\theta,\tau)$ as in (3.6) and (3.7) respectively. Assumption C(a) holds by applying the uniform WLLN given in Andrews (1992, Thm. 3(a)) and verifying its Assumption BD, P-WLLN, and W-LIP with the metric

$$\rho((\theta_a,\tau_a),(\theta_b,\tau_b)) = \|\theta_a - \theta_b\| + \sup_{x \in \mathscr{X}^*} |\tau_{a1}(x) - \tau_{b1}(x)| + \sup_{\theta \in \Theta} \sup_{x \in \mathscr{X}^*} |\tau_{a2}(x,\theta) - \tau_{b2}(x,\theta)| + \sup_{x \in \mathscr{X}^*} |\tau_{a3}(x) - \tau_{b3}(x)|.$$

P-WLLN is verified using Andrews (1988, Thm. 1, Remark 4 of Sec. 3). Assumption C(b) holds using (A.20)–(A.22) and various parts of WLS/PPR. Assumption C(c) holds automatically because $d(m,\gamma)$ is the identity function. Assumption C(d) holds by WLS/PPR(e). Theorem A-1 now implies that $\hat{\theta} \xrightarrow{p} \theta_0$. An analogous argument using the condition on $m_a(\theta, \tau_0^*)$ given in WLS/PPR(e) yields $\theta^* \xrightarrow{p} \theta_0$, where θ^* is the nonweighted LS estimator that is used in the definition of $\hat{\tau}_3(x)$.

Assumption N(a) holds by Theorem A-1 and WLS/PPR(a). Assumption N(b) holds by (A.20)–(A.22). Assumption N(c) holds by (A.20)–(A.22), because by standard inequalities we can write

$$\text{(A.23)} \quad \left\| \sqrt{T} \frac{\partial}{\partial \theta} d(\bar{m}_T^*(\theta_0, \hat{\tau}), \hat{\gamma}) \right\| \leqslant \sqrt{T} \Big(\sup_{x \in \mathscr{X}^*} |\hat{\tau}_1(x) - \tau_{10}(x)| + \sup_{x \in \mathscr{X}^*} |\hat{\tau}_2(x, \theta_0) - \tau_{20}(x, \theta_0)| \Big) \times \sup_{x \in \mathscr{X}^*} \left\| \frac{\partial}{\partial \theta} \hat{\tau}_2(x, \theta_0) - \frac{\partial}{\partial \theta} \tau_{20}(x, \theta_0) \right\| \Big/ \inf_{\tau \in \mathscr{T},\, x \in \mathscr{X}^*} \tau_3(x).$$

Assumption N(d) holds using WLS/PPR(b) and (c) by a CLT of Herrndorf (1984, Cor. 1). Assumption N(e) holds by the Proposition of Section 2 using the definition of $\mathscr{T}$ and WLS/PPR(a) and (b). Assumption N(f) holds trivially since $d(m, \gamma) = m'm/2$. The continuity of $m(\theta, \tau)$ and $M(\theta, \tau)$ specified in Assumption N(g) holds with the metrics given in (4.10) and (4.11) using (A.20)–(A.22) to show that $(\hat{\theta}, \hat{\tau}) \xrightarrow{p} (\theta_0, \tau_0)$. The uniform WLLNs specified in Assumption N(g) hold by applying the generic uniform WLLN of Andrews (1992, Thm. 3(a)) in the same manner as for Assumption C(a). *Q.E.D.*

REFERENCES

ANDREWS, D. W. K. (1987): "Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers," *Econometrica*, 55, 1465–1472.

——— (1988): "Laws of Large Numbers for Dependent Non-identically Distributed Random Variables," *Econometric Theory*, 4, 458–467.

——— (1989a): "Asymptotics for Semiparametric Econometric Models: I. Estimation and Testing," Cowles Foundation Discussion Paper No. 908R, Yale University.

——— (1989b): "Stochastic Equicontinuity for Semiparametric Models," unpublished manuscript, Cowles Foundation, Yale University.

——— (1991a): "An Empirical Process Central Limit Theorem for Dependent Non-identically Distributed Random Variables," *Journal of Multivariate Analysis*, 38, 187–203.

——— (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

——— (1991c): "Asymptotics for Kernel-based Non-orthogonal Semiparametric Estimators," unpublished manuscript, Cowles Foundation, Yale University.

——— (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241–257.

——— (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.

——— (1994a): "Empirical Process Methods in Econometrics," Chapter 2 in *Handbook of Econometrics*, Vol. 4. New York: North Holland, forthcoming.

——— (1994b): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 10, forthcoming.

ANDREWS, D. W. K., AND J. C. MONAHAN (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966.

BICKEL, P. J. (1982): "On Adaptive Estimation," *Annals of Statistics*, 10, 647–671.

BIERENS, H. J. (1987): "Kernel Estimators of Regression Functions," in *Advances in Econometrics: Fifth World Congress*, Vol. I, ed. by T. F. Bewley. New York: Cambridge University Press.

BILLINGSLEY, P. (1968): *Convergence of Probability Measures*. New York: Wiley.

CHAMBERLAIN, G. (1992): "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567–596.

COX, D. R. (1975): "Partial Likelihood," *Biometrika*, 62, 269–276.

GALLANT, A. R. (1987): *Nonlinear Statistical Models*. New York: Wiley.

HALL, P., AND C. C. HEYDE (1980): *Martingale Limit Theory and its Applications*. New York: Academic Press.

HAN, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316.

HERRNDORF, N. (1984): "An Invariance Principle for Weakly Dependent Sequences of Random Variables," *Annals of Probability*, 12, 141–153.

HOROWITZ, J. L. (1988): "Semiparametric *M*-estimation of Censored Linear Regression Models," *Advances in Econometrics*, 7, 45–83.

ICHIMURA, H., AND L.-F. LEE (1990): "Semiparametric Estimation of Multiple Index Models," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J. L. Powell, and G. Tauchen. New York: Cambridge University Press.

JENNRICH, R. (1969): "Asymptotic Properties of Nonlinear Least Squares Estimators," *Annals of Mathematical Statistics*, 40, 633–643.

KLEIN, R. W., AND R. H. SPADY (1993): "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica*, 61, 387–421.

KREISS, J.-P. (1987): "On Adaptive Estimation in Stationary ARMA Processes," *Annals of Statistics*, 15, 112–133.

LEE, L.-F. (1989): "Semiparametric Maximum Profile Likelihood Estimation of Polytomous and Sequential Choice Models," Discussion Paper No. 253, Department of Economics, University of Minnesota.

MANSKI, C. F. (1984): "Adaptive Estimation of Non-linear Regression Models," *Econometric Reviews*, 3, 145–194.

MCLEISH, D. L. (1975a): "A Maximal Inequality and Dependent Strong Laws," *Annals of Probability*, 3, 826–836.

——— (1975b): "Invariance Principles for Dependent Variables," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebeite*, 32, 165–178.

——— (1977): "On the Invariance Principle for Nonstationary Mixingales," *Annals of Probability*, 5, 616–621.

NEWEY, W. K. (1988): "Two Step Estimation of Sample Selection Models," unpublished manuscript, Department of Economics, Princeton University.

——— (1989): "The Asymptotic Variance of Semiparametric Estimators," unpublished manuscript, Department of Economics, Princeton University.

——— (1990a): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837.

——— (1990b): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161–1167.

NEWEY, W. K., AND K. D. WEST (1987): "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

PAGAN, A., AND A. ULLAH (1988): "The Econometric Analysis of Models with Risk Terms," *Journal of Applied Econometrics*, 3, 87–105.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. New York: Springer Verlag.

——— (1990): *Empirical Processes: Theory and Applications*. CBMS Conference Series in Probability and Statistics, Vol. 2. Hayward, CA: Institute of Mathematical Statistics.

PÖTSCHER, B. M., AND I. R. PRUCHA (1989): "A Uniform Law of Large Numbers for Dependent and Heterogeneous Data Processes," *Econometrica*, 57, 675–683.

POWELL, J. L. (1987): "Semiparametric Estimation of Bivariate Latent Variable Models," SSRI Working Paper No. 8704, University of Wisconsin, Madison.

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

PROHOROV, YU. V. (1956): "Convergence of Random Processes and Limit Theorems in Probability Theory," *Theory of Probability and Its Applications*, 1, 157–214.

ROBERTS, R. B., G. S. MADDALA, AND G. ENHOLM (1978): "Determinants of the Requested Rate of Return and the Rate of Return Granted in a Formal Regulatory Process," *Bell Journal of Economics*, 9, 611–621.

ROBINSON, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

——— (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

SEVERINI, T. A., AND W. H. WONG (1987): "Profile Likelihood and Semiparametric Models," unpublished manuscript, Department of Statistics, University of Chicago.

Stein, E. M. (1970): *Singular Integrals and Differentiability Properties of Functions.* Princeton, NJ: Princeton University Press.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10, 1040–1053.

Whang, Y.-J., and D. W. K. Andrews (1993): "Tests of Model Specifications for Parametric and Semiparametric Models," *Journal of Econometrics*, 57, 277–318.

White, H. (1984): *Asymptotic Theory for Econometricians.* Orlando, FL: Academic Press.

Wooldridge, J. M., and H. White (1988a): "Central Limit Theorems for Dependent, Heterogeneous Processes with Trending Moments," Department of Economics, M.I.T.

——— (1988b): "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes," *Econometric Theory*, 4, 210–230.