COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1066

A LIMIT THEOREM FOR A SMOOTH CLASS OF SEMIPARAMETRIC ESTIMATORS

Ariel Pakes and Steven Olley

January 1994

First Draft, September 1991.

This Draft, August 1993.

A Limit Theorem for A Smooth Class of Semiparametric Estimators[1]

Ariel Pakes and Steven Olley,

Yale University and the NBER, and New York University and the NBER.

JEL Classification; C14(semiparametric and nonparametric methods), C24(truncated and censored models), L11(production and market structure).

Keywords. Semiparametric m–estimators, selection and simultaneity biases in production functions.

---

[1]We have benefitted from the comments of D. Andrews, M. Buchinsky, Z. Griliches, O.Linton, and Y. Ritov, and are grateful for partial financial support from the NSF through grants SES–8821733 and SES–9122672. All errors remain, of course, our own responsibility. This paper is to appear in a volume in honor of Zvi Griliches.

ABSTRACT

We consider an econometric model based on a set of moment conditions which are indexed by both a finite dimensional parameter vector of interest, θ, and an infinite dimensional parameter, h, which in turn depends upon both θ and another infinite dimensional parameter, τ. The model assumes that the moment conditions equal zero at the true value of all unknown parameters. Estimators of θ are obtained by forming nonparametric estimates of h and τ, substituting them into the sample analog of the moment conditions, and choosing that value of θ that makes the sample moments as "close as possible" to zero. Using independence and smoothness assumptions the paper provides consistency, $\sqrt{n}$ consistency, and asymptotic normality proofs for the resultant estimator. As an example, we consider Olley and Pakes'(1991) use of semiparametric techniques to control for both simultanaeity and selection biases in estimating production functions. This example illustrates how semiparametric techniques can be used to overcome both computational problems, and the need for strong functional form restrictions, in obtaining estimates from structural models. We also provide two additional sets of empirical results for this example. First we compare the estimators of θ obtained using different estimtors for the nonparametric components of the problem, and then we compare alternative estimators for the estimated standard errors of those estimators.

A Limit Theorem for A Smooth Class of Semiparametric Estimators

We consider an econometric model that specifies a set of conditions on a vector of population moments

$$G(\theta) \equiv \int m[z,h_0\{z,\tau_0(z),\theta\},\theta]\, P(dz),$$

and assumes they equal zero at $\theta = \theta_0$. Here both h_0 and τ_0 are unknown functions. Estimators of θ are obtained by drawing a random sample of size n from the distribution $P(\cdot)$, forming nonparametric estimates of h_0 and τ_0, say h_n and τ_n, and finding that value of θ that makes the sample moment

$$G_n(\theta) \equiv n^{-1}\Sigma_i\, m[z_i,h_n\{z_i,\tau_n(z_i),\theta\},\theta]$$

as close as possible to zero. The generality obtained by allowing the unknown functions to be indexed by both the parameters of interest, and by other unknown functions, is essential for the examples we have in mind (see below).

Section I provides consistency, root–n consistency, and asymptotic normality results for a "smooth" class of such estimators. Section II discusses prior results which justify the additional assumptions (beyond smoothness) needed for our limit theorems, and section III concludes with an empirical example.

The example, taken from Olley and Pakes(1991), uses semiparametric estimation techniques to control for both simultanaeity and selection biases in estimating production functions. It starts from Griliches'(1967) notion that one cannot obtain interpretable estimates of production function parameters without a complete model of firm decision

making, and then illustrates how semiparametric techniques can be used to overcome both the computational problems and the need for strong functional form assumptions that arise in obtaining parameter estimates that are consistent with such a model. In this paper we provide the limit distribution of Olley and Pakes'(1991) estimator, and then compare empirical results obtained from both, i)alternative estimators for the nonparametric components of their problem, and ii)alternative estimators for the standard errors of those estimators.

We assume at the outset that all functions are sufficiently smooth in all of their arguments, and that the data are an i.i.d. sample from some population. Though, as pointed out below, these assumptions could be relaxed in an increasing number of ways, they allow us to provide proofs for the limit theorems, and an exposition of the conditions that will suffice for them, that are reasonably transparent.

In addition, use of the moment conditions framework allows us to use familiar arguments from prior work (see, for e.g. Hansen,1982, Hardle and Stoker,1989, Manski,1991, and Pakes and Pollard,1989), and this simplifies the proofs of the propositions. In particular the proofs presented here are structured similarly to the proofs given in section III of Pakes and Pollard (1989) for the parametric case, so that the reader who is familiar with that material should have little difficulty with the material that follows. This structure of proof has the additional advantage that it can also be used for cases that do not satisfy our smoothness restrictions.

The problem dealt with in Section I, and the results obtained there, are, however, more similar to Newey (1991; see also Andrews 1994). Indeed, if we were to extend Newey's definition of a semiparametric M–estimator (his 3.1) to allow the unknown functions to depend on other unknown functions, and extend his definition of the "pathwise" derivative analogously, then Newey's formula for the limit distribution of the parameters of interest would be identical to ours (and Newey does not require the smoothness conditions used here). The incremental contribution of section 1 of this paper

is in providing a set of assumptions that, for the smooth case, allows us to justify our limit theorems when there are unknown functions that are indexed both by the vector of parameters of interest and by other unknown functions, and in the simplicity of the method of proof. We conclude the first section by noting first that given consistency (and our smoothness and independence assumptions), the additional assumptions needed for the limit distribution of our estimator are automatically satisfied if "orthogonality" conditions similar to those used in Andrews(1994) are satisfied (a similar result is presented in Newey,1991). As noted by Andrews(1994), in this special case we need not adjust the variance–covariance of our estimated parameters for the fact that we use estimated (rather than actual) values of h_0 and τ_0 in the definition of our moment conditions. We then extend this discussion and provide conditions under which we need not adjust the variance covariance for the fact that we use an estimate of τ_0 in the definition of our moment conditions, even though an adjustment for the fact that we use an estimate of h_0 is still required.

Section II, which discusses conditions on the primitives of our problem (in particular on the nonparametric estimators) which insure that the assumptions needed for the limit theorems of section I are satisfied, is largely a short summary of relevant results in prior work by Andrews (1991,1993,1994,forthcoming), Newey(1991 and forthcoming), Pollard (1991), and Powell, Stock, and Stoker (1989). It makes particularly intensive use of the results in Andrews(1993). Section II deals explicitly only with the case where h_0 and τ_0 are conditional expectations, as this is the case we need for our example. However, as noted in section II, the literature provides conditions which justify the assumptions used in section I, and the precise form of the resulting limit distribution, for a variety of other cases.

The empirical example in Section III uses theory to generate investment and exit rules that allow one to correct estimates of production–function parameters for both the simultaneity problem induced by endogenous input demands, and the selection problem induced by exit behavior. The estimation algorithm leaves the exit rule and the

investment equation, as well as the process generating differences in productivity over time, as nuisance functions to be accounted for by nonparametric techniques. As is frequently the case in estimation problems derived directly from a behavioral model, the relevant moment conditions involve unknown functions which are indexed by both the parameters of interest and by other unknown functions.

We present several estimators of the production function parameters that differ in the way the nonparametric components of the problem are estimated. Also, for each such estimator, we compare standard errors estimated in three ways; once using a bootstrap, once using an analytic form which produces a consistent estimator of the variance–covariance matrix generated by our limit theorems, and once using an analytic form which ignores the fact that h_0 and τ_0 are not known at the outset. The alternative estimators include both kernel and series estimators, and among the kernel estimators we present estimates based on both bias–reducing kernels with bandwidths obtained from a cross–validation procedure, and estimates obtained from a standard Normal kernel.

Of these alternatives the only one that, to our knowledge, is currently known to satisfy all the assumptions of the limit theorems is the estimator based on the bias–reducing kernels. The results, however, do not differ much among estimation techniques. The correction to the standard error formula that is required to adjust for the presence of the estimates of the nonparametric functions (in the case where orthogonality conditions do not hold) does seem to increase the estimated variances, but the parameters of major interest are still estimated with a fair amount of precision. Interestingly, the bootstraps produced estimates of standard errors which tended to be larger than the estimates obtained from the analytic formula. Also worth pointing out is that the bias–reducing kernels produced a distribution of bootstrapped coefficients with a slightly larger variance, and a significantly larger kurtosis, then did the non bias–reducing kernels.

We note that our empirical example is quite complex, involving nonparametric estimates of three unknown functions, and structural parameters that are buried inside

them (though in an intuitive way). Thus the fact that the estimated standard errors were within traditionally acceptable bounds for production function estimators (though the sample was not terribly large by modern microeconometric standards) is indicative of the potential usefulness of semiparametric techniques in alleviating both restrictive assumptions, and computational bottlenecks, in empirical work on micro data sets. In particular, estimation of any reasonable specification of a parametric version for our problem (and by this we mean any specification which allowed for serial correlation in the productivities of a plant over time), would have led to a computational problem which would have probably been impractical even using the current generation of supercomputers (see the discussion in sections 3 and 4 in Pakes, forthcoming). In contrast, some of the nonparametric specifications ran in under an hour on our 486 personal computer. The computational advantages of semipararametric techniques also come out clearly from the recent work of Hotz and Miller (1991).

Notation.

The symbol $\|\cdot\|$ denotes not only the usual Euclidean norm but also a matrix norm: $\|(b_{ij})\| = (\Sigma_{ij}\, b_{ij}^2)^{1/2}$. It has the useful property that $\|Bx\| \leq \|B\|\,\|x\|$, for each vector x and each conformable matrix B. The symbol $\rightarrow_d$ will denote convergence in distribution.

I.Limit Theorems.

The population moment conditions are

$$(1a) \quad G(\theta) \equiv E\, m[z, h_0\{v_1, \tau_0(v_2), \theta\}, \theta] ,$$

with $G(\theta)$ an $\mathbb{R}^1$ valued function which satisfies $G(\theta)=0$ at $\theta=\theta_0$.

Here the expectation operator is with respect to P, a distribution for z with support $\mathcal{Z} \subset \mathbb{R}^d$, and $v'=(v_1', v_2')$ is a subvector of z whose support contains $\mathcal{V} \subset \mathbb{R}^{d_v}$ ($d_v \leq d$). v_1

and v_2 have supports $\mathcal{V}_1 \subset \mathbb{R}^{d1}$ and $\mathcal{V}_2 \subset \mathbb{R}^{d2}$, respectively. $\theta \in \Theta$, a bounded subset of $\mathbb{R}^k$. $\tau_0(\cdot) \in \mathcal{T}$ and $h_0(\cdot) \in \mathcal{H}$ where $\mathcal{H}$ and $\mathcal{T}$ are pseudo metric spaces of functions from $\mathcal{V}_1 \times T \times \Theta \to H \subset \mathbb{R}^h$, and from $\mathcal{V}_2 \to T \subset \mathbb{R}^t$, respectively. Thus $m(\cdot) : \mathcal{Z} \times H \times \Theta \to \mathbb{R}^k$; $h_0 : \mathcal{V}_1 \times T \times \Theta \to H$; and $\tau_0(\cdot) : \mathcal{V}_2 \to T$.

We note that it will be assumed that $m(\cdot)$ can be written as

$$(1b) \qquad m[z,h_0\{v_1,\tau_0(v_2),\theta\},\theta] = m^*[z,h_0\{v_1,\tau_0(v_2),\theta\},\theta]I\{v \in \mathcal{V}\},$$

where $I\{\cdot\}$ is an indicator function which takes the value of one if the condition inside it is satisfied and zero elsewhere. The indicator function, $I\{\cdot\}$, allows us to trim out, or delete, those observations with v values that are "outliers" in the sense that they lie in subsets of $\mathbb{R}^{dv}$ where there are likely to be few observations; too few to allow us to estimate the associated values of $h(\cdot)$ or $\tau(\cdot)$ precisely. Though we needed to introduce the indicator function here for completeness, it has no direct role to play in the discussion of the limit theorems of this section. We come back to it in section 2 where we provide restrictions on $\mathcal{V}$ which insure that the rate conditions used as assumptions in the proofs of the limit theorems of this section are indeed satsified. More detailed notation will be introduced when needed.

Our problem is to estimate θ, and we would use a method of moments estimator if h_0 and τ_0 were known. Because these two functions are not known, we plug preliminary estimates of them, say $\tau_n(\cdot)$, and $h_n(v,\tau_n,\theta) \equiv h_n\{v_1,\tau_n(v_2),\theta\}$, into $G(\cdot)$ and consider minimizing a distance in

$$(2) \qquad G_n(\theta) \equiv n^{-1} \Sigma_i \, m[z_i,h_n(v_i,\tau_n,\theta),\theta] \, .$$

Starting with Euclidean distance then, θ_n, our estimate of θ, will be assumed to satisfy

$$(3) \quad \| G_n(\theta_n) \| \equiv \inf_{\theta \in \Theta} \| G_n(\theta) \| + o_p(1/\sqrt{n}).$$

Note that in general $h_n(\cdot)$ has to be reestimated for each different value of θ.

Our goal is to provide a limit distribution for a θ_n which satisfies (3). To do so we show that any such θ_n is also a solution to a different problem, and then show that the limit properties of the solutions to this simpler problem can be analyzed quite easily. We justify the use of this method of proof by assuming that the functions $m(\cdot)$ and $h(\cdot)$ are sufficiently smooth (though, again, less restrictive justifications could be used at a cost of increasing the complexity of the proofs). Briefly we assume that: $m(\cdot)$ is twice continuously differentiable in $h=h(\cdot)$, and once continuously differentiable in θ; and that $h(\cdot)$ is twice continuously differentiable in $t=\tau(\cdot)$. Further these derivatives are assumed to be continuously differentiable in θ in a region of θ_0 and bounded by functions (envelopes) which are square integrable with respect to P (see below). Also for expositional convenience we consider the case where $h(\cdot)$ and $\tau(\cdot)$ are both functions into a subset of $\mathbb{R}^1$. The example will extend the discussion to higher dimensional spaces.

We introduce these assumptions now. To simplify the exposition we use the following conventions. Capital letters are used to designate functions which are constructed as derivatives of their lower case counterparts, with sub or super scripted capitals used to differentiate different derivatives where there is a chance of confusion, and iterated capitals used for second derivatives. Also we subscript functions by n when we are referring to a nonparametric estimate of the function from a sample of size n, and we omit a function from an index set when we evaluate that function at its true value.

Assumption R. (regularity conditions)

For each $(\theta,h,\tau) \in \Theta \times \mathcal{H} \times \Upsilon$ let

$$m(z,h,\tau,\theta) \equiv m[z,h=h(v,\tau,\theta),\theta],$$
$$m(z,\theta)\equiv m[z,h_0(v,\tau_0,\theta),\theta],$$
$$M(z,h,\tau,\theta)\equiv \partial m(z,h,\theta)/\partial h \,|\, h=h(v,\tau,\theta),$$
$$M(z,\theta)\equiv M(z,h_0,\tau_0,\theta),$$
$$MM(z,h,\tau,\theta)\equiv \partial^2 m(z,h,\theta)/\partial^2 h \,|\, h=h(v,\tau,\theta),$$

and assume that $M(z,h,\tau,\theta)$ and $MM(z,h,\tau,\theta)$ exist (a.e.P). Also assume that there are envelopes $\overline{M}(\cdot)$, and $\overline{MM}(\cdot)$, with the property that

$$|M(z,\theta)| \leq \overline{M}(z), \quad |MM(z,h,\tau,\theta)| \leq \overline{MM}(z),$$

and

$$\int \overline{M}(z)^2 P(dz) \leq \kappa, \quad \int \overline{MM}(z)^2 P(dz) \leq \kappa, \quad \text{for some } \kappa < \infty.$$

The functions $h(v,\tau,\theta)$, and $h(v,\theta)$ are defined analogously. Further their derivatives with respect to τ, $H_t(v,\tau,\theta)$, $H_t(v,\theta)$, and $HH_{tt}(v,\tau,\theta)$, are assumed to exist (a.e. P) and be bounded by the square integrable envelopes $\overline{H}_t(\cdot)$, and $\overline{HH}_{tt}(\cdot)$.

$m(z,\theta)$, $M(z,\theta)$, $h(v,\theta)$, and $H_t(v,\theta)$ are all continuously differentiable in θ (a.e. P). Moreover, for all θ in some neighborhood of θ_0, $M^\theta(v,\theta)$ and $H_t{}^\theta(v,\theta)$ exist and are bounded by the square integrable functions $\overline{M}^\theta(v)$ and $\overline{H}_t{}^\theta(v)$, while $\|H^\theta(v,\theta) - H^\theta(v,\theta_0)\| \leq \|\overline{H}^\theta(v)\|\|\theta-\theta_0\|$, with $\overline{H}^\theta(v)$ integrable.

Finally as n grows large the nonparametric estimate, $h_n[\tau_n]$, will be contained in $\mathscr{H}[I]$ with probability tending to one, and θ_0 is in the interior of Θ. □.

Section II will provide conditions which insure that as n grows large $h_n[\tau_n]$ will be contained in $\mathscr{H}[I]$ with probability tending to one. Other than that, assumption R will be maintained throughout the rest of the paper without further comment. Note that it, together with a Taylor's expansion, imply that we can write $m(\cdot)$ and $h(\cdot)$ at each

$(z_i,h_n,\tau_n,\theta) \in \mathbb{Z} \times \mathcal{H} \times \mathbb{T} \times \Theta$ as

$$(4a) \quad m[z_i,h_n(v_i,\tau_n,\theta),\theta] =$$

$$m(z_i,\theta) + M(z_i,\theta)[h_n(v_i,\tau_n,\theta) - h_0(v_i,\theta)] + (1/2)MM[z_i,\bar{h}_n(v_i,\tau_n,h_n,\theta),\theta][h_n(v_i,\tau_n,\theta) - h_0(v_i,\theta)]^2,$$

where $\bar{h}_n(v_i,\tau_n,h_n,\theta) \in [h_n(v_i,\tau_n,\theta),h_0(v_i,\theta)]$, and

$$(4b) \quad h_n(v_i,\tau_n,\theta) =$$

$$h_n(v_i,\theta) + H_{tn}(v_i,\theta)[\tau_n(v_i) - \tau_0(v_i)] + (1/2)HH_{ttn}[v_i,\bar{t}(v_i,\tau_n),\theta][\tau_n(v_i) - \tau_0(v_i)]^2 =$$

$$h_n(v_i,\theta) + H_{t0}(v_i,\theta)[\tau_n(v_i) - \tau_0(v_i)] + [H_{tn}(v_i,\theta) - H_{t0}(v_i,\theta)][\tau_n(v_i) - \tau_0(v_i)]$$

$$+ (1/2)HH_{ttn}[v_i,\bar{t}(v_i,\tau_n),\theta][\tau_n(v_i) - \tau_0(v_i)]^2,$$

where $\bar{t}(v_i,\tau_n) \in [\tau_n(v_i),\tau_0(v_i)]$.

We now use the first two terms in (4a) and (4b) to define a different minimization problem which is easier to work with, and then provide conditions under which the θ_n which satisfies (3) also solves the new minimization problem. Let

$$(5) \quad m(z,h_n,\tau_n,\theta) =$$

$$\{m(z,\theta) + M(z,\theta)[h_n(v,\theta) - h_0(v,\theta)] + M(z,\theta)H_{t0}(v,\theta)[\tau_n(v) - \tau_0(v)]\}$$

and

$$(6) \quad \mathcal{G}_n(\theta) = 1/n \sum m(z_i,h_n,\tau_n,\theta).$$

Below we provide conditions which insure that

$$\sup_\theta \| \mathcal{G}_n(\theta) - G_n(\theta) \| = o_p(1/\sqrt{n}).$$

A standard argument will then insure that if θ_n is an estimator which satisfies

$$(7) \qquad \mathcal{G}_n(\theta_n) = \inf_\theta \| \mathcal{G}_n(\theta)\| + o_p(1/\sqrt{n}) ,$$

it will also satisfy our original problem (3), and, conversely, if θ^*_n is an estimator which satisfies (7) it must also satisfy (3). Thus we can analyze our estimator by analyzing the set of estimators which satisfy either (7) or (3), and it is easier to work with those that satisfy (7).

We now introduce a set of three assumptions on the rates of convergence of the estimators of h_0 and τ_0 (A1a to A1c), that together insure that $\sup_\theta \| \mathcal{G}_n(\theta) - G_n(\theta)\| = o_p(1/\sqrt{n})$. Section II provides conditions which insure that the rates in A1a to A1c are attained.

Assumption 1. (rates of convergence)

$$(A1a) \qquad n^{a_1} \sup_{(v \in \mathcal{V},\, \theta \in \Theta)} \| h_n(v,\theta) - h_0(v,\theta)\| = O_p(1),$$

$$(A1b) \qquad n^{a_2} \sup_{(v \in \mathcal{V})} \| \tau_n(v) - \tau_0(v) \| = O_p(1),$$

$$(A1c) \qquad n^{a_3} \sup_{(v \in \mathcal{V},\, \theta \in \Theta)} \| H_{tn}(v,\theta) - H_{t0}(v,\theta) \| = O_p(1),$$

with

$$a_i > 1/4, \text{ and } a_i + a_3 > 1/2, \quad \text{for } i=1,2,$$

and

$$(A1d) \qquad \sup_{(v \in \mathcal{V})} \| H^\theta{}_n(v,\theta_0) - H^\theta{}_0(v,\theta_0) \| = o_p(1), \qquad \square.$$

(A1d) is introduced here for convenience, as it follows from (A1c) in all the examples we are aware of and helps simplify the asymptotic normality proof below.[2]

---

[2]In the examples $h_0(\cdot)$ is an unknown function of a known (possibly vector valued) function of τ,v, and θ, say $x(\tau,v,\theta)$ [for more details see the discussion in section II]. Then (A1c)

Only (A1a) to (A1c) are needed for Lemma 8.

8.Lemma.

$$\sup_\theta \| \mathcal{G}_n(\theta) - G_n(\theta) \| = o_p(1/\sqrt{n}).$$

Proof.

From (2),(4),(5),(6) and the triangle inequality

$$\begin{aligned}
&\| G_n(\theta) - \mathcal{G}_n(\theta) \| \leq \\
&\| n^{-1} \Sigma_i M(z_i,\theta)[H_{tn}(v_i,\theta) - H_{t0}(v_i,\theta)][\tau_n(v_i) - \tau_0(v_i)] \| + \\
&\| n^{-1} \Sigma_i M(z_i,\theta) H H_{ttn}[v_i, \bar{t}(v_i,\tau_n), \theta][\tau_n(v_i) - \tau_0(v_i)]^2 \| + \\
&\| n^{-1} \Sigma_i MM[z_i, \bar{h}_n(v_i,\tau_n,h_n,\theta), \theta]\{h_n(v_i,\theta) - h_0(v_i,\theta) + H_{t0}(v_i,\theta)[\tau_n(v_i) - \tau_0(v_i)] \\
&+ [H_{tn}(v_i,\theta) - H_{t0}(v_i,\theta)][\tau_n(v_i) - \tau_0(v_i)] + H H_{ttn}[v_i, \bar{t}(v_i,\tau_n), \theta][\tau_n(v_i) - \tau_0(v_i)]^2\}^2 \|,
\end{aligned}$$

with $\bar{t}(\cdot)$, and $\bar{h}(\cdot)$ defined as in (4). We consider only the first term in this expression. The other terms can be handled analogously. Since

$$\begin{aligned}
&\| M(z_i,\theta)[H_{tn}(v_i,\theta) - H_{t0}(v_i,\theta)][\tau_n(v_i) - \tau_0(v_i)] \| \\
&\leq \| \overline{M}(z_i) \| \times \| [H_{tn}(v_i,\theta) - H_{t0}(v_i,\theta)] \| \times \| [\tau_n(v_i) - \tau_0(v_i)] \| ,
\end{aligned}$$

the triangle inequality implies that the supremum with respect to θ of the left hand side of the first expression is

$$\begin{aligned}
&\leq \sup_{(\theta)} \{ (1/n) \Sigma_i \| \overline{M}(z_i) \| \times \| [H_{tn}(v_i,\theta) - H_{t0}(v_i,\theta)] \| \times \| [\tau_n(v_i) - \tau_0(v_i)] \| \} \\
&\leq n^{-(a_2 + a_3)} O_p(1) \times (1/n) \Sigma_i \| \overline{M}(z_i) \| \leq o_p(1/\sqrt{n}),
\end{aligned}$$

---

will imply (A1d) provided $\partial x(\tau_0, v, \theta_0)/\partial \tau \neq 0$ whenever $\partial x(\tau_0, v, \theta_0)/\partial \theta \neq 0$.

where the inequalities are due to A1 (parts b and c), and the law of large numbers for i.i.d. deviates since, by assumption, $\int \overline{M}(z)^2 P(dz) < \infty$. □.

Equation (7) (with θ_n defined as in 3) follows immediately. We now use that equation to prove consistency, $\sqrt{n}$ consistency, and asymptotic normality of θ_n. Assumption 1 could have been replaced by weaker conditions for the consistency proof, but we will use something close to it later on, and by employing assumption 1 here we can provide a consistency proof which relies on steps that are familiar from parametric problems.

9.Theorem. (consistency)

The definition of the estimator in (3), Assumptions R and 1, and the identification condition that for any $\delta > 0$

$$\inf_{\|\theta-\theta_0\|>\delta} \| G(\theta) \| > 0,$$

imply that $\theta_n - \theta_0 = o_p(1)$.

Proof.

From the definitions of $\mathcal{G}_n(\theta)$ and $G(\theta)$ and an argument analogous to that used in the proof of (8) we have

$$\sup_\theta \| \mathcal{G}_n(\theta) - G(\theta) \| \leq \sup_\theta \| n^{-1} \Sigma_i [m(z_i,\theta) - Em(z_i,\theta)] \| + o_p(1).$$

Thus we need only work with the finite dimensional parameter. The fact that the r.h.s. of this expression is $o_p(1)$ then follows from the differentiability of $m(\cdot,\theta)$ in θ and the

boundedness of θ since they imply that $\sup_\theta \|n^{-1}\Sigma[m(z_i,\theta)-Em(z_i,\theta)]\|=o_p(1)$ (a uniform law of large numbers based on these conditions is provided, for example, in lemma 2.13 of Pakes and Pollard,1989). This together with the identification condition and the definition of the estimator in (7), imply the result (see for example, Pakes and Pollard Corollary 3.2). □.

We now move to the proof of $\sqrt{n}$ consistency. Here we begin by proving lemma 10, which, together with assumption 2 (stated immediately thereafter), makes the proof of $\sqrt{n}$ consistency straightforward. However, before moving to the proof of this lemma two comments might prove helpful. First it is the most detailed argument in the paper so a reader who prefers to begin with an overview of where we are going may prefer to jump ahead to the $\sqrt{n}$ consistency proof and return to lemma 10 thereafter. Second, lemma 10 can be shown to be true under less restrictive assumptions than those we are assuming using results on stochastically equicontinuous families of functions. As a result, *after* providing $\sqrt{n}$ consistency and asymptotic normality proofs using only the more familiar concepts introduced thus far, we turn to a brief introduction to the notion of stochastic equicontinuity and provide references to a literature which insures that lemma 10 holds for a wider variety of cases than those considered here.

10.Lemma.

For any sequence $\{\overline{\theta}_n\}$ such that $\overline{\theta}_n - \theta_0 = o_p(1)$,

$$\|\sqrt{n}[\mathcal{G}_n(\overline{\theta}_n)-G(\overline{\theta}_n)] - \sqrt{n}\,\mathcal{G}_n(\theta_0)\| = o_p(1)[1 + \|\sqrt{n}(\overline{\theta}_n-\theta_0)\|]. \quad \square.$$

Proof.

From the definitions of $\mathcal{G}_n$ and G and the triangle inequality

$$(10^*) \quad \|\sqrt{n}[\mathcal{G}_n(\overline{\theta}_n)-G(\overline{\theta}_n)] - \sqrt{n}\,\mathcal{G}_n(\theta_0)\| \leq$$
$$\|(1/\sqrt{n})\Sigma[m(z_i,\overline{\theta}_n)-Em(z_i,\overline{\theta}_n)]-(1/\sqrt{n})\Sigma m(z_i,\theta_0\}\|$$
$$+ \|(1/\sqrt{n})\Sigma M(z_i,\overline{\theta}_n)[h_n(v_i,\overline{\theta}_n)-h_0(v_i,\overline{\theta}_n)] - (1/\sqrt{n})\Sigma M(z_i,\theta_0)[h_n(v_i,\theta_0)-h_0(v_i,\theta_0)]\|$$
$$+\|(1/\sqrt{n})\Sigma M(z_i,\overline{\theta}_n)H_{0t}(v_i,\overline{\theta}_n)[\tau_n(v_i)-\tau_0(v_i)]-(1/\sqrt{n})\Sigma M(z_i,\theta_0)H_{0t}(v_i,\theta_0)[\tau_n(v_i)-\tau_0(v_i)]\|.$$

We prove that for any sequence $\{\overline{\theta}_n\}$ such that $\overline{\theta}_n-\theta_0=o_p(1)$, the second term is $o_p(1)[1+\|\sqrt{n}(\overline{\theta}_n-\theta_0)\|]$ (the proofs that the other two terms are also have a similar structure, but require less detail). Note that this term is

$$\leq \|(1/\sqrt{n})\Sigma M(z_i,\theta_0)\{[h_0(v_i,\theta_0)-h_0(v_i,\overline{\theta}_n)] - [h_n(v_i,\theta_0)-h_n(v_i,\overline{\theta}_n)]\}\|$$
$$+ \|(1/\sqrt{n})\Sigma[M(z_i,\overline{\theta}_n)-M(z_i,\theta_0)][h_n(v_i,\overline{\theta}_n)-h_0(v_i,\overline{\theta}_n)]\|.$$

Given that $\overline{\theta}_n-\theta_0=o_p(1)$, assumption R insures that the last term in this expression is, with probability tending to one

$$\leq \|\sqrt{n}(\overline{\theta}_n-\theta_0)\|\ \sup_{v,\theta}\|h_n(v,\theta)-h_0(v,\theta)\|\ (1/n)\ \Sigma\ \|\overline{M}^{\theta}(z_i)\|$$
$$=o_p(1)\ \|\sqrt{n}(\overline{\theta}_n-\theta_0)\|\ ,$$

where the equality is a result of assumption 1a, the square integrability of $\overline{M}^{\theta}(z_i)$, and the law of large numbers for i.i.d. deviates. In addition

$$(1/\sqrt{n})\ \Sigma\ \{\ M(z,\theta_0)\ \{\ [h_n(v,\overline{\theta}_n)-h_n(v,\theta_0)] - [h_0(v,\overline{\theta}_n)-h_0(v,\theta_0)]\ \}$$
$$\leq (1/n)\ \Sigma\ \|M(z,\theta_0)\|\ \{\ \|H_n^{\theta}(v,\theta_0)-H_0^{\theta}(v,\theta_0)\| +$$
$$\|H_n^{\theta}[v,\theta_n^*(\overline{\theta}_n,v)]-H_n^{\theta}(v,\theta_0)\| + \|H_0^{\theta}[v,\theta_0^*(\overline{\theta}_n,v)]-H_0^{\theta}(v,\theta_0)\}\|\ \}\ \|\sqrt{n}(\overline{\theta}_n-\theta_0)\|$$

where both $\theta_n^*(\overline{\theta}_n,v)$ and $\theta_0^*(\overline{\theta}_n,v)$ are in the interval $[\theta_0,\overline{\theta}_n]$ for each v, and the inequality is due to the differentiablity conditions in Assumption R.

Since independence and square integrability implies that $(1/n)\Sigma\|M(z,\theta_0)\|=O_p(1)$, and assumption 1d insures that $(1/n)\Sigma\|H_n^\theta(v,\theta_0)-H_0^\theta(v,\theta_0)\|=o_p(1)$, Holder's inequality implies that we need only show that $(1/n)\Sigma\|H_n^\theta[v,\theta_n^*(\overline{\theta}_n,v)]-H_n^\theta(v,\theta_0)\|$, and $(1/n)\Sigma\|H_0^\theta[v,\theta_0^*(\overline{\theta}_n,v)]-H_0^\theta(v,\theta_0)\}\|$ are $o_p(1)$ to complete the proof of the lemma. Since the probability that either of these expressions is less than ϵ is, for arbitrary δ, less than or equal to

$$(11)\quad \Pr\{\sup_{(h\in\mathcal{H})}\sup_{(\|\theta-\theta_0\|\le\delta)}(1/n)\Sigma\|H^\theta(v,\theta)-H^\theta(v,\theta_0)\|\ge\epsilon\} + \Pr\{\|\overline{\theta}_n-\theta_0\|\ge\delta\}+\Pr\{h_n\in\mathcal{H}\},$$

it will suffice to show that we can make this expression less than ϵ by choosing n large enough. The fact that $\overline{\theta}_n-\theta_0 = o_p(1)$ together with assumption R (which implies that $\Pr\{h_n\in\mathcal{H}\}\to 1$ as $n\to\infty$) insures that the last two terms can be made arbitrarily small ($\le\epsilon/3$) by choosing n large enough. Now recall that for θ near θ_0, $\|H^\theta(v,\theta)-H^\theta(v,\theta_0)\| \le \|\overline{H}^\theta(v)\|\|\theta-\theta_0\|$, from which it follows that

$$\sup_{(h\in\mathcal{H})}\sup_{(\|\theta-\theta_0\|\le\delta)}(1/n)\Sigma\|H^\theta(v,\theta)-H^\theta(v,\theta_0)\|\le\delta(1/n)\Sigma\|\overline{H}^\theta(v)\|=\delta E[\|\overline{H}^\theta(v)\|]+o_p(1),$$

which, provided δ is chosen small enough $\{<\epsilon/(3E[\|\overline{H}^\theta(v)\|])\}$, insures that for n large enough the first expression in (11) will be less than $\epsilon/3$ (for arbitrary ϵ). □.

Lemma 10 and assumption 2 (that $\mathcal{G}_n(\theta_0)$ is stochastically bounded) provide the basis for the $\sqrt{n}$ consistency proof. Since assumption 2 follows from assumption 3 (the normality assumption used in our asymptotic normality proof), we delay justification of it until assumption 3 is introduced.

Assumption 2.

$$\mathcal{G}_n(\theta_0) = O_p(1/\sqrt{n}) \; . \quad \square.$$

11. Theorem. ($\sqrt{n}$ consistency).

Let $D(\theta_1) \equiv \partial G(\theta)/\partial\theta' \,|\, \theta=\theta_1$, and assume that $D(\theta_0)$ is of rank k. Then the consistency result in (9), A2, and Lemma 10 imply that

$$\theta_n - \theta_0 = O_p(1/\sqrt{n}).$$

Proof.

Use the triangle inequality twice to show that

$$\begin{aligned} \|\sqrt{n}[\mathcal{G}_n(\theta_n) - G(\theta_n)] - \sqrt{n}\,\mathcal{G}_n(\theta_0)\| &\geq \|\sqrt{n}G(\theta_n)\| - \|\sqrt{n}\,\mathcal{G}_n(\theta_n)\| - \|\sqrt{n}\,\mathcal{G}_n(\theta_0)\| \\ &\geq \|\sqrt{n}G(\theta_n)\| - 2\|\sqrt{n}\,\mathcal{G}_n(\theta_0)\| \; , \end{aligned}$$

where the last inequality follows from the fact that θ_n minimizes the objective function. Recall that for any matrix $B=[b_{ij}]$, $\|B\|=(\Sigma_{ij}b_{ij}^2)^{1/2}$. Thus the above inequality, A2, and lemma 10 imply that

$$O_p(1) + o_p(1)[1 + \|\sqrt{n}(\theta_n - \theta_0)\|] \geq \|\sqrt{n}G(\theta_n)\| = \|D(\theta_{0n})[\sqrt{n}(\theta_n - \theta_0)]\|,$$

where $\theta_{0n} \in [\theta_n, \theta_0]$. Continuity of $D(\theta)$, consistency of θ_n, and the fact that $D(\theta_0)$ has rank k imply that

$$\|D(\theta_{0n})[\sqrt{n}(\theta_n - \theta_0)]\| = \|[D(\theta_0) + o_p(1)][\sqrt{n}(\theta_n - \theta_0)]\| \geq c\|\sqrt{n}(\theta_n - \theta_0)\| - o_p(1)\|\sqrt{n}(\theta_n - \theta_0)\|,$$

for some $c > 0$. Rearranging terms we have $O_p(1) \geq \|\sqrt{n}(\theta_n - \theta_0)\|$, as required. $\square$.

The argument for asymptotic normality is now essentially the same as in the parametric case. We provide it for completeness (the argument parallels the proof of the second part of Theorem 3.3 in Pakes and Pollard). Let $D \equiv D(\theta_0)$ $(=\partial G(\theta)/\partial\theta_{|\theta=\theta_0})$, assume it has full rank, and consider the quadratic form $\| L_n(\theta) \|$ where

$$(12a) \qquad L_n(\theta) = D(\theta-\theta_0) + \mathcal{G}_n(\theta_0).$$

The value of θ which minimizes $\| L_n(\theta) \|$, say θ_n^*, can be solved for explicitly as

$$(12b) \qquad \sqrt{n}(\theta_n^* - \theta_0) = -(D'D)^{-1}D'\sqrt{n}\,\mathcal{G}_n(\theta_0).$$

Consequently the limit distribution of $\sqrt{n}(\theta_n^* - \theta_0)$ follows directly from an assumption on the asymptotic normality of $\sqrt{n}\,\mathcal{G}_n(\theta_0)$ [assumption 3]. This will also provide the limit distribution of $\sqrt{n}(\theta_n - \theta_0)$ provided

$$(13) \quad \sqrt{n}(\theta_n^* - \theta_0) = \sqrt{n}(\theta_n - \theta_0) + o_p(1).$$

So what the asymptotic normality proof requires is (13) and assumption 3. We provide a discussion of assumption 3 (of the more primitive conditions which might suffice for it, as well as of the form of the variance covariance matrix) immediately after the asymptotic normality proof.

Assumption 3.

$$\sqrt{n}\,\mathcal{G}_n(\theta_0) \to_d \mathcal{N}(0,V),$$

with $\|V\| < \infty$. □.

Theorem 14. (Asymptotic Normality).

Assume that $\theta_n-\theta_0=O_p(1/\sqrt{n})$, that D has full rank, and assumptions R,1, and 3. Then

$$\sqrt{n}(\theta_n - \theta_0) \to_d \mathcal{N}[0,(D'D)^{-1}D'VD(D'D)^{-1}].$$

Proof.

We begin by showing that

$$\| L_n(\theta_n) - \mathcal{G}_n(\theta_n) \| = o_p(1/\sqrt{n}) = \| L_n(\theta^*_n) - \mathcal{G}_n(\theta^*_n) \| .$$

Use the triangle inequality, lemma 10, and differentiability of $G(\cdot)$ to show that for any $\{\bar{\theta}_n\}$ which satisfies $\bar{\theta}_n-\theta_0=O_p(1\sqrt{n})$

$$\| L_n(\bar{\theta}_n) - \mathcal{G}_n(\bar{\theta}_n) \| \le \| \mathcal{G}_n(\bar{\theta}_n) - G(\bar{\theta}_n) - \mathcal{G}_n(\theta_0) \| + \| L_n(\bar{\theta}_n) - G(\bar{\theta}_n) - \mathcal{G}_n(\theta_0) \| =$$
$$o_p(1/\sqrt{n}) + \| L_n(\bar{\theta}_n) - G(\bar{\theta}_n) - \mathcal{G}_n(\theta_0)\| = o_p(1/\sqrt{n}) + \| D(\bar{\theta}_n-\theta_0) - G(\theta_n) \| = o_p(1/\sqrt{n}).$$

Now theorem (11) [$\theta_n-\theta_0=O_p(1\sqrt{n})$], and the combination of A3 and (12b) [which together imply $\theta^*_n-\theta_0=O_p(1\sqrt{n})$], give the desired result. Consequently

$$\| L_n(\theta^*_n) \| \le \| L_n(\theta_n) \| + o_p(1/\sqrt{n}) = \| \mathcal{G}_n(\theta_n) \| + o_p(1/\sqrt{n})$$
$$\le \| \mathcal{G}_n(\theta^*_n) \| + o_p(1/\sqrt{n}) = \| L_n(\theta^*_n) \| + o_p(1/\sqrt{n}),$$

where the two inequalities follow from the fact that that θ^*_n and θ_n minimize $\| L_n(\theta) \|$ and $\| \mathcal{G}_n(\theta) \|$ respectively. Now note that because θ^*_n is obtained as a projection

$$L_n(\theta_n) = L_n(\theta^*_n) + D(\theta_n-\theta^*_n), \quad \text{with} \quad (\theta_n-\theta^*_n)'D'L_n(\theta^*_n) = 0.$$

As a result

$$\| L_n(\theta_n) \| = \| L_n(\theta_n^*) \| + \|D(\theta_n - \theta_n^*)\|.$$

Combining this with the fact, from above, that $\|L_n(\theta_n)\| - \|L_n(\theta_n^*)\| = o_p(1/\sqrt{n})$, gives us

$$\|D(\theta_n - \theta_n^*)\| = o_p(1/\sqrt{n}),$$

which, since D has full rank, implies (14). The theorem is then a direct implication of A3, and the Lindberg Levy central limit theorem (Rao,1973, section 2c.5). □.

We now return to assumption 3. It requires that

$$\sqrt{n}\ \mathcal{G}_n(\theta_0) = (1/\sqrt{n})\ \Sigma_i \{ m(z_i,\theta_0) + M(z_i,\theta_0)[h_n(v_i,\theta_0) - h_0(v_i,\theta_0)] + M(z_i,\theta_0)H_0(v_i,\theta_0)[\tau_n(v_i) - \tau_0(v_i)]\},$$

has a limit Normal distribution. It will be sufficient therefore to provide conditions which insure that

$$(15a) \qquad \sqrt{n}^{-1}\ \Sigma_i M(z_i,\theta_0)[h_n(v_i,\theta_0) - h_0(v_i,\theta_0)] = \sqrt{n}^{-1}\ \Sigma_i f_1(z_i) + o_p(1),$$

and

$$(15b) \qquad \sqrt{n}^{-1}\ \Sigma_i M(z_i,\theta_0)H_{0t}(v_i,\theta_0)[\tau_n(v_i) - \tau(v_i)] = \sqrt{n}^{-1}\ \Sigma_i f_2(z_i) + o_p(1),$$

with

$$E[f_j(z_i)] = 0, \quad \text{and } E[f_j(z_i)^2] = \sigma_j^2 < \infty,$$

for j=1,2.

That is if (15) is satisfied then

$$\sqrt{n}\ \mathcal{G}_n(\theta_0) = (1/\sqrt{n})\Sigma_i \{\ m(z_i,\theta_0) + f_1(z_i) + f_2(z_i)\} + o_p(1),$$

so the Lindberg Levy central limit theorem, together with our assumption on the boundedness of E $m(z_i,\theta_0)m(z_i,\theta_0)'$, will imply assumption 3 with

$$(16)\quad V = E\{\ [m(z_i,\theta_0) + f_1(z_i) + f_2(z_i)][m(z_i,\theta_0) + f_1(z_i) + f_2(z_i)]\ '\}.$$

Section II discusses conditions which insure 15a and 15b when $\tau_n(\cdot)$ and $h_n(\cdot)$ are either series or kernel estimators of regression functions; i.e., when there exists random variables $y_1(z,\tau_0,\theta_0)$ and $y_2(z)$ such that

$$(17a)\quad E[y_1(z_i,\tau_0,\theta_0)\,|\,v] = h_0(v_i,\tau_0,\theta_0),\ \text{ and }\ E[y_2(z_i)\,|\,v_i] = \tau_0(v_i).$$

In this case, given sufficient regularity, (15) holds with

$$(17b)\quad f_1(z_i) = M(v_i,\theta_0)[y_1(z_i,\tau_0,\theta_0) - h_0(v_i,\tau_0,\theta_0)],\ \text{ and}$$
$$(17c)\quad f_2(z_i) = M(v_i,\theta_0)H_{0t}(v_i,\tau_0,\theta_0)[y_2(z_i) - \tau_0(v_i)],$$

Not surprisingly the contribution of the variance in the estimate of the nonparametric component to the variance of the estimator of θ depends directly on both the conditional variance in the unknown regression functions, and on the derivative of the moment condition with respect to the value of that regression function. We note that Andrews(1991,1993,1994) and Newey (1991 and forthcoming) consider a variety of other cases (including densities and integrals of conditional expectations, as well as derivatives of these objects), and alternative proofs of assumption (3). Also Newey (forthcoming,b) considers an extension wherein the object enterring into the moment conditions is not the value of the unknown function per se (our h_0), but rather a functional of h_0 (eg. an integral,

or weighted integral, of h_0 over a subset of $\mathcal{V}$).

Theorems 9,11, and 14 insure the consistency and asymptotic normality of our semiparametric estimator given assumptions R, 1 and 3. Before going on to sets of primitive conditions which insure that the latter two assumptions are satisfied it is useful to pause and review the notion of stochastic equicontinuity. This for two reasons. First, as noted above, existing results allow us to use the notion of stochastic equicontinuity to verify lemma 10 directly under less restrictive assumptions then those contained in assumption R. In addition, that notion underlies much of the literature on primitives which suffice for our assumption 3 above (the normality assumption; these conditions are reviewed in the next section). On the other hand, stochastic equicontinuity is not used directly in the proofs of our major results. Thus the reader who is willing to suffice with references for the primitives which insure our assumptions should be able to omit this discussion (go directly to Lemma 20) and have no trouble with the rest of the paper.

Let

$$\begin{aligned} \mathrm{m}_1(\mathrm{z},\theta) &\equiv \mathrm{m}(\mathrm{z},\theta), \\ \mathrm{m}_2(\mathrm{z},\theta,h) &\equiv \mathrm{M}(\mathrm{z},\theta)[h(\mathrm{v},\theta)-h_0(\mathrm{v},\theta)] , \\ \mathrm{m}_3(\mathrm{z},\theta,\tau) &\equiv \mathrm{M}(\mathrm{z},\theta)H_{0\mathrm{t}}(\mathrm{v},\theta)[\tau(\mathrm{v})-\tau_0(\mathrm{v})], \end{aligned}$$

define the following families of functions,

$$\begin{aligned} M_1 &= \{\mathrm{m}_1(\cdot,\theta),\ \theta \in \Theta\}, \\ M_2 &= \{\mathrm{m}_2(\cdot,\theta,h),\ \theta \times h \in \Theta \times \mathcal{H}\}, \text{ and} \\ M_3 &= \{\mathrm{m}_3(\cdot,\theta,\tau),\ \theta \times \tau \in \Theta \times \Upsilon\} \end{aligned}$$

and endow each with the metric

$$(18) \qquad \rho_j(\gamma_{1j},\gamma_{2j}) = \|m_j(\gamma_{1j})-m_j(\gamma_{2j})\|$$

where $\|f(\gamma_1)\| = [Ef(z,\gamma_1)^2]^{1/2}$, and $\gamma_j \in \Gamma_j$ is the appropriate index set ($\Gamma_1=\Theta$, $\Gamma_2=\Theta \times \mathcal{H}$, $\Gamma_3 = \Theta \times \Upsilon$).

The empirical processes we associate with these families and the random sequence $\{z_i\}$ are defined as

$$(19a) \qquad \nu_{1n}(\theta) = (1/\sqrt{n}) \Sigma\, [m(z,\theta) - Em(z,\theta)],$$

$$(19b) \qquad \nu_{2n}(\theta \times h) = (1/\sqrt{n}) \Sigma\, [m_2(z,\theta,h) - Em_2(z,\theta,h)], \text{ and}$$

$$(19c) \qquad \nu_{3n}(\theta \times \tau) = (1\sqrt{n}) \Sigma\, [m_3(z,\theta,\tau) - Em_3(z,\theta,\tau)]$$

where here and below all expectations are with respect to P.

An empirical process defined on a metric space is called stochastically equicontinuous with respect to its index set if for any two sequences of random indices, say $\{\gamma_{1n}\}$, $\{\gamma_{2n}\}$, we have $\nu_n(\gamma_{1n})-\nu_n(\gamma_{2n})=o_p(1)$, whenever $\rho(\gamma_{1n},\gamma_{2n})=o_p(1)$ [this presumes that both γ_{1n} and γ_{2n} will be in the space with probability tending to one]. There is a reasonably large literature which provides metric spaces of functions which are stochastically equicontinuous in their index sets, and sets out rules which allow one to combine functions from different spaces to produce new spaces which inherit the property of equicontinuity from the original spaces (for a good overview with econometric applications see Andrews,forthcoming, and the literature cited there).

To illustrate the usefulness of the notion of stochastic equicontinuity go back to (10*) and note that the first term in that expression is just

$$\nu_{1n}(\overline{\theta}_n)-\nu_{1n}(\theta_0).$$

I.e. provided the process $\{\nu_{1n}(\theta):\theta\in\Theta\}$ is stochastically equicontinuous the first term in

(10*) is $o_p(1)$, which is small enough for lemma 10. The condition that $\{\nu_{1n}(\theta):\theta\in\Theta\}$ be stochastically equicontinuous has a simple interpretation. That process provides the disturbance generated by $\sqrt{n}$ times the difference between the sample and the population mean of a (random) function that is indexed by a parameter value, at different values of that parameter. $\nu_{1n}(\overline{\theta}_n)-\nu_{1n}(\theta_0)$ is comparing the value of that disturbance at two values in the index set. Stochastic equicontinuity implies that whenever the difference in the index set values converges in probability to zero, the difference in the disturbance also converges to zero. We would expect this to be true if the underlying families of functions were sufficiently "smooth" in the index set. This is the case, and our differentiability conditions together with the assumption that Θ is bounded insure that $\{\nu_{1n}(\theta):\theta\in\Theta\}$ is stochastically equicontinuous (see Pakes and Pollard, lemmas 2.13 and 2.17). Continuity of $m(\cdot,\theta)$ is not, however, necessary for the stochastic equicontinuity of $\{\nu_{1n}(\theta):\theta\in\Theta\}$ (for a more detailed discussion see Pollard,1991, and the literature cited their). As a result one can use methods of proof similar to those provided here for problems with discontinuities in the objective function (see for eg. Pakes and Pollard,1989).

Provided that $\tau(v)$ is sufficiently smooth in v, one can also show that our assumptions imply that $\{\nu_{3n}(\theta\times\tau):\ \theta\times\tau\in\Theta\times\Upsilon\}$ is stochastically equicontinuous in its index set (see below). Since the third term in (10*) can be written as $[\nu_{3n}(\theta_n\times\tau_n)-\nu_{3n}(\theta_0\times\tau_n)]-\sqrt{n}E\{m_3(z,\theta_n,\tau_n)-m_3(z,\theta_0,\tau_n)\}$, and assumption 1 implies that $\rho_3[(\theta_n\times\tau_n),(\theta_0\times\tau_n)]=o_p(1)$, if we prove that $\sqrt{n}E\{m_3(z,\theta_n,\tau_n)-m_3(z,\theta_0,\tau_n)\}=o_p(1)[1+\|\sqrt{n}(\overline{\theta}_n-\theta_0)\|]$, we will have shown that the third term in (10*) is $o_p(1)$. An argument similar to the proof of the first part of lemma 10 shows that this expectation satisfies the needed condition.

The second term in (10*) is a bit more problematic. Since each h is indexed by θ, and m_2 is indexed by h, the process $\{\nu_{2n}(\theta\times h):\ \theta\times\mathcal{H}\in\Theta\times\mathcal{H}\}$ is constructed by composing functions of one index set with functions of another index set and we do not know of a general rule which implies that such a composite family inherits stochastic equicontinuity. What the proof of lemma 10 shows is that the existence of integrable envelopes for the

derivatives of both $M(z,\theta)$ and $H^{\theta}(v,\theta)$ with respect to θ in a neighborhood of θ_0 imply that $\nu_{2n}(\theta_n \times h_n) - \nu_{2n}(\theta_0 \times h_n) = o_p(1)[1+\|\sqrt{n}(\overline{\theta}_n - \theta_0)\|]$, provided $\sqrt{n}E\{m_2(z,\theta_n,h_n) - m_2(z,\theta_0,h_n)\} = o_p(1)[1+\|\sqrt{n}(\overline{\theta}_n - \theta_0)\|]$, and that the latter is indeed true.

Finally note that the asymptotic normality assumption (3 above) depends only on $\nu_{1n}(\theta_0)$, $\nu_{2n}(\theta_0 \times h_n)$, and $\nu_{3n}(\theta_0 \times \tau_n)$ (and the expectations of the latter two processes). As a result in reviewing conditions for assumption 3 below we can set $\theta = \theta_0$ and worry only about the equicontinuity of $\nu_{2n}(\cdot)$ and $\nu_{3n}(\cdot)$ as we vary τ_n and h_n. We provide conditions which insure that $\{\nu_{2n}(\theta_0 \times h): h \in \mathcal{H}\}$ and $\{\nu_{3n}(\theta_0 \times \tau): \tau \in \Upsilon\}$ are stochastically equicontinuous in the next section.

There are two other points that we would like to make before concluding this section. First we illustrate how the results simplify when alternative orthogonality conditions hold. The first is a condition similar to that introduced by Andrews(1994) [and our result here is similar to Theorem 5.4 of Newey(1991)], and allows us to specify conditions under which we need not adjust the variance–covariance matrix of our estimate of θ for the fact that we have used estimated, rather than the actual, values of our unknown functions in our moment conditions. The second allows us to specify conditions under which we need not adjust the variance covariance of θ for the fact that $\tau(\cdot)$ has been estimated, even though we *do* need to make an adjustment for the fact that $h(\cdot)$ has been estimated. Both of these conditions seem to occur frequently in empirical applications (see section III below).

These orthogonality conditions are a direct result of the following asymptotically equivalent expression for $\sqrt{n}\,\mathcal{G}_n(\theta_0)$.

20. Lemma.

Let

$$M(v,\theta) = E\,[M(z,\theta)\,|\,v],$$

$$\mathrm{MH}(\mathrm{v}_2,\theta_0) = \mathrm{E}\,[\mathrm{M}(\mathrm{z},\theta_0)H(\mathrm{v},\theta_0)\,|\,\mathrm{v}_2],$$

assume that both $\{\nu_{2n}(\theta_0\times h)\colon h\in\mathscr{H}\}$ and $\{\nu_{3n}(\theta_0\times\tau)\colon\tau\in\Upsilon\}$ are stochastically equicontinuous, and recall that $\tau_0(\mathrm{v})=\tau_0(\mathrm{v}_2)$ a.e. P. Then

$$\sqrt{n}\ \mathscr{G}_n(\theta_0) = \nu_{1n}(\theta_0) + \sqrt{n}\int \mathrm{M}(\mathrm{v},\theta_0)[h_n(\mathrm{v},\theta_0)-h_0(\mathrm{v},\theta_0)]\mathrm{P}(d\mathrm{v})$$
$$+ \sqrt{n}\int \mathrm{MH}(\mathrm{v}_2,\theta_0)[\tau_n(\mathrm{v}_2)-\tau_0(\mathrm{v}_2)]\mathrm{P}(d\mathrm{v}_2) + o_p(1).$$

Proof.

Note that equation (5), the definitions of the empirical processes in (19), and the equicontinuity conditions combined with assumption 1 allow us to write

$$\sqrt{n}\ \mathscr{G}_n(\theta_0) = \nu_{1n}(\theta_0) + \sqrt{n}\int \mathrm{M}(\mathrm{z},\theta_0)[h_n(\mathrm{v},\theta_0)-h_0(\mathrm{v},\theta_0)]\mathrm{P}(d\mathrm{z})$$
$$+ \sqrt{n}\int \mathrm{M}(\mathrm{z},\theta_0)H(\mathrm{v},\theta_0)[\tau_n(\mathrm{v})-\tau_0(\mathrm{v})]\mathrm{P}(d\mathrm{z}) + o_p(1).$$

We consider only the second term in this expression. The third term can be treated in an analogous way.

$$\sqrt{n}\int \mathrm{M}(\mathrm{z},\theta_0)[h_n(\mathrm{v},\theta_0)-h_0(\mathrm{v},\theta_0)]\mathrm{P}(d\mathrm{z})$$
$$= \sqrt{n}\int [\mathrm{M}(\mathrm{z},\theta_0)-\mathrm{M}(\mathrm{v},\theta_0)][h_n(\mathrm{v},\theta_0)-h(\mathrm{v},\theta_0)]\mathrm{P}(d\mathrm{z})$$
$$+ \sqrt{n}\int \mathrm{M}(\mathrm{v},\theta_0)[h_n(\mathrm{v},\theta_0)-h_0(\mathrm{v},\theta_0)]\mathrm{P}(d\mathrm{z}).$$

But

$$\left|\sqrt{n}\int [\mathrm{M}(\mathrm{z},\theta_0)-\mathrm{M}(\mathrm{v},\theta_0)][h_n(\mathrm{v},\theta_0)-h_0(\mathrm{v},\theta_0)]\mathrm{P}(d\mathrm{z})\right| \le$$
$$\sup_{h\in\mathscr{H}}\left|\sqrt{n}\int [\mathrm{M}(\mathrm{z},\theta_0)-\mathrm{M}(\mathrm{v},\theta_0)][h(\mathrm{v},\theta_0)-h_0(\mathrm{v},\theta_0)]\mathrm{P}(d\mathrm{z}\,|\,\mathrm{v})\mathrm{P}(d\mathrm{v})\right| = 0,$$

which completes the proof. □.

The equicontinuity conditions given in the statement of the lemma are generally also required for Assumption 3, and so do not really detract from the generality of the lemma. Sufficient conditions for the equiconinuity conditions are given in the next section.

21.Corollary. (orthogonality conditions).

(a) If $M(v,\theta_0)=0$ a.e. P, assumptions R and 1 together with the consistency result imply the limit theorem in 14, with $V=Em(z_i,\theta_0)m(z_i,\theta_0)'$.

(b) If $MH(v_2,\theta_0)=0$ a.e. P, then assumptions R and 1 together with the consistency result and (15a) imply the limit theorem in 14, with $V=E\,[m(z_i,\theta_0) + f_1(z_i)][m(z_i,\theta_0) + f_1(z_i)]'$.

Proof.

We prove only part (a). The proof of part (b) is similar. If $M(v,\theta_0)=0$ then

$$\sqrt{n}\ \mathcal{G}_n(\theta_0) = (1/\sqrt{n})\ \Sigma\ m(z_i,\theta_0) + o_p(1).$$

But then assumption (4) is an immediate consequence of our regularity conditions, since they imply the conditions of the Lindberg–Levy central limit theorem for $(1/\sqrt{n})\ \Sigma\ m(z_i,\theta_0)$. □.

The impact on the variance of our estimate of θ_0 of substituting an estimator (rather than the actual value) of $h(\cdot)$ in the moment condition depends on the derivative of the moment conditions with respect to $h=h(\cdot)$ at the true value of the estimated parameters [it depends on $M(\cdot,\theta_0)$]. Moreovoer since $h(\cdot)$ only takes on distinct values at distinct values of v, it only depends on $\int M(z,\theta_0)P(dz\,|\,v)$. The condition in 21(a) states that $\int M(z,\theta_0)P(dz\,|\,v) = 0$ (a.e. P), and therefore implies that we do not need to adjust the variance covariance matrix of our estimates of θ_0 for the fact that we have used estimates

of h_0 and τ_0 rather than their actual values. Condition (b) insures that the derivative with respect to $t=\tau(\cdot)$ is zero a.e.P, even though the derivative with respect to $h(\cdot)$ may not be, in which case we do not have to adjust for the fact that τ_0 is estimated (though we do have to adjust for the fact that h_0 is estimated). In general, however, an adjustment for the perturbation induced by the variance in both estimated nonparametric components is required.

Finally we note that there was no need to use Euclidean norm in the analysis. For a nonsingular matrix A, define a norm by $\|x\|_A=\|Ax\|$, and let $\{A_n(\theta)\}$ be a sequence of matrices whose elements are random variables that depend on θ. Lemmas 3.4 and 3.5 in Pakes and Pollard (1989) provide conditions on $\{A_n(\theta)\}$ which insure that our limit theorems will be true if, instead of minimizing $\|G_n(\theta)\|$, we minimized $\|G_n(\theta)\|_{A_n(\theta)}$. As in Hansen (1982), and discussed in the context of semiparametric estimators in Newey (1991), the asymptotic efficiency of one's estimator can be improved by an appropriate choice of norm.

II. Primitives for the Assumptions of Section I.

There is a large literature with alternative sets of conditions which suffice for the assumptions used in Section I; too large to be surveyed in a subsection of this paper. Our strategy will be to provide a brief verbal exposition of a set of conditions which suffice when the h_n and the τ_n functions which appear in the moment condition in (3) are nonparametric estimates of regression functions, emphasizing the aspects of those conditions which impact on the actual computation of the estimator, and then refer the reader to references which deal with a variety of other cases. Notationally recall that: $v'=[v_1',v_2'] \in \mathcal{V} \subset \mathbb{R}^{dv}$ with $v_i \in \mathcal{V}_i \subset \mathbb{R}^{dvi}$ for $i=1,2$; $\tau_0 \in \Upsilon$, a pseudo metric space of functions from $\mathbb{R}^{dv2} \to T \subset \mathbb{R}^1$; $h_0 \in \mathcal{H}$ a pseudo metric space of functions from $\mathbb{R}^{dv1} \times \mathbb{R}^1 \times \Theta \to H \subset \mathbb{R}^1$; and that $\mathcal{V} \subset \mathbb{R}^{dv}$ defines the trimming function $I\{v \in \mathcal{V}\}$ which sets $m(\cdot)$ to zero if

$v \notin \mathcal{V}$.

We begin with conditions which insure that assumption 1 is satisfied when our h_n and τ_n are *kernel* estimates of the regression functions, h_0 and τ_0. Good introductions to kernel estimators, introductions which include discussions of computational techniques and their performance as well as of the analytic properties of the estimates, can be found in Silverman,1986, Hardle,1991, and Stoker,1991b. What we require is appropriate rates of convergence for those estimates. In particular assumptions 1a and 1c require that

$$\text{(A1a)} \quad n^{a_1} \sup_{(v \in \mathcal{V}, \theta \in \Theta)} \| h_n(v,\theta) - h_0(v,\theta) \| = O_p(1)$$

$$\text{(A1c)} \quad n^{a_3} \sup_{(v \in \mathcal{V}, \theta \in \Theta)} \| H_{nt}(v,\theta) - H_{0t}(v,\theta) \| = O_p(1),$$

where $a_1 > 1/4$, and $a_1 + a_3 > 1/2$, while assumption 1d requires

$$\text{(A1d)} \quad \sup_{(v \in \mathcal{V})} \| H_n^{\theta}(v,\theta_0) - H_0^{\theta}(v,\theta_0) \| = o_p(1).$$

[Recall that $h(v,\theta) \equiv h[v_1, \tau_0(v_2), \theta]$, $H_t(v,\theta) \equiv \partial h(v_1,t,\theta)/\partial t_{|t=\tau_0(v_2)}$, and $H^{\theta}(v,\theta) \equiv \partial h(v,\theta)/\partial \theta$].

Note that in (A1a) and (A1c) the convergence must be uniform over the product space constructed from $\mathcal{V}$, a subset of the support of v, and the index set Θ. Andrews (1993) details conditions which insure rates of convergence of kernel estimators of functions of the data and an index set which are uniform over this product space. The dependence on the index set is built in through the construction of regressors or dependent variables that depend on the value of the index. This is general enough to cover all empirical examples we are aware of. Accordingly, assume that there is a *known* vector function $x(\cdot)$: $\mathbb{R}^{d_v} \times \mathbb{R}^1 \times \Theta \to \mathbb{R}^x$, and a known function $y(\cdot)$: $\mathbb{R}^{d_z} \times \Theta \to \mathbb{R}^1$ such that

$$h_0\{x_1[v,\tau_0(v_2),\theta],\ldots,x_x[v,\tau_0(v_2),\theta],\theta\} = E[y(z,\theta) \mid v].$$

The estimate of h_0, $h_n(\cdot)$, is obtained as a kernel estimate of the regression of $y(\cdot)$ on $x(\cdot)$, while

$$H_{0t}(\cdot) \equiv \Sigma_j [\partial h_0[\cdot, x_j, \cdot]/\partial x_j |_{x_j = x_j(\cdot)}] \partial x_j(\cdot, t, \cdot)/\partial t |_{t=\tau_0(\cdot)},$$

and $H_{nt}(\cdot)$, $H_0^{\theta}(\cdot)$, and $H_n^{\theta}(\cdot)$ are obtained analogously.

We will assume that each $x_j[v, \tau_0(v_2), \theta]$ is differentiable in $t=\tau_0(v_2)$ and in θ, and that these derivatives are bounded uniformly over $(v, \theta) \in \mathcal{V} \times \Theta$. Thus to prove both A1c and A1d it will suffice to show that

$$(A1c') \qquad n^{a_3} \sup_{(x,\theta)} \| \partial h_n(x,\theta)/\partial x - \partial h_0(x,\theta)/\partial x \| = O_p(1),$$

where the supremum is taken over $(x, \theta) \in X \times \Theta$ and

$$X = \{x \in \mathbb{R}^{dx} : x = x[v, \tau_0(v_2), \theta] \text{ for some } v \in \mathcal{V} \text{ and some } \theta \in \Theta\}$$

[note that since $x(\cdot)$ is continuous, if Θ and $\mathcal{V}$ are compact, so is X]. To focus our discussion we will carry along the example in the next section wherein $x[v, \tau_0(v_2), \theta] = [v, \tau_0(v_2)]' \theta$, and τ_0 is continuous in v_2.

We note that Andrews(1993) also provides conditions which insure rates of uniform convergence for the more general case where *both* $y(\cdot)$ and $x(\cdot)$ depend on an infinite (rather than on a finite) dimensional index set [here only $x(\cdot)$ does], and for cases in which h_0 is a density function, or a higher order derivative of either a density function or of a conditional expectation[3].

The sufficient conditions we provide include: i) additional smoothness conditions on

---

[3]Conceptually the proofs for the other cases are quite similar to those for the regression function case, but to deal with them we would have to add alot of notation.

the functions of interest [$y(\cdot)$,$x(\cdot)$,and $h(\cdot)$; additional to the conditions in assumption R]; ii) additional restrictions on the choice of the kernel (conditions which insure a smooth bias–reducing kernel of appropriate order)[4]; and iii) restrictions on X [we will choose $\mathcal{Y}$, and hence X, so that the density of $x(v,\tau_0,\theta)$ is bounded away from zero over all $v\in\mathcal{Y}$ and $\theta\in\Theta$; this will trim out subsets of the data which could generate values of x with densities that are too low to allow us to obtain precise enough estimates of the regression function, h_0]. We first provide a summary of the conditions we need (this is a verbal summary of material in section 4 of Andrews simplified for the case we are considerring), and then comment on their impact on the computation of the estimator. The rates provided here are not necesserily sharp, nor is there any claim that the conditions we are providing are in any sense minimal.

We summarize on the smoothness conditions first. In addition to assumption R, it is assumed that for each $\theta\in\Theta$: $x(\cdot)$ has a density w.r.t. Lebsegue measure [say $f_x(x)$]; that both $h(x,\theta)$ and $f_x(x)$ are continuously differentiable in x with bounded derivative on $\mathbb{R}^x$ [the bound being uniform in (x,θ)] to order at least 2 (further differentiability conditions on these functions are given below); that both $h(x,\theta)$ and $x(v,\theta)$ are differentiable in θ (a.e.) with derivatives that are bounded by square integrable functions of v; and that for some $\kappa>0$ and every $\theta\in\Theta$, $y(z,\theta)$ is both bounded by a function of z which is integrable to power $2+\kappa$, and has a derivative with respect to θ which is bounded by a square integrable function of z. For the example in the next section $x[v,\tau_0(v_2),\theta] = [v,\tau_0(v_2)]'\theta$, $h(x,\theta)=h(x)$, $y(z,\theta)=z_1-z_2\theta$ and this satisfies all our conditions provided $\tau_0(v_2)$, and the density of v, $f_v(v)$, are sufficiently smooth in their arguments.

The kernel estimate of the density and regression functions at point x are given, respectively, by

---

[4]We shall focus on the use of bias reducing kernels to insure a given rate of uniform convergence for our estimates, but alternative bias reduction techniques could also have been used (see for eg. Newey, Hsieh, and Robins,1993, or Jones, Linton, and Robins,1993).

$$f_n(x,\theta)=n^{-1}\Sigma_i \hat{K}[(x-x(v_i,\theta))/\sigma_n]/(\sigma_n)^k, \text{ and}$$
$$h_n(x,\theta)=\{n^{-1}\Sigma_i\, y(z_i,\theta)\hat{K}[(x-x(v_i,\theta))/\sigma_n]/(\sigma_n)^k\}/f_n(x,\theta),$$

where $\hat{K}(x)=\det\hat{\Omega}^{-1/2}\, K(\hat{\Omega}^{-1/2}x)$, $K(\cdot)$ is a non random function on $\mathbb{R}^x$, σ_n is a (possibly random) bandwidth parameter, and $\sqrt{n}(\hat{\Omega}-\Omega)=O_p(1)$ for some positive definite matrix Ω. Usually $\hat{\Omega}$ is taken to be an estimate of the variance covariance matrix of x, or a diagonal matrix with estimates of the variances of the individual elements of x on the diagonal.

Letting μ be a vector of nonnegative integers with $|\mu| = \Sigma_j \mu_j$, the smoothness and bias–reducing conditions on the kernel $K(\cdot)$ can be formulated as follows: $\int K(x)dx = 1$; $\int x^{\mu}K(x)dx = 0$, where $x^{\mu} \equiv x_1^{\mu_1} \times \ldots \times x_x^{\mu_x}$ for all μ with $|\mu|\leq b$; $\partial^{|\mu|}K(x) / \partial^{\mu_1}x_1 \times \ldots \times \partial^{\mu_x}x_x$ is bounded (uniformly in x) and continuous for all μ with $|\mu|\leq b+1$; and $K(x)\to 0$ as $\|x\| \to \infty$. Here b is the order of bias reduction of the kernel. To allow for data dependent bandwidths (eg. cross validation, see our example below), σ_n is permitted to be random, but there is assumed to be two sequences of numbers, $\{\sigma_{in}\}$ for i=1,2, and constants C_1 and C_2, such that $C_1\sigma_{1n}\leq\sigma_n\leq C_2\sigma_{2n}$, with probability tending to one. We assume $\sigma_{in}=n^{-\psi}$ in what follows, and consider setting ψ to enable us to insure the rate conditions in (A1a) and (A1c') by appropriate choice of b.

Recall that what we need to insure is a convergence rate for $h_n(x,\theta)$ and for $\partial h_n(x,\theta)/\partial\theta$ that is uniform over $(x,\theta)\in(X,\Theta)$. To obtain them we will need to also insure convergence rates for the density $f_n(x,\theta)$ and the derivative of the density $\partial f_n(x,\theta)/\partial\theta$ which are uniform over the same space. We consider first the convergence rate for a density or a regression function at a given point (and of their derivatives at that point; more detail on these calculations can be found, for eg., in Silverman,1986). To do so fix x and compute the squared bias and the variance of the kernel estimates of $E[y(z,\theta)|x]f_x(x)$ and $f_x(x)$ as a function of the bandwidth, the degree of bias reduction of the kernel (b), and the sample size. The bias will be increasing, and the variance will be decreasing, in the bandwidth. Given a sample size, and a degree of bias reduction for the kernel, one can

calculate the bandwidth which minimizes the mean sqare error of the estimates. The rate at which these bandwidths go to zero as we increase sample size, or ψ in the formula above, is $1/(2b + x + 2q)$, where, q is the order of the derivative we need to estimate consistently (in our case q is one or zero), and, x is the number of regressors. This presumes that both the density and the regression function are differentiable to order (b+1) in x. To form the estimate of the regression function we divide the estimate $E[y(z,\theta)|x]f(x)$ by the estimate of f(x). For this procedure to produce a consistent estimate of $E[y(z,\theta)|x]$, the density,f(x), must be bounded away from zero. The rate of convergence of the kernel estimate of the regression function obtained in this way is then $1/[2 + (x+2q)/(b+1)]$. If (in addition to the assumptions stated above): i)these conditions [both the density and the regression function are differentiable to order b+1, and $f(x)\geq\kappa>0$] are met at every $(x,\theta)\in X \times \Theta$; ii)the latter space is totally bounded; and iii) $\|y(z,\theta_1)-y(z,\theta_2)\|\leq\rho(z)\|\theta_1-\theta_2\|$ for a square integrable $\rho(\cdot)$ and the kernel estimate of the regression function is continous uniformly over both $x\in X$ and $h_n(\cdot)\in\mathcal{H}$ (the space of possible estimates of the regression function), then the rate given above will also be uniform. That is, these conditions insure that the a_1 in (A1a) will be $1/\{2 + [x/(b+1)]\}$ and the a_3 in (A1c') will be $1/[2 + (x+2)/(b+1)]$. So for $a_1>1/4$ we require $b > x/2 -1$, and for $a_1+a_3>1/2$, $b\geq x/2$ is more than enough.

A few comments on empirical implementation are in order. First note that the normal based system of bias–reducing kernels discussed in Bierens (1987, equation 2.2.36) will satisfy the smoothness and bias reduction conditions in these assumptions provided his bias reduction parameter is set appropriately. Second, the trimming conditions are placed on the domain of x rather than on the domain of v. In particular it is assumed that there is a bounded subset of $\mathbb{R}^x$, say X_2, and a $\kappa>0$ with an associated $X_1 = \{ x\in\mathbb{R}^x : \inf_{\theta\in\Theta} f(x,\theta) \geq \kappa\}$, and then shown that if $X = X_1 \cap X_2$ the convergence will be uniform over $x \in X$. Thus setting $\mathcal{V}= \{v: x(v,\theta) \in X \text{ for all } \theta \in \Theta\}$ will insure that the trimming conditions are satisfied. lly note that a special case of this framework is $x(v,\theta)=v_2$ for all θ, which implies that conditions analogous to those given above will also imply that $\tau_n(v)$ is also

obtained as a kernel estimator

$$\text{(A1b)} \quad n^{a_2} \sup_{(v \in \mathcal{V})} \| \tau_n(v) - \tau_0(v) \| = O_p(1),$$

with $a_2 > 1/4$ and $\alpha_2 + \alpha_3 > 1/2$.

Both Andrews(1991a) and Newey(forthcoming) provide conditions which suffice for (A1b), that is for convergence rates that are uniform over a subset of the data, for *series* estimates of regresssion functions and their derivatives. We do not review the details here[5]. We are *not aware* of any results on rates of convergence for series estimators which are also uniform over an index set (as required by assumptions A1a and A1c).

A brief summary on assumption 1 is in order before proceeding. The conditions detailed above insure that it is indeed satisfied if τ_n and h_n are estimated using bias–reducing kernels of appropriate orders. Conditions are available which insure that assumption 1 will also be satisfied if τ_n is estimated using a series estimator, and h_n is estimated using a kernel. Finally, there *may* be conditions which insure the required convergence rates if *both* τ_n and h_n are estimated using series estimators, but we are not aware of a formal statement of them (though the empirical example in the next section provides some indication that this may just be a technical problem).

Next we briefly review conditions which insure that the empirical processes generated by the families $M_1 = \{m_1(z,\theta) : \theta \in \Theta\}$, $M_{20} = \{m_2(z,\theta_0,h) : h \in \mathcal{H}\}$ and $M_3 = \{m_3(z,\theta,\tau) : \theta \times \tau \in \Theta \times \mathcal{T}\}$ are stochastically equicontinuous in their index sets. We begin by noting that a sufficient condition for a family of functions to generate an empirical process which is stochastically equicontinuous is that the family have an envelope which is

---

[5]They require; i)smoothness of the underlying functions, ii)conditions on the sequence of basis functions to insure invertibility and convergence, iii)rates at which the number of basis functions used in the series expansion must grow as a function of sample size (because of arbitrary constants, these do not constrain the choice of the number of terms used in any empirical example), and iv) bounds on the density over the set $\mathcal{V}$.

square integrable with respect to P and satisfy Pollard's (1991) entropy condition. F is an envelope for a family $\mathcal{F}$ if $F \geq |f|$ for all $f \in \mathcal{F}$. The existence of a square integrable envelope for our families follows directly from Assumption R. Thus all that is needed is to show that M_1, M_{20}, and M_3 satisfy the entropy conditon.

From (13), the M_3 family of functions is formed as the product of two functions one of which depends only on the finite dimensional component of the index set, and the other depends only on the infinite dimensional component. That is if $f \in M_{30}$, then $f = f_{31} \otimes f_{32}$ where $f_{31} \in \mathcal{F}_{31} = \{M(\cdot,\theta)H_{0t}(\cdot,\theta);\ \theta \in \Theta\}$, $f_{32} \in \mathcal{F}_{32} = \{\tau(\cdot) - \tau_0(\cdot);\ \tau \in \Upsilon\}$, and $\otimes$ is the product operator. Since families formed as the product of elements of two different families each of which satisfies Pollard's entropy condition will satisfy Pollard's entropy condition, the empirical process $\{\nu_{3n}(\theta \times \tau): \theta \times \tau \in \Theta \times \Upsilon\}$ will be stochastically equicontinuous if each of the primitive families of functions satisfies the entropy condition. A finite dimensional family satisfies Pollard's entropy condition if it is Euclidean in the sense of Pakes and Pollard(1989). For our example it suffices to note that a family is Euclidean if it is differentiable in its index set (Θ), and Θ is bounded (Pakes and Pollard, lemma 2.13). An infinite dimensional family will satisfy Pollard's entropy condition if it is a type III family of functions as defined by Andrews (forthcoming, section 3). An infinite dimensional family from a subset of $\mathbb{R}^{dv2}$ into $T \subset \mathbb{R}$ will be a type III family provided each of its members is differentiable to order greater than $d_{v2}/2$, say $[d_{v2}/2]$, with derivatives of order $d_{v2}/2$ that satisfy a Lipschitz condition on a compact connected subset of $\mathbb{R}^{dv2}$ and take on a constant value outside of this subset, say $\mathcal{V}_2^*$. Recall that a compact $\mathcal{V}$ defines our trimming set, so that provided the set $\mathcal{V}_2 = \{v_2 \in \mathbb{R}^{dv2}: v=(v_1,v_2), \text{ for some } v \in \mathcal{V}\} \subset \mathcal{V}_2^*$, it is irrelvant how $\tau(\cdot)$ is defined outside of $\mathcal{V}_2^*$. That is, all we require is the ability to form a compact connected subset of $\mathbb{R}^{dv2}$ on which $\tau(\cdot)$ is sufficiently smooth and which contains $\mathcal{V}_2$.

Similar smoothness conditions on the family M_{20} guarantee that the process $\{\nu_{2n}(\theta_0 \times h):\ h \in \mathcal{H}\}$ is stochastically equicontinuous. Finally the Euclidean argument for

finite dimensional index sets given above also establishes that the empirical process $\nu_1(\theta)$ formed from M_1 is stochastically equicontinuous in θ. For alternative primitives which guarantee Pollard's entropy conditions, and rules for building families of functions which inherit them, see Andrews (forthcoming), Pollard (1991), and the literature cited in these references.

Before going on to assumption 3 we note that use of both assumption R, and of the equicontinuity conditions, in our proofs is premised on the assumption that $\Pr\{\tau_n \in \Upsilon\} \to 1$ and $\Pr\{h_n \in \mathcal{H}\} \to 1$ as $n \to \infty$; where, recall, Υ and $\mathcal{H}$ are families of functions with bounded derivatives up to a certain order. For kernel estimates, existence of derivatives is insured by the choice of the kernel. An easy way to insure that these derivatives will be bounded (with probability tending to one) is to insure that the derivatives of the estimated functions converge in probability to their true values (these satisfy the restrictions by assumption). For bandwidths of the form discussed here, a standard argument will insure mean square convergence of the derivatives provided; $0 < \psi < [1/(x+2d)]$, where d is the order of the derivative needed, and x is the number of regressors. Thus, to insure convergence of derivatives of order $x/2+1$ (which is more than enough for the equicontinuity results, see above), it suffices for $[1/2(x+1)] \geq \psi$. Recall that for any given degree of bias reduction (b), there will be a lower bound to the ψ that will allow us to attain the rates in assumption 1. In particular, $\psi > [1/(2b+x)]$ will suffice for the rate conditions, so provided $b > (x/2+1)$, there will be a choice of ψ which satisfies all our conditions simultaneously. For series estimators one can insure the existence of bounded derivatives by constraining the space of acceptable coefficient vectors.

We now turn to assumption 3, and provide references for conditions which suffice for (15a) and (15b) above. Again we focus on the case where $E[y(z_i, \tau_0, \theta_0) | v] = h_0(v_i, \theta_0, \tau_0)$ and $E[y_2(z_i) | v_i] = \tau_0(v_i)$, so that h_n and τ_n are obtained as either a series estimator or a kernel estimator of a regression function. Both Andrews (1991a) and Newey (forthcoming; see also the discussion in Newey,1991a) provide (slightly different) sets of

sufficient conditions for (15a) and (15b) when h_n and τ_n are series estimators. The $f_1(\cdot)$ and $f_2(\cdot)$ functions appearing in (15) are then given by

$$f_1(z_i) = M(v_i,\theta_0)[y_1(z_i,\theta_0)-h_0(v_i,\theta_0,\tau_0)], \text{ and}$$
$$f_2(z_i) = M(v_i,\theta_0)H_{0t}(v_i,\theta_0,\tau_0)[y_2(z_i)-\tau_0(v_i)].$$

Andrews (1992) extends the results in Powell Stock and Stoker (1989) to provide a simple proof that, again provided certain regularity conditions are satisfied, then (18a) and (18b) are also satisfied if h_n [τ_n] are obtained using a bias–reducing kernel estimator[6]. The formula for $f_1(\cdot)$ and $f_2(\cdot)$ are identical to those given above for the series estimator. That is, the limit distribution does not depend on the form of the estimator for the nonparametric component of the moment conditions (a result which accords with proposition 1 of Newey,1991). We note that the regularity conditions used in these articles to do not generate any further restrictions on the way we form our estimators.

Recall that the formula given above for the limit distribution of our semiparametric moment estimator is the natural extension of the pathwise derivative formula provided in Newey(1991) to the case where one or more of the unknown functions depends on another unknown function. This implies that smoothness, independence, and our A1 to A3 are sufficient for using the pathwise derivative formula in cases where one or more of the unknown functions depends on θ and/or on another unknown function. Since our regularity conditions suffice for theorem 5.5 in Newey(1991), it also implies that (one, of several possible) consistent estimator for the variance covariance matrix in (19) can be obtained by substituting estimated (denoted by a $\hat{}$) for actual functions in the definitions of $f_1(\cdot)$, $f_2(\cdot)$, and $m(\cdot)$, and computing

---

[6]An early version of this paper contained an appendix which provides an alternative proof of this result.

$$1/n\ \Sigma_i\{[\hat{m}(z_i,\hat{\theta}_n) + \hat{f}_1(z_i) + \hat{f}_2(z_i)][\hat{m}(z_i,\hat{\theta}_n) + \hat{f}_1(z_i) + \hat{f}_2(z_i)]'\}.$$

III. An Empirical Example.

Olley and Pakes (1991) develop a semiparametric estimator to account for simultanaeity and selection biases in estimates of production function parameters (the former due to the endogeneity of input demands and the latter due to exit behavior). We begin by formalizing their argument for the limit distribution of that estimator, and then consider the effect of alternative estimators for the semiparametric components, and for the standard errors of the parametric components, on the results reported in the Olley/Pakes article.

A brief description of their problem is in order. They assume that the industry produces a homogeneous product with Cobb–Douglas technology, and that the factors underlying the profitability differences among firms are neutral efficiency differences. Therefore the production function is written as

$$(22) \qquad y_{it} = \beta_0 + \beta_a a_{it} + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \eta_{it}$$

where y_{it} is the log of output (value added) from plant i at time t, a_{it} is its age, k_{it} is the log of its capital input, l_{it} is the log of its labor input, ω_{it} is its productivity, and η_{it} is either measurement error in output or a shock to productivity which is independent over time and realized after all input decisions are made. Here both ω and η are unobserved. The distinction between them is that input and exit decisions can depend on ω while those decisions will be independent of realizations of η (formally ω is an unobserved state variable known to the economic agent but not to the econometrician).

Labor is assumed to be the only variable factor (so its choice can be affected by the current value of ω_t). The other two inputs, k_t and a_t, are fixed factors and are only affected by the distribution of ω_t conditional on information at time t–1, and past values of

ω. k_t is assumed to be built up from past investments through the traditional accumulation relationship; $k_t = (1-\delta)k_{t-1} + i_{t-1}$, where i_t is investment at time t and δ is a known depreciation rate. ω_t is assumed to evolve as an exogenous first order Markov process. That is if $\mathcal{A}_{-1}$ contains all information known in period t–1 then for any x, $\Pr\{\omega_t \leq x | \mathcal{A}_{-1}\} = \Pr\{\omega_t \leq x | \omega_{t-1}\}$, and this latter probability is determined by the family of distributions $\mathbb{P}_\omega = \{ P(\cdot|\omega), \omega \in \Omega \}$.

Olley and Pakes (1991) consider a Markov Perfect Nash equilibrium which generates investment and exit policies of the following form. If χ is an indicator function which takes the value one if a firm continues in operation and zero if it exits, then

$$(23a)\quad \chi_t = \begin{cases} 1 & \text{if } \omega_t \geq \underline{\omega}_t(a_t,k_t) \\ 0 & \text{otherwise} \end{cases}; \quad (23b)\quad i_t = \begin{cases} & \text{if } \chi_t=0 \\ i_t(\omega_t,a_t,k_t) & \text{if } \chi_t=1 \end{cases}$$

with $i_t(\cdot)$ strictly increasing in ω whenever $i_t>0$.

The functions $\underline{\omega}_t(\cdot)$ and $i_t(\cdot)$ which set the exit and investment rules are determined as part of the Markov Perfect Nash equilibrium, and their form depends on a host of auxiliary assumptions (on both the nature of the spot market equilibrium for current output, and on functional forms). Moreover, even given the assumptions needed to determine the investment and exit rules as a function of a small number of parameters, the forms of these policies will be very difficult to compute; so difficult that attempting to compute them iteratively at each different function evaluation needed for a nonlinear search routine would probably be impractical even on the most sophisticated of computing equipment.

We now turn to a semiparametric technique which treats the investment policy $[i_t(\cdot)]$, the exit policy $[\chi_t]$, and the Markov process $[\mathbb{P}_\omega]$, as nuisance functions to be estimated nonparametrically. Note first that the fact that the investment function is strictly increasing in ω whenever $i_t>0$ lets us invert (23b) and write

$$(24) \qquad \omega_t = q_t(i_t,a_t,k_t) \ .$$

Substituting (24) into the production function (22) gives us

$$(25) \qquad y_{it} = \beta_l l_{it} + \phi_t(i_{it},a_{it},k_{it}) + \eta_{it}$$

where,

$$\phi_t(i_{it},a_{it},k_{it}) = \beta_0 + \beta_a a_{it} + \beta_k k_{it} + q_t(i_{it},a_{it},k_{it}).$$

(25) is an example of a partially linear model and it can be used to estimate β_l. Note, however, that the production function coefficients of capital and age, β_a and β_k, can not be identified from this equation since the equation does not allow us to separate out the effect of capital and age on the investment decision, from there effect on output.

To identify the age and capital coefficients we have to use the panel structure of the data and the model's implications regarding the relationship between the productivities of a given firm over time. Moreover since we only observe the subsequent years' data for those plants that survive, we will need the probability of survival. That probability is given by

$$\begin{aligned} (26) \qquad & \Pr\{\chi_t = 1 \mid \underline{\omega}_t(k_t,a_t),\ \mathcal{A}_{-1}\} \\ & = \Pr\{\omega_t \geq \underline{\omega}_t(k_t,a_t) \mid \underline{\omega}_t(k_t,a_t),\ \omega_{t-1}\} \\ & = P_{t-1}\{\underline{\omega}_t(k_t,a_t),\omega_{t-1}\} = P_{t-1}(k_{t-1},a_{t-1},i_{t-1}) \equiv \mathcal{A}_{-1}, \end{aligned}$$

where the first equality in the third line follows from the fact that (k_t,a_t,ω_{t-1}) can be written as a function of $(k_{t-1},a_{t-1},i_{t-1})$, we have dropped the dependence of the variables on the individual subscript (i) for notational convenience, and $\mathcal{A}_{-1}$ represents all information known at time t–1.

We complete the system to be estimated by considering the expectation of $y_t - \beta_l l_t$

conditional on information at t−1 and survival. This equation, when combined with the estimates of β_l, ϕ_{t-1}, and $\mathcal{R}_{-1}$, from (25) and (26) will allow us to identify β_a and β_k . We have

$$(27)\quad E[y_t - \beta_l l_t \mid \mathcal{A}_{-1}, \chi_t = 1] = \beta_0 + \beta_a a_t + \beta_k k_t + E[\omega_t \mid \omega_{t-1}, \chi_t = 1]$$
$$= \beta_a a_t + \beta_k k_t + g(\mathcal{R}_{-1}, \omega_{t-1})$$

where

$$g(\mathcal{R}_{-1}, \omega_{t-1}) \equiv \beta_0 + \int_{\underline{\omega}_t} \{\omega_t P(d\omega_t \mid \omega_{t-1}\} / \int_{\underline{\omega}_t} P(d\omega_t \mid \omega_{t-1}\}.$$

and the last equality assumes that the function giving the probability of survival, $\mathcal{R}_{-1} = P\{\underline{\omega}_t(k_t, a_t) \mid \omega_{t-1}\}$, is invertible for almost every ω, allowing us to write $\underline{\omega}_t(\cdot)$ as a function of $\mathcal{R}_{-1}$ and ω_{t-1}.

We now consider alternative estimators of the system in (25), (26), and (27). The data consist of a thirteen year panel of enterprises in the telecommunications equipment industry. For an overview of both the events that took place in this industry during the period of the study, and of the data set used, see Olley and Pakes (1991), and the literature cited there.

Equation (25) is an example of a partially linear model. The limit properties of its estimator of β_l have been analyzed using both kernel (see,Robinson, 1988) and series (Andrews 1991 and Newey forthcoming) estimators of the nonparametric component, $\phi_t(\cdot)$. For simplicity we use a polynomial series estimator in all of what follows. That is we project y_t on l_t and a polynomial in the triple (a_t, k_t, i_t). All empirical results described below that are based on series approximations use a fourth order polynomial (with a full set of interactions) as the approximating function. In no case was there any noticeable change in either the estimates of the coefficients of interest, or in the minimand, when we went from a third to a fourth order approximation.

Next we consider the estimation of the selection equation in (26); the equation giving the probability of survival as a function of (i_t,k_t,a_t). Here we use three different estimators of the survival probability and compare results; a series estimator, an estimator based on a normal kernel, and an estimator based on a bias–reducing kernel. The series approximation was constructed by using a polynomial series in (i_t,k_t,a_t) as regressors in a probit estimation algorithm (that this is just a series approximation follows from the fact that the formula the computer uses to compute the normal distribution is a series approximation to that distribution). Whenever we employ a normal kernel we use a diagonal covariance matrix with the inverse of the variance of the regressors as the diagonal elements, and a bandwidth of one. Whenever we employ a bias–reducing kernel we use the family of normal based bias–reducing kernels discussed in Bierens (1987) with; a diagonal $\hat{\Omega}$ with the inverse of the variance of the regressors as the diagonal elements, a bandwidth chosen by cross validation, and a degree of bias reduction equal to four.

The third step of the estimation procedure takes the estimates of β_1, q_{t-1} and $\mathcal{R}_{-1}$ from the first two steps, substitutes them into equation (27) for the true β_1, q_{t-1}, and $\mathcal{R}_{-1}$, and then obtains estimates of β_a, β_k, and the $g(\cdot)$ function by minimizing the sum of squared residuals in the resulting equation. As above we tried three estimators for the unknown $g(\mathcal{R}_{-1}, q_{t-1})$ function; a series estimator, a normal kernel, and a bias–reducing kernel estimator. Recall that it is ϕ_{t-1} that is estimated in the first stage, and $q_{t-1} = \phi_{t-1}-\beta_a a_{t-1}-\beta_k k_{t-1}$, so that the values of the regressors in the nonparametric function in the third stage depend upon the values of the parameters of interest. Consequently the diagonal elements of the covariance matrix of the kernel, and the cross validation procedure used to choose the bandwidth for the bias–reducing kernel, could be recomputed for every alternative value of the parameter vector evaluated in the minimization subroutine. This proved to be too computer time intensive. Instead we used the series estimates of the coefficients to both set the diagonal elements, and to cross validate for the bandwidth, and then held both the covariance matrix and the bandwidth fixed throughout

the remainder of the estimation algorithm.

If, for expositional simplicity, we temporarily ignore trimming, then the population analogue of the sample moment conditions that the nonlinear least squares estimation procedures sets equal to zero are

$$(28a)\quad E\Big[\chi_t\{y_t-l_t\beta_l-k_t\beta_k-a_t\beta_a-g[\mathcal{R}_{t-1},\phi_{t-1}-k_{t-1}\beta_k-a_{t-1}\beta_a]\} \times \{k_t-[\partial g(\cdot)/\partial\phi]k_{t-1},a_t-[\partial g(\cdot)/\partial\phi]a_{t-1}\}'\Big].$$

Rewriting this in the notation of the first two sections of the paper with $z'=(y_t,l_t,\chi_t,k_t,a_t,k_{t-1},a_{t-1},i_{t-1})$, $v_{1t}' = v_{2t}' = v'=(k_{t-1},a_{t-1},i_{t-1})$, and $\theta' = (\beta_k,\beta_a)$, we have

$$(28b)\quad G(\theta) = Em\{z,h_{10},h_{20}[v,\tau_{10}(v),\tau_{20}(v),\theta],\partial h_{20}[\cdot]/\partial\tau_{20},\theta],$$

where

$$\tau_{10}(v) = \mathcal{R}_{t-1}(k_{t-1},a_{t-1},i_{t-1}),$$
$$\tau_{20}(v) = \phi_{t-1}(k_{t-1},a_{t-1},i_{t-1}),$$
$$h_{10}=\beta_l$$
$$h_{20}[v,\tau_{10}(v),\tau_{20}(v),\theta] = g[\mathcal{R}_{t-1}(\cdot),\phi_{t-1}(\cdot)-k_{t-1}\beta_k-a_{t-1}\beta_a],$$

and it is understood that both $\tau_{10}(\cdot)$ and $\tau_{20}(\cdot)$ are indexed by time. Note that (28) implies that $G(\theta_0)=0$.

Our estimation procedure substitutes preliminary estimators of $\tau_1(\cdot)$, $\tau_2(\cdot)$, h_1 and $h_2(\cdot)$ discussed above into $m(\cdot)$ and then chooses θ to minimize

$$(29)\quad \|G_n(\theta)\| = \|(1/n)\Sigma_i m\{z_i,h_{1n},h_{2n}[v_i,\tau_{1n}(v_i),\tau_{2n}(v_i),\theta],\partial h_{2n}[\cdot]/\partial\tau_2,\theta\}\|.$$

If we define $h_{30}(\cdot)\equiv\partial h_{20}(\cdot)/\partial\tau_2$, and h_{3n} accordingly, then (29) is in the form of equation (2) with $\mathcal{H}$ mapping into a subset of $\mathbb{R}^3$, and Υ into a subset of $\mathbb{R}^2$. We now

provide an expression for V, the variance covariance of $\sqrt{n}\,\mathcal{G}_n(\theta_0)$, for the example in (29). An estimate of it will then be combined with an estimate of

$$(30) \qquad D = \partial G(\theta)/\partial\theta' \,|\, \theta=\theta_0 \equiv \partial\, E\, m(z,\theta_o)/\partial\theta \,|\, \theta=\theta_0 ,$$

to obtain an expression for the variance covariance of θ appearing in our limit theorum (theorum 14).

Let $M_j(\cdot)$ be the (row) vector of derivatives of $m(\cdot)$ with respect to the value of h_j for j=1,2,3; and $H_{js}(\cdot)$ be the derivative of $h_j(\cdot)$ with respect to $\tau_s(\cdot)$ for s,j = 1,2. Now use equation (27) to show that $E[M_3(z,\theta)\,|\,v] = 0$. It follows that we can treat h_3 as if it were a known function in computing the variance of $\sqrt{n}\,\mathcal{G}_n(\theta_0)$ [corollary 20]. As a result

$$(31) \qquad \mathcal{G}_n(\theta_0) = (1/n)\Sigma_i\{\, m(z_i,\theta_0) + M_1(v_i,\theta_0)[\beta_{1n} - \beta_1] +$$
$$M_2(v_i,\theta_0)[h_{2n}(v_i,\theta_0) - h_{20}(v_i,\theta_0)] + M_2(z_i,\theta_0)H_{21}(v_i,\theta_0)[\tau_{1n}(v_i)-\tau_{10}(v_i)]$$
$$+ M_2(z_i,\theta_0)H_{22}(v_i,\theta_0)[\tau_{2n}(v_i)-\tau_{20}(v_i)]\,\} + o_p(1/\sqrt{n}).$$

Moreover, since $h_2(\cdot)$, $\tau_1(\cdot)$, and $\tau_2(\cdot)$ are all conditional expectations, we have

$$(32a) \quad (1/\sqrt{n})\Sigma_i M_2(v_i,\theta_0)[h_{2n}(v_i,\theta_0)-h_{20}(v_i\ \theta_0)] = (1/\sqrt{n})\Sigma_i f_2(z_i) + o_p(1/\sqrt{n}),$$
$$(1/\sqrt{n})\Sigma_i M_2(v_i,\theta_0)H_{21}(v_i,\theta_0)[\tau_{1n}(v_i)-\tau_{10}(v_i)] = (1/\sqrt{n})\Sigma_i f_3(z_i) + o_p(1/\sqrt{n})$$

and

$$(1/\sqrt{n})\Sigma_i M_2(v_i,\theta_0)H_{22}(v_i,\theta_0)[\tau_{2n}(v_i)-\tau_{20}(v_i)] = (1/\sqrt{n})\Sigma_i f_4(z_i) + o_p(1/\sqrt{n})$$

where if

$$Q(v_i,\theta_0) = \chi_t\{k_t-[\partial g(\cdot)/\partial\phi]k_{t-1},a_t-[\partial g(\cdot)/\partial\phi]a_{t-1}\}'$$

then

$$f_2(z) = Q(v_i,\theta_0) \times \{y_t-l_t\beta_1-k_t\beta_k-a_t\beta_a-g[\mathcal{P},\phi_{t-1}-k_{t-1}\beta_k-a_{t-1}\beta_a]\}\ ,$$
$$f_3(z) = Q(v_i,\theta_0) \times [\partial g(\cdot)/\partial\mathcal{P}][\chi_t-\ \mathcal{R}_t]$$

and,

$$f_4(z) = Q(v_i,\theta_0) \times [\partial g(\cdot)/\partial \phi]\ \eta_{t-1},$$

where, again, $\eta_{t-1}=[y_{t-1}-\beta_1 l_{t-1}-\phi_{t-1}]$. Finally we note that the moment condition which defines the estimate of β_1 (Robinson,1986) is

$$E\ \{\ \beta_1\ [l-E(l|i,a,k)]^2 - [y-E(y|i,a,k)][l-E(l|i,a,k)]\ \}.$$

Substituting first stage nonparametric estimators of $E(l|i,a,k)]$ and $E(y|i,a,k)$ into the sample analogue of this equation and then analyzing the resulting estimator of β_1 as in the last section we find that

$$\sqrt{n}(\beta_{1n}-\beta_0) = \{E[l-E(l|i,a,k)]^2\}^{-1}\ (1/\sqrt{n})\ \Sigma\ \eta[l-E(l|i,a,k)]$$

so that

$$(32b) \qquad (1/\sqrt{n})\ \Sigma_i M_i(v_i,\theta_0)[\beta_{1n}-\beta_1] = (1/\sqrt{n})\Sigma_i f_1(z_i) + o_p(1)$$

where

$$f_1(z_i) = \{E[l-E(l|i,a,k)]^2\}^{-1}\overline{Q}(\theta_0)\eta[l-E(l|i,a,k)], \text{ and } \overline{Q}(\theta_0)= EQ(v,\theta_0).$$

Equation (32), our regularity conditions, and the Lindberg Levy central limit theorum imply that

$$(33) \qquad \sqrt{n}\ \mathcal{G}_n(\theta_0) \rightarrow_d \mathcal{N}(0,V),$$

with

$$V=E[m(z_i,\theta_0)+\Sigma_{j=1}^{4} f_j(z_i)][m(z_i,\theta_0)+\Sigma_{j=1}^{4} f_j(z_i)]'.$$

Equations (30) and (33) provide the analytic formula needed for the variance covariance matrix in the central limit theorum in (14). We obtain our estimate of that

variance covariance matrix by treating the estimated as the actual values of unknown parameters, and forming the sample analogues of the expectations needed for D and V. The estimated variance covariance for each different estimator uses that estimator's estimates of the nonparametric components of the problem.

Table 1 provides alternative estimates of the capital, age, and time coefficients, and their standard errors. The different columns correspond to estimation algorithms which use different estimators for the nonparmametric components of the problem. As noted above all columns use the same series estimator for the partially linear model in (24). The first row of the table specifies the form of the nonparametric estimate of the survival probability used in obtaining the results in that column, while the second row specifies the form of the estimate of the $g(\mathcal{P},h)$ function in (27).[7] The table lists three estimates of the standard error of each estimated coefficient. The first (labelled NCE for not corrected estimate) is the estimate one would obtain if one did not correct for the fact that β_1, $\mathcal{P}$, h, and g are estimated objects (it sets $f_j(z)=0$, for j=1,2,3,4 and all i). The second provides the standard errors estimated from the analytic formula given above (it is labelled PDE for pathwise derivative estimate).

The final estimate of the standard error is obtained from a bootstrap procedure (and is labelled BSE). The Census data set is a "rolling" panel. Every five years a new probability sample is drawn from the latest Census of Manufacturing and those plants are followed either over the next five years, or until they exit. The sampling weights depend on the size of the establishment, and in some years, on the size of the firm. The base sample is augmented by a sample of new entrants (taken from the standard establishment list or SSEL) every year. The sampling procedure used for the bootstraps mimicked the sampling procedure used by the Census. We were satsified that we had reproduced their

---

[7]We also combined series and standard kernel estimates for $\mathcal{P}$ with bias reducing kernel estimates of $g(\cdot)$. Since the estimates obtained in this fashion did not differ much from those reported in the table we do not discuss them.

procedure with sufficient accuracy when we began generating samples that had characteristics that were quite close to the characteristics of the original sample (eg. the number of plant–year observations generated, the number of plants active in different periods,...). We note that the samples used for the bootstraps were drawn independently for each different estimation procedure. The number of bootstrapped coefficient estimates used to compute the standard errors is given in the last row of the table.

Two points should be noted before proceeding. First, it is only the third column whose estimates are known to abide by all the regularity conditions needed for our limit theorems. On the other hand, there is a strong presumption that the series estimators do also, and if prior monte carlo work is to be taken as a guide, any bias caused by using the normal (instead of the bias–reducing) kernel should be small (and may be offset by smaller variances in finite samples; see Powell, Stock, and Stoker, 1989). Second, we ran several of the estimators with and without trimming. There was only one case in all of our experiments for which the trimming had any substantial impact on the results, and that impact was not on the coefficient values, but rather on the PDE of their standard error. Since the case in question was an early run not reported here, we ignore trimming in what follows.

The first stage estimate of the labor coefficient was .615, with a PDE of .027, and a BSE between .028 and .031 in the three bootstrapped samples. As to the other coefficients, the Table seems to make several points. Perhaps most important, the PDE's of all coefficients, with the possible exception of the time coefficient, seem to be within tolerable bounds for standard errors of those coefficients. We should note that the focus of interest was on the capital coefficient (as there was a theoretical reason to expect it to be underestimated in procedures which do not account for selection), and partly as a result both the investment and the selection equations were allowed to vary freely across time periods. Hence we did not expect a precisely estimated time coefficient. Also, though this is a large sample (it is an unbalanced panel with 1763 plant/time observations), it is not

unusually large by modern econometric standards, and the estimation problem itself is quite complex (it requires us to estimate three nonparametric functions along with the coefficients of interest). At least in this instance, then, semiparametric techniques seem to be quite helpful in both; ameliorating the need for auxiliary assumptions, and in simplifying the computational burden of the estimation algorithm.

Second, the coefficients and the PDE estimates of their standard errors do not seem to differ "too much" between estimation techniques. By "too much" here we simply mean differences of a magnitude which are likely to have an impact on the empirical implications of the parameter estimates (see the discussion in Olley and Pakes,1991). So any biases induced by either the series or by the standard kernel estimation procedures do not seem to generate large differences in results for our example. Note that this is so even though the fit from the equation used to estimate these coefficients does seem to differ somewhat as we vary the estimators for the nonparametric components of the problem (this was a bit surprising, and we do not really have a good idea of why the kernels, particularly the bias–reducing kernel, fit better in our application).

It is important to realize that in problems such as ours where a nonparametric function depends on the parameter of interest, when we use kernels we must recalculate the kernel every time we evaluate a different parameter vector in the minimization subroutine. As a result the kernel estimators can require substantially more computer time than the series estimators. In our example the series estimator took under one hour to run on our 486, while the bias–reducing kernel estimator generally took over twelve hours. Thus even if one did not want to stop with the series estimators, one might consider using them as starting values for the more comptationally intensive kernel estimators.

We now move on to consider the estimates of the standard errors in more detail. The first point that comes out clearly here is that it does seem important to correct the standard errors for the fact that we are using estimated, instead of true, values of (β_1, $\mathcal{P}$, q, g). The NCE estimates of the standard errors are all lower than the PDE's, and the

difference between them is often, though not always, quite large.

Second the BSE's are uniformly larger than the PDE's. A couple of caveats should be noted before attributing too much to these results. First, we should keep in mind that though we are producing three different vectors of estimators, there is only one underlying data set, so the difference between the two estimates of the standard errors are correlated across the columns of the table. Second, we note that to obtain the actual estimates in table 1 we ran our estimation procedure *several* times using different starting values and different minimization subroutines in the different runs, and then chose the estimates that minimized *over* the runs. We simply did not have the computational resources that would be needed to do this for the bootstraps in a reasonable time period, so the possibility that the minimization routine used in the bootstrap procedure periodically picked out a local minima which was not global cannot be ruled out (a possibility which we would expect to increase variance). However, we did not seem to have any trouble finding the global minima for the estimator in the series column of the table, and here also the BSE was about fifty per cent larger than the PDE. Also, the fact that the BSE's are larger than the PDE's is reminiscent of the results in Stoker (1991) on smoothing bias in density and derivative estimators. That is, the derivative estimator used in our computations of the variance may just have a "smoothing bias" which causes a systematic tendency to produce smaller standard errors than the true standard errors when the bandwidth is held at a positive constant. As a result a deeper investigation of the relationship between the PDE's and the BSE's, though beyond the scope of this paper, may well be worthwhile.

Table 2 provides a comparison of the distribution of the bootstrap estimates of the capital coefficient from alternative estimators. Of particular interest here is a comparison of the distribution of the bootstrap estimates from the bias–reducing to those from the standard normal, kernels. Their means, standard errors, and skewness are all very close, but the kurtosis in the two distribution is very different. In particular the distribution of the estimates from the standard kernel looks very much like a normal distribution, but the

distribution from the bias–reducing kernel has much fatter tails (indeed all the difference in the two distributions is outside of the interquartile range). A risk averse researcher might, then, want to stick with the standard normal kernel.

References

Andrews, D.W.K. (1991a); "Asymptotic Normality of Series Estimators for Various Nonparametric and Semiparametric Models", Econometrica, 59, 307–45.

Andrews, D.W.K. (1992); "Asymptotics for Kernel–based Non–orthogonal Semiparametric Estimators", unpublished manuscript, Cowles Foundation, Yale, University.

Andrews, D.W.K. (1993); "Nonparametric Kernel Estimation for Semiparametric Models", Revision of Cowles Foundation Discussion Paper No.909R, Yale University.

Andrews, D.W.K. (1994); "Asymptotics for Semiparametric Econometric Models Via Stochastic Equiconintuity", forthcoming Econometrica.

Andrews, D.W.K. (forthcoming); "Empirical Process Methods in Econometrics", the Handbook of Econometrics, vol.iv, ed. R.Engle and D.McFadden.

Bierens, H.J. (1987); "Kernel Estimators of Regression Functions", in Advances in Econometrics, Vol.I., T.F.Bewley (ed), Cambridge University Press.

Goldstein L., and K. Messer (1990); "Optimal Plug–In Estimators for Nonparametric Functional Estimation", Technical Report No. 277, Department of Statistics, Stanford University.

Griliches Z. (1967): "Production Functions in Manufacturing: Some Preliminary Results," in The Theory of Empirical Analysis of Production, Murray Brown, ed., Columbia University Press for the N.B.E.R.

Hansen L. (1982): "Large Sample Properties of Method of Moments Estimators", Econometrica, 50, 1029–54.

Hardle W. (1991); Applied Nonparmetric Regression, Econometric Society Monograph Series, Cambridge University Press.

Hardle W., and T. Stoker (1989); "Investigation of Smooth Multiple Regression by the Method of Average Derivatives", Journal of the American Statistical Association, 84, 986–995.

Hausman J., and W. Newey (1991); "Nonparametric Estimates of Exact Consumer Surplus and Deadweight Loss", mimeo, M.I.T. Department of Economics.

Hotz J., and R. Miller (1991); "Conditional Choice Probabilities and the Estimation of Dynamic Models", mimeo,GSIA, Carnegie Mellon University.

Jones M.,Linton O.,and Nielsen J.(1993);"A Simple and Effective Bias Reduction Method for Density and Regression Estimation", mimeo, Yale University.

Manski, C.F.(1991); Analogue Estimation Methods in Econometrics.

Newey W.K.(1993); "Convergence Rates for Series Estimators," forthcoming in Statistical Methods of Economics and Quanitative Economics: Essays in Honor of C. R. Rao, G. S. Maddala and P. C. B. Phillips, eds.

Newey W.K.(1991);"The Asymptotic Variance of Semiparametric Estimators", Revision of MIT Department of Economics Working Paper No. 583.

Newey W.K.(forthcoming,b); "Kernel Estimation of Partial Means and a General Variance Estimator", Econometric Theory .

Newey W.,Hsieh F.,and Robins J.(1993); "Bias Corrected Semi–Parametric Estimation", mimeo, Department of Economics, M.I.T.

Olley S., and A. Pakes (1991); "The Dynamics of Productivity in the Telecommunications Equipment Industry", preprint, Department of Economics, Yale University.

Pakes A. (forthcoming); "Dynamic Structural Models, Problems and Prospects; Mixed Continuous Discrete Controls and Market Interactions", in Advances in Econometrics, proceedings of the sixth World Congress of the Econometric Society, edited by C. Sims.

Pakes A., and D. Pollard (1989); "Simulation and the Asymptotics of Optimization Estimators", Econometrica, 57, pp 1027–57.

Powell J.L., J.H.Stock, and T.M.Stoker (1989); "Semiparametric Estimation of Index Coefficients", Econometrica, 57, pp 1403–30.

Pollard D. (1991); Empirical Processes, Theory and Applications, CBMS/NSF Regional Conference Series, Volume 2, Institute of Mathematical Statistics.

Rao C.R.(1973); Linear Statistical Inference and its Applications, John Wiley and Sons.

Robinson P.(1988); "Root–N–Consistent Semiparametric Regression", Econometrica, Vol.56, pp 931–54.

Silverman B.W. (1986); Density Estimation, Chapman and Hall Monographs on Statistics and Applied Probability.

Stoker T.(1991a); "Smoothing Bias in Density Derivative Estimation", mimeo, Sloan School of Management, M.I.T.

Stoker T. (1991b); Lectures on Semiparametric Econometrics, Core Lecture Series, Core Foundation, Univerisite Catholique de Louvain.

Table 1: Alternative Estimates of Production Function Coefficients and Their Standard Errors[1]

	1	2	3
Selection Equation	Probit/Series	Standard Kernel	Bias Reducing Kernel
Semiparametric[2] Moment Condition	Series	Standard Kernel	Bias Reducing Kernel
Capital Coefficient	.33	.31	.35
NCE[3]	.032	.014	.004
PDE	.035	.036	.045
BSE	.052	.092	.097
Age Coefficient	-.001	-.009	.01
NCE	.005	.003	.002
PDE	.006	.011	.014
BSE	.014	.021	.017
Time Coefficient	.012	.038	.04
NCE	.026	.016	.004
PDE	.027	.037	.038
BSE	.030	.042	.046
SSR (Semiparametric Moment Condition)	582.8	572.9	553.7
No. of Bootstraps	44	60	60

[1] The first stage estimator of the labor coefficient was .615 with a PDN estimated standard error of .027 and a Bootstrap estimated standard error of .031.

[2] All series estimators use a fourth order polynomial with full interactions among regressors. All standard kernel estimators use normal kernels with a diagonal convariance matrix equal to the inverse of the variance of the regressors and a bandwidth equal to one. All bias reducing kernel estimators use the normal based bias reducing kernels in Bierens (1987) with $\hat{\Omega}$ equal to a diagonal matrix with the inverse of the variance of the regressors on the diagonal, a degree of bias reduction equal to four, and a bandwidth chosen by cross-validation.

[3] NCE = no correction for the nonparametric estimators
PDE = analytic formula for estimators of standard errors given in the text
BSE = Bootstrap estimate of standard errors

Table 2

Characteristics of the Distribution of Capital Coefficients Obtained from the Boostrap

	Bias Reducing Kernel	Standard Kernel	Series
Summary Statistics			
Mean	.36	.36	.29
Standard Deviation	.097	.092	.056
Skew	.27	-.26	-.55
Kurtosis	2.63	.04	.52
Quantiles			
.99	.69	.54	.43
.95	.47	.51	.36
.90	.44	.49	.36
.75	.41	.42	.32
.50	.37	.37	.30
.25	.30	.30	.26
.10	.25	.23	.22
.05	.18	.19	.17
.01	.11	.13	.14