working paper
department
of economics

The Asymptotic Variance of Semiparametric Estimators

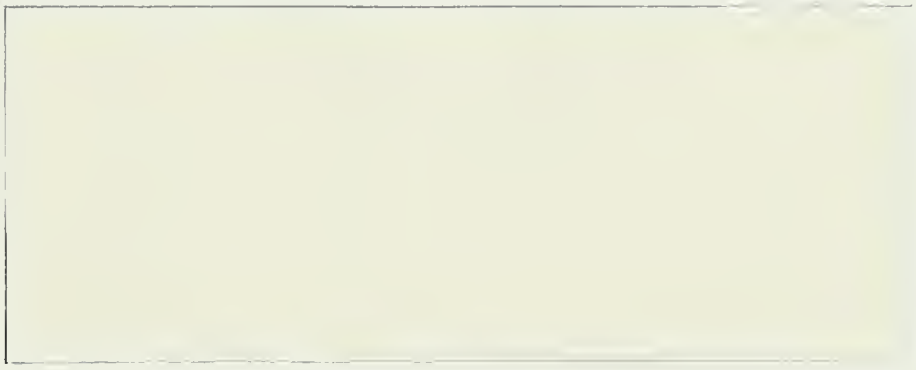
Whitney K. Newey

No. 583

Rev. July 1991

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139



The Asymptotic Variance of Semiparametric Estimators

Whitney K. Newey

No. 583 Rev. July 1991

MIT Working Paper 583

The Asymptotic Variance of Semiparametric Estimators

by

Whitney K. Newey

MIT Department of Economics

August 1989

Revised, July 1991

Helpful comments were provided by referees and M. Arellano, L. Hansen, J. Hausman, R. Klein, P.C.B. Phillips, J. Powell, J. Robins, T. Stoker. Financial support was provided by the NSF, the Sloan Foundation, and BellCore.

Abstract

Knowledge of the asymptotic variance of an estimator is important for large sample inference, efficiency, and as a guide to the specification of regularity conditions. The purpose of this paper is the presentation of a general formula for the asymptotic variance of a semiparametric estimator. A particularly important feature of this formula is a way of accounting for the presence of nonparametric estimates of nuisance functions. The general form of an adjustment factor for nonparametric estimates is derived and analyzed.

The usefulness of the formula is illustrated by deriving propositions on asymptotic equivalence for different nonparametric estimators of the same function, conditions for estimation of the nuisance functions to have no effect on the asymptotic variance, and the form of a correction term for the presence of linear function of a conditional expectation estimator, or other projection estimator (e.g. partially linear and/or additive nonparametric projections), and for a function of a density. Specific results cover a semiparametric random effects model for binary panel data, nonparametric consumer surplus, nonparametric prediction, and average derivatives. Regularity conditions are given for many of the propositions. These include primitive conditions for $\sqrt{n}$-consistency, asymptotic normality, and consistency of an asymptotic variance estimator with series estimators of conditional expectations (or projections), in each of the examples.

Keywords: semiparametric estimation, asymptotic variance, nonparametric regression, series estimation, panel data, consumer surplus, average derivative.

1. Introduction

This paper develops a general form for the asymptotic variance of semiparametric estimators. Despite the complicated nature of such estimators, which can depend on estimators of functions, the formula is straightforward to derive in many cases, requiring only some calculus. Although the formula is not based on primitive conditions, it should be useful for semiparametric models, just as analogous formulae are for parametric models, such as Huber (1967) for m-estimators. It gives the form of remainder terms, which facilitates specification of primitive conditions. It also can be used to make asymptotic efficiency comparisons, in order to find an efficient estimator in some class.

The usefulness of this formula is illustrated in several ways. New examples are considered throughout, in order to emphasize that it can be a useful tool for further work in semiparametric estimation, and not just a way of "unifying" existing results. A number of Propositions are derived, and primitive conditions are given for many of them. The propositions include showing that the method of estimating a function (e.g. kernel or polynomial regression) does not affect the asymptotic variance of the estimator. Also, two sufficient conditions are given for the absence of an effect on the asymptotic variance from the presence of a function estimator. One is that the limit of the function estimator maximizes the same expected objective function as the population parameter, i.e. the function has been "concentrated out." The other is a certain orthogonality condition.

Several propositions are given on the form of correction terms for the presence of function estimates. One has sufficient conditions for this adjustment to take the form of the projection on the tangent set (the mean-square closure of all scores for parametric models of the nuisance

functions) for a semiparametric model. More specific results are given for the case of conditional expectations, or other mean square projections, and for densities. A characterization of the correction term for estimators of linear functions of projections and densities is given, with specific formula given for semiparametric individual effects regression for binary panel data, nonparametric consumer surplus, and Stock's (1989) nonparametric prediction estimator.

Regularity conditions for $\sqrt{n}$-consistency and asymptotic normality are formulated. The discussion is organized around a few "high-level" assumptions. Times series are covered, including weighted autocovariance estimation of the asymptotic variance, with data-based lag choice. Primitive conditions are given for power series estimators of conditional expectations and other projections, including several examples.

The formula builds on previous work, including that on Von Mises (1947) estimators, i.e. functionals of the empirical distribution, by Reeds (1976), Boos and Serfling (1980), and Fernholz (1983). The formula here allows for explicit dependence on nonparametric functions estimators, such as conditional expectations or densities, which are difficult to allow for in the Gateaux derivative formula for Von-Mises estimators. It is based on calculating the semiparametric efficiency bound, as in Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), and Van der Vaart (1991) for the functional the estimator is a nonparametric estimator of, as discussed in the next section. Also, some of the examples build on previous work on semiparametric estimation, including Bickel, Klaassen, Ritov, and Wellner (1990), Hardle and Stoker (1989), Klein and Spady (1987), Powell, Stock, and Stoker (1989), Robinson (1988), Stock (1989), and others cited below.

Section 2 gives the formula for the asymptotic variance. Section 3 and 4 apply this formula to derive some propositions on the effect of preliminary

nonparametric estimators on the asymptotic variance. Some high-level regularity conditions are collected in Section 5. Section 6 gives general regularity conditions for $\sqrt{n}$-consistency and asymptotic normality when an estimator depends on a series estimator of a conditional expectation or other projection, and Section 7 applies these results to specify primitive regularity conditions for several examples.

2. The Pathwise Derivative Formula for the Asymptotic Variance

The formula is based on the observation that $\sqrt{n}$-consistent nonparametric estimators are often efficient. For example, the sample mean is known to be an efficient estimator of the population mean in a nonparametric model where no restrictions, other than regularity conditions (e.g. existence of the second moment) are placed on the distribution of the data. The idea here is to use this observation to calculate the asymptotic variance of a semiparametric estimator, by finding the *functional* that it nonparametrically estimates, i.e. the object that it converges to under general misspecification, and calculating the semiparametric variance bound for this functional.

To be more precise, let $\hat{\beta}$ be an estimator, and suppose that one can associate with it the triple,

$$(2.1)\quad \hat{\beta} \longrightarrow \begin{cases} z\ ; & \text{finite dimensional data vector,} \\ \mathcal{F} = \{F_z\}; & \text{unrestricted family of distributions of } z, \\ \mu : \mathcal{F} \longrightarrow \mathbb{R}^q; & \mu(F_z) = \operatorname{plim}(\hat{\beta}) \text{ when } F_z \text{ is true.} \end{cases}$$

That is, $\hat{\beta}$ is a nonparametric estimator of $\mu(F_z)$, having this as its probability limit for all distributions of z belonging to a family that is unrestricted, except for regularity conditions. In other words, $\mu(F_z)$ is the

object estimated by $\hat{\beta}$ under general misspecification, when the distribution of z does not necessarily satisfy restrictions on which $\hat{\beta}$ is based. The asymptotic variance formula discussed here is taken to be the variance bound bound for estimation of $\mu(F_z)$, $F_z \in \mathcal{F}$. This formula is an alternative to the Gateaux derivative for Von-Mises estimators, because the domain of $\mu(F_z)$ need not include all distributions, e.g. so that $\mu(F_z)$ can depend explicitly on a density function. In the technical conditions to follow, this feature of the formula results from F_z having a density with respect to a measure for which the true distribution also has a density.

The formula for calculating the variance bound for $\mu(F_z)$ is that given in previous work by Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), and others. Following Van der Vaart (1991), let $\{F_{z\theta} : \theta \in (0, \varepsilon) \subset \mathbb{R}, \ \varepsilon > 0, \ F_{z\theta} \in \mathcal{F}\}$, denote a one-dimensional subfamily of $\mathcal{F}$, i.e. a path in $\mathcal{F}$, such that the true distribution and each member of this subfamily are absolutely continuous with respect to the same σ-finite measure. Let $E[\cdot]$ be the expectation under the true distribution F_{z0}, and let $dF_{z\theta}$ and dF_{z0} be the densities with respect to the common dominating measure, and dz integration with respect to that measure. Let $\mathcal{P}$ denote a set of paths such that for each one there is a random variable $S_\theta(z)$ with $E[S_\theta(z)^2] < \infty$ and

$$\int [\theta^{-1}(dF_{z\theta}^{1/2} - dF_{z0}^{1/2}) - \tfrac{1}{2} S_\theta(z) dF_{z0}^{1/2}]^2 dz \longrightarrow 0, \quad \text{as} \quad \theta \longrightarrow 0,$$

Here $S_\theta(z)$ is a "mean-square version" of the score $\partial \ln(dF_{z\theta})/\partial\theta|_{\theta=0}$ associated with the path, which quantifies a direction of departure from the truth allowed by $\mathcal{F}$. The requirement that $\mathcal{F}$ be unrestricted is formalized in the condition that there is a set of paths $\mathcal{P}$ with associated set of scores $\mathcal{S}$ satisfying the following property:

Assumption 2.1: $\mathcal{S}$ is linear and for any $s(z)$ with $E[s(z)] = 0$, $E[s(z)^2] < \infty$, and any $\varepsilon > 0$ there is $S_\theta(z) \in \mathcal{S}$ such that $E[\{s(z) - S_\theta(z)\}^2] < \varepsilon$.

That is, the mean-square closure of the set of scores is all mean-zero random variables, i.e. $\mathcal{T}$ allows for any direction of departure from the truth.

The functional $\mu(F_z)$ is *pathwise differentiable* if there is a mapping $\mu_F(S_\theta) : \mathcal{S} \longrightarrow \mathbb{R}^q$ that is linear and mean-square continuous with respect to the true distribution (i.e. for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\|\mu_F(S_\theta)\| < \varepsilon$ if $E[S_\theta(z)^2] < \delta$), such that for each path,

$$(2.2) \qquad \theta^{-1}[\mu(F_{z\theta}) - \mu(F_{z0})] \longrightarrow \mu_F(S_\theta) \quad \text{as} \quad \theta \longrightarrow 0,$$

i.e. the derivative from the right of $\mu(F_{z\theta})$ at the truth ($\theta = 0$) is $\mu_F(S_\theta)$. The linearity and mean-square continuity of $\mu_F(S_\theta)$, Assumption 2.1, and the Riesz representation theorem imply the existence of a unique (up to the usual a.s. equivalence) random vector $d(z)$, the *pathwise derivative*, such that $E[d(z)] = 0$, $E[d(z)^2] < \infty$, and

$$(2.3) \qquad \mu_F(S_\theta) = E[d(z)S_\theta(z)].$$

Under Assumption 2.1 and with i.i.d. data the asymptotic variance bound for estimators of $\mu(F_z)$ is $E[d(z)d(z)']$. Hence, the formula for the asymptotic variance of $\hat{\beta}$ suggested here is the variance of the pathwise derivative of the functional $\hat{\beta}$ is a nonparametric estimator of.

A stronger justification for regarding the pathwise derivative of $\mu(F_z)$ as a correct formula for the asymptotic variance of $\hat{\beta}$ is available when $\hat{\beta}$ is asymptotically equivalent to a sample average. Define $\hat{\beta}$ to be *asymptotically linear* with influence function $\psi(z)$ if at the truth,

$$(2.4) \qquad \sqrt{n}(\hat{\beta} - \beta_0) = \sum_{i=1}^{n}\psi(z_i)/\sqrt{n} + o_p(1), \quad E[\psi(z)] = 0, \quad \text{Var}(\psi(z)) \text{ finite.}$$

This condition is satisfied by many semiparametric estimators, under sufficient regularity conditions. For i.i.d. data, asymptotic linearity and

the central limit theorem imply $\hat{\beta}$ is asymptotically normal with variance $\text{Var}(\psi(z))$. Define $\hat{\beta}$ to be a *regular* estimator of $\mu(F_z)$ if for any path in $\mathcal{P}$, and $\theta_n = O(1/\sqrt{n})$, when z_i has distribution F_{θ_n}, $\sqrt{n}(\hat{\beta}-\mu(F_{\theta_n}))$ has a limiting distribution that does not depend on $\{\theta_n\}_{n=1}^{\infty}$ or on the path. Regularity is the precise condition that specifies that $\hat{\beta}$ is a nonparametric estimator of $\mu(F_z)$, because it requires that $\hat{\beta}$ is asymptotically, locally consistent for $\mu(F_z)$.

Theorem 2.1: Suppose that $z_1, z_2, \ldots$ *are i.i.d,* $\hat{\beta}$ *is asymptotically linear and regular for* $\mathcal{P}$, *and Assumption 2.1 is satisfied. Then* $\mu(F_z)$ *is pathwise differentiable and* $\psi(z) = d(z)$.

The thing that seems to be novel here is the idea of applying this result to the functional $\mu(F_z)$ that is nonparametrically estimated by $\hat{\beta}$. The fact that asymptotic linearity and regularity imply pathwise differentiability follows by Van der Vaart (1991, Theorem 2.1), and the fact that Assumption 2.1 implies that there is only one influence function and that it equals the pathwise derivative, is a small additional step that has been discussed in Newey (1990a).

This result can also be used to detect whether an estimator is $\sqrt{n}$-consistent. As shown by Van der Vaart (1991), if equation (2.2) is satisfied but $\mu_F(S_\theta)$ is not mean-square continuous (i.e. $d(z)$ satisfying equation (2.3) does not exist) then no $\sqrt{n}$-consistent, regular estimator exists. For example, the value of a density function at a point does not have a mean-square continuous derivative, and neither does the functional that is nonparametrically estimated by Manski's (1975) maximum score estimator. The pathwise derivative does not help in finding the asymptotic distribution (at a slower than $\sqrt{n}$ rate) of such estimators, which can be quite complicated:

e.g. see Kim and Pollard (1989).

The hypotheses of Theorem 2.1 are not primitive, but the point of Theorem 2.1 is to formalize the statement that "under sufficient regularity conditions" the influence function of a semiparametric estimator is the pathwise derivative of the functional that is nonparametrically estimated by $\hat{\beta}$. In Sections 3 and 4, this result and some pathwise derivative calculations are used to derive propositions about semiparametric estimators. These results are labeled as "propositions" because primitive conditions for their validity are not given in Sections 3 and 4. They might also be labeled as "conjectures," although this word does not convey the same sense that the validity of the results only requires regularity conditions. In Sections 3 and 4, the solution to equation (2.3) is calculated using the chain rule of calculus, differentiation under integrals, integration, and $\partial \int a(z) dF_\theta / \partial \theta |_{\theta=0} = E[a(z)S_\theta(z)]$ for $a(z)$ with finite mean square where ever needed, and then in Sections 5 - 7 conditions for implied remainder terms to be small are given. This approach, with formal calculation followed by regularity conditions, is similar to that used in parametric asymptotic theory (e.g. for Edgeworth expansions), and is meant to illustrate the usefulness of the pathwise derivative calculation.

3. Semiparametric M-Estimators

The rest of the paper will focus on a class semiparametric m-estimators, obtained from moment conditions that can depend on estimated functions. Let $m(z,\beta,h)$ be a vector of functions with the same dimension as β, depending on a data observation z and a vector of unknown functions h. Let $\hat{h}(\beta)$

denote an estimator of h, with corresponding $m(z,\beta,\hat{h}(\beta))$. A semiparametric m-estimator $\hat{\beta}$ is one which solves an asymptotic moment equation

$$(3.1) \qquad \sum_{i=1}^{n} m(z_i,\beta,\hat{h}(\beta))/n = 0.$$

The general idea here is that $\hat{\beta}$ is obtained by a procedure that "plugs-in" an estimated function $\hat{h}(\beta)$, that can depend on β.

An early and important example is the Buckley and James (1979) estimator for censored regression. Other examples are Robinson's (1988) semiparametric regression estimator and Powell, Stock, and Stoker's (1989) weighted average derivative estimator. For a new example, consider a semi-linear model with additive nonparametric component, $E[y|x,v] = x'\beta_0 + \rho_1(v_1) + \rho_2(v_2)$, where $v = (v_1,v_2)$. The motivation for this model is that if v is high dimensional the asymptotic properties of $\hat{\beta}$ could be adversely affected if additivity of $\rho_1(v_1) + \rho_2(v_2)$ is true but not imposed: see Section 4 for further discussion. Assume that the set of additive functions in v_1 and v_2 with finite mean-square is closed in mean square, and let $\Pi(\cdot|v_1,v_2)$ denote the mean-square (Hilbert space) projection on this set. Also, let $\hat{\Pi}(\cdot|v_1,v_2)$ denote an estimator of this projection, such as the series estimator considered in Stone (1985) and in Section 6, or the alternating conditional expectation estimator in Breiman and Friedman (1985). Consider

$$(3.2) \qquad \hat{\beta} = \operatorname{argmin}_{\beta}\{\sum_{i=1}^{n}[y_i - x_i'\beta - \hat{h}(v_i,\beta)]^2/2\},$$

$$\hat{h}(v,\beta) = \hat{\Pi}(y|v_1,v_2) - \hat{\Pi}(x|v_1,v_2)'\beta.$$

This is a semiparametric m-estimator with $m(z,\beta,\hat{h}(\beta))$
$= [x + \partial\hat{h}(v,\beta)/\partial\beta][y - x'\beta - \hat{h}(v,\beta)]$.

It is possible, at the level of generality of equation (3.1), to derive a

number of propositions. To use the pathwise derivative formula in this derivation, it is necessary to identify the functional that is nonparametrically estimated by $\hat{\beta}$. Let $h(\beta,F)$ denote the limit of $\hat{h}(\beta)$ for a general distribution $F = F_z$, where the z subscript is suppressed henceforth for notational convenience. By the usual method of moments reasoning, the limit $\mu(F)$ of $\hat{\beta}$ for a general F should be the solution to

$$(3.3) \qquad E_F[m(z,\mu,h(\mu,F))] = 0.$$

That is, equation (3.1) sets $\hat{\beta}$ so that sample moments are zero, and the sample moments have a limit of $E_F[m(z,\beta,h(\beta,F))]$ (by the law of large numbers and $h(\beta,F)$ equal to the limit of $\hat{h}(\beta)$), so that $\hat{\beta}$ is consistent for that value of μ that sets the population moments to zero.

Before computing the pathwise derivative, it is interesting to note that it will depend only on the limit $h(\beta,F)$, and not on the particular form of the estimator $\hat{h}(\beta)$. Thus, different nonparametric estimators of the same functions should result in the same asymptotic variance. For example, this reasoning explains why replacing the kernel estimator of Robinson (1988) by series estimators gives an asymptotically equivalent estimator, as shown by Newey (1990b), and suggests that for estimation of the additive model above, the distribution is invariant to the estimator of the projection. Also, two estimators may not be asymptotically equivalent if the nuisance functions estimate different objects nonparametrically.

Proposition 1: The asymptotic variance of semiparametric estimators depends only on the function that is nonparametrically estimated, and not on the type of estimator (such as kernel or series nonparametric regression).

To obtain more results, it is useful to be more specific about the form

of the pathwise derivative. Suppose that h has J components, $h = (h_1,\ldots,h_J)$. For a path F_θ, equal to the truth when $\theta = \theta_0 = 0$, let $E_\theta[\cdot]$ denote the expectation with respect to F_θ, $h_j(\beta,\theta) = h_j(\beta,F_\theta)$, $h_j(\beta) = h_j(\beta,\theta_0)$, $h_j(\theta) = h_j(\beta_0,\theta)$, and let the same expressions without the j subscript denote corresponding vectors. For a path, $\mu(\theta)$ will be the functional satisfying the parametric version of equation (3.2),

$$E_\theta[m(z,\mu,h(\mu,\theta))] = 0. \tag{3.4}$$

Then for $m(z,h(\theta)) = m(z,\beta_0,h(\theta))$, differentiation gives

$$\partial E_\theta[m(z,h(\theta_0))]/\partial\theta|_{\theta_0} = \int m(z,h(\theta_0))[\partial dF_\theta/\partial\theta]dz|_{\theta_0} = E[m(z,h(\theta_0))S_\theta(z)].$$

Then, applying the chain rule to $E_{\theta_1}[m(z,h(\theta_2))]$, it follows that

$$\partial E_\theta[m(z,h(\theta))]/\partial\theta|_{\theta_0} = E[m(z,h(\theta_0))S_\theta(z)] + \partial E[m(z,h(\theta))]/\partial\theta|_{\theta_0}.$$

Assuming $D \equiv \partial E[m(z,\beta,h(\beta,\theta_0))]/\partial\beta|_{\beta_0}$ is nonsingular, the implicit function theorem gives

$$\partial\mu(\theta)/\partial\theta|_{\theta_0} = -D^{-1}\{E[m(z,h(\theta_0))S_\theta(z)] + \partial E[m(z,h(\theta))]/\partial\theta|_{\theta_0}\}. \tag{3.5}$$

The first term is already in outer product form of equation (2.3), so that the pathwise derivative will exist if the second term can be put in a similar form. Suppose there are $\alpha_j(z)$ such that for each $j = 1, \ldots, J$,

$$\partial E[m(z,h_1(\theta_0),\ldots,h_j(\theta),\ldots,h_J(\theta_0))]/\partial\theta|_{\theta_0} = E[\alpha_j(z)S_\theta(z)]. \tag{3.6}$$

Then, applying the chain rule to $E[m(z,h_1(\theta_1),\ldots,h_j(\theta_j),\ldots,h_J(\theta_J))]$ with each θ_j equal to θ, it follows that

$$\partial E[m(z,h(\theta))]/\partial\theta|_{\theta_0} = \sum_{j=1}^{J} \partial E[m(z,h_1(\theta_0),\ldots,h_j(\theta),\ldots,h_J(\theta_0))]/\partial\theta|_{\theta_0}$$

$$= \sum_{j=1}^{J} E[\alpha_j(z)S_\theta(z)] = E[\{\sum_{j=1}^{J}\alpha_j(z)\}S_\theta(z)],$$

giving the outer product form. Then, moving $-D^{-1}$ inside the expectation, it follow that the pathwise derivative is $d(z) = -D^{-1}[m(z,h(\theta_0)) + \{\sum_{j=1}^{J}\alpha_j(z)\}]$, so that by Theorem 2.1 the influence function of $\hat{\beta}$ equals

$$\psi(z) = -D^{-1}\{m(z,\beta_0,h(\beta_0)) + \sum_{j=1}^{J}[\alpha_j(z) - E[\alpha_j(z)]]\}. \tag{3.7}$$

This influence function has an interesting structure. The leading term $-D^{-1}m(z,\beta_0,h(\beta_0))$ is the usual Huber (1967) formula for the influence function of an m-estimator with moment functions $m(z,\beta,h(\beta))$, i.e. the formula that would be obtained if $\hat{h}(\beta)$ were equal to $h(\beta)$. Thus, the second term is an adjustment term for the estimation of $h(\beta)$, a nonparametric analog of adjustments that are familiar for two-step parametric estimators. It can also be interpreted as the pathwise derivative of the functional $D^{-1}E[m(z,\beta_0,h(\beta_0,F))]$, or as the influence function of $D^{-1}\int m(z,\beta_0,\hat{h}(\beta_0))dF(z)$. Furthermore, the adjustment contains exactly one term for each component of h, and the j^{th} adjustment can be interpreted as the pathwise derivative of $E[m(z,\beta_0,h_1(\beta_0),\ldots,h_j(\beta_0,F),\ldots,h_J(\beta_0))]$. This property is useful, because the adjustment terms can be calculated for each function h_j, holding the other functions fixed at their true values, and then the total adjustment formed as the sum. For this reason the j subscript will be dropped in the rest of Sections 3 and 4, with the understanding that the results can be applied to individual h_j terms, and then combined to derive the total adjustment (e.g. when some adjustment terms are zero and others are not).

It is useful to know when an adjustment term is zero. In such cases, it should not be necessary to account for the presence of $\hat{h}(\beta)$, i.e. $\hat{h}(\beta)$ can be treated as if it were equal to $h(\beta)$, greatly simplifying the calculation of the asymptotic variance and finding a consistent estimator of it. One case where an adjustment term will be zero is when equation (3.1) is the first-order condition to a maximization problem, and $\hat{h}(\beta)$ has a limit that maximizes the population value of the same function. To be specific, suppose that there is a function $q(z,\beta,h(\beta))$ and a set of functions $\mathcal{H}(\beta)$, possibly depending on β but not on the distribution F of z, such that

$$(3.8)\quad m(z,\beta,h(\beta)) = \partial q(z,\beta,h(\beta))/\partial\beta,\ h(\beta,F) = \operatorname{argmax}_{\tilde{h}(\beta)\in\mathcal{H}(\beta)} E_F[q(z,\beta,\tilde{h}(\beta))].$$

The interpretation of this condition is that $m(z,\beta,h(\beta))$ are the first order conditions for a stationary point of the function q and that $h(\beta,F)$ maximizes the expected value of the same function, i.e. that $h(\beta,F)$ has been "concentrated out." Then for any parametric model F_θ, since $h(\beta,\theta) = h(\beta,F_\theta)$, it follows that $E[q(z,\beta,h(\beta,\theta))]$ is maximized at θ_0. The first order conditions for this maximization are $\partial E[q(z,\beta,h(\beta,\theta))]/\partial\theta|_{\theta_0} = 0$, *identically* in β. Differentiating again with respect to β,

$$(3.9)\quad 0 = \partial^2 E[q(z,\beta,h(\beta,\theta))]/\partial\theta\partial\beta|_{\theta_0} = \partial E[\partial q(z,\beta,h(\beta,\theta))/\partial\beta]/\partial\theta|_{\theta_0}$$

$$= \partial E[m(z,\beta,h(\beta,\theta))]/\partial\theta|_{\theta_0}.$$

Evaluating this equation at β_0, it follows that $\alpha(z) = 0$ will solve equation (3.6), and hence the adjustment term is zero. Summarizing:

Proposition 2: If equation (3.8) is satisfied, then the estimation of h can be ignored in calculating the asymptotic variance, i.e. it is the same as if $\hat{h}(\beta) = h(\beta)$.

Examples of estimators that satisfy the hypotheses of this proposition are those of Robinson (1988), Ichimura (1987), Klein and Spady (1987), and Ai (1990). A new example is the additive semi-linear estimator of equation (3.2). Suppose that the set $\mathcal{H}$ of additive functions is closed under any $F \in \mathcal{F}$ and is invariant to F, and let $\Pi_F(\cdot|v_1,v_2)$ denote the projection under F. Then $\hat{h}(\beta)$ is a nonparametric estimator of $\Pi_F(y|v_1,v_2) - \Pi_F(x|v_1,v_2)'\beta = \Pi_F(y-x'\beta|v_1,v_2)$, which minimizes $E_F[(y - x'\beta - h(v,\beta))^2]$, the same objective function minimized by the limit of $\hat{\beta}$. Therefore, by Proposition 2, estimation of $\Pi_F(y|v_1,v_2)$ and $\Pi_F(x|v_1,v_2)$ should have no effect on the asymptotic variance of $\hat{\beta}$. Thus, for $\varepsilon = y-x'\beta_0-\rho_1(v_1)-\rho_2(v_2)$, the formula for the influence function is

$$(3.10) \qquad \psi(z) = (E[\{x-\Pi[x|v_1,v_2]\}\{x-\Pi[x|v_1,v_2]\}'])^{-1}\{x-\Pi[x|v_1,v_2]\}\varepsilon.$$

Primitive conditions for this result are given in Newey (1991), and somewhat weaker conditions could be formulated using the results of Section 6.

There is another, more direct condition under which estimation of the nuisance function does not affect the asymptotic variance. To formulate this condition, suppose that $m(z,h)$ depends on h only through the value it takes on as a function $h(v)$ of a subvector v of z, i.e. h is a real vector argument in $m(z,h)$. The additive semi-linear example has this property if $\hat{h}(\beta)$ is redefined to include $\hat{\Pi}[x|v_1,v_2]$. Let $h(v,\theta)$ denote the limiting value of $\hat{h}(v,\beta_0)$ for a path. For $M(z) = \partial m(z,\beta_0,h)/\partial h|_{h=h(v)}$, differentiation gives

$$(3.11) \qquad \partial E[m(z,h(\theta))]/\partial\theta|_{\theta_0} = E[M(z)\partial h(v,\theta)/\partial\theta|_{\theta_0}] = \partial E[M(z)h(v,\theta)]/\partial\theta|_{\theta_0}.$$

If the term on the right-hand side is zero, then $\alpha(z) = 0$ will solve equation (3.6), and the adjustment term is zero. One simple condition for

this is that $E[M(z)|v] = 0$. More generally, the adjustment term will be zero if $h(v,\theta)$ is an element of a set to which $M(z)$ is orthogonal.

Proposition 3: If $E[M(z)|v] = 0$, or more generally $h(v,F)$ is an element of a set $\mathcal{H}$ such that $E[M(z)\tilde{h}(v)] = 0$ for all $\tilde{h} \in \mathcal{H}$, then estimation of h can be ignored in calculating the asymptotic variance.

The semi-linear, additive model is also an example here.

In cases where the correction term is nonzero, its form will depend on the limit of $\hat{h}(\beta)$. Therefore, it is difficult to give a general characterization of the correction term. One result that does not depend on completely specifying the form of h can be obtained in semiparametric models where the data, $z_1, \ldots, z_n$ are i.i.d. and z_i is restricted to have a density function of the form $f(z|\beta,g)$, where g is a nonparametric (functional) component. Let $S_\eta(z) = \partial \ln f(z|\beta_0, g(\eta))/\partial\eta|_{\eta_0}$ denote the score for a finite-dimensional parameterization of g with $g(\eta_0) = g_0$, (g_0 is the truth), and let $S_\beta(z) = \partial \ln f(z|\beta, g_0)/\partial\beta|_{\beta_0}$. Also, let A denote a constant matrix with number of rows equal to the number of elements of β. The tangent set $\mathcal{T}$ is defined as the mean-square closure of the set of all linear combinations $AS_\eta(z)$. The tangent set is useful in calculating the asymptotic variance bound for estimators of β in the semiparametric model $f(z|\beta,g)$. The form of this bound is $V = (E[S(z)S(z)'])^{-1}$, where $S(z) = S_\beta(z) - \Pi(S_\beta(z)|\mathcal{T})$ and $\Pi(\cdot|\mathcal{T})$ denotes the mean-square projection on the tangent set. See, for example, Newey (1990a) for further discussion.

Under certain conditions, $\alpha(z) = -\Pi(m(z)|\mathcal{T})$ will solve equation (3.6), for $m(z) = m(z,\beta_0,h(\beta_0))$, so that the correction term can be calculated from this projection. Let θ be the parameter of an unrestricted path, as discussed in Section 2 (θ *does not* have anything to do with β or η).

Suppose that there is $g(\theta)$ such that,

(3.12) $\int m(z,h(\theta))f(z|\beta_0,g(\theta))dz = 0.$

In words, for the limit of $\hat{h}(\beta_0)$ under a general distribution there is a corresponding value of the nonparametric component of the semiparametric model where the population moment conditions (corresponding to equation (3.1)) are satisfied. Let $S_{g\theta}(z) = \partial \ln f_0(z|g(\theta))/\partial\theta$, and note that $S_{g\theta}(z)$ is an element of the tangent set, implying $E[m(z)S_{g\theta}(z)] = E[\Pi(m(z)|\mathcal{T})S_{g\theta}(z)]$. Suppose that $S_{g\theta}(z) = \Pi(S_\theta(z)|\mathcal{T})$. Then differentiating equation (3.12) with respect to θ,

$$\partial E[m(z,h(\theta))]/\partial\theta|_{\theta_0} = -E[m(z)S_{g\theta}(z)] = -E[m(z)\Pi(S_\theta(z)|\mathcal{T})]$$

$$= -E[\Pi(m(z)|\mathcal{T})\Pi(S_\theta(z)|\mathcal{T})] = E[-\Pi(m(z)|\mathcal{T})S_\theta(z)].$$

Thus, under the previous conditions, $\alpha(z) = -\Pi(m(z)|\mathcal{T})$ satisfies equation (3.6). Summarizing:

Proposition 4: If for all unrestricted paths $F_{z\theta}$ there exists $g(\theta)$ such that equation (3.12) is satisfied, and $\partial \ln f(z|\beta_0,g(\theta))/\partial\theta = \Pi(S_\theta(z)|\mathcal{T})$, then $\alpha(z) = -\Pi(m(z,\beta_0,h(\beta_0))|\mathcal{T})$ and the influence function of $\hat{\beta}$ is $-D^{-1}[m(z,\beta_0,h(\beta_0)) - \Pi(m(z,\beta_0,h(\beta_0))|\mathcal{T})]$.

This form of the correction term has previously been derived by Bickel, Klaassen, Ritov, and Wellner (1990) and Newey (1990a). The contribution of Proposition 4 is to give a general formulation for this result in terms of the pathwise derivative calculation developed in Section 2.

This result leads to a sufficient condition for asymptotic efficiency of a semiparametric m-estimator, that the hypotheses of Proposition 4 are

satisfied and $m(z) = S_\beta(z)$. In this case, $m(z) + \alpha(z) - E[\alpha] = S_\beta(z) - \Pi(S_\beta(z)|\mathcal{T}) = S(z)$. Furthermore, any semiparametric m-estimator that is regular under the semiparametric model $f(z|\beta,g)$ and has influence function $-D^{-1}(m(z)+\alpha(z)-E[\alpha])$ will satisfy

$$D = -E[(m(z)+\alpha(z)-E[\alpha])S(z)'], \tag{3.13}$$

as discussed in Newey (1990a). Thus, the influence function of $\hat{\beta}$ is $(E[S(z)S(z)'])^{-1}S(z) = VS(z)$, with corresponding asymptotic variance $VE[S(z)S(z)']V = V$, which equals the lower bound.

4. Functions of Mean-Square Projections and Densities

In this section, the form of the correction term is derived when the nuisance functions are linear functions of conditional expectations or other mean-square projections, such as additive or partially linear regressions, and for densities. Let y be a random variable with finite second moment and x an $r \times 1$ vector. Let $\mathcal{G}$ denote a linear set of functions of x that is closed in mean-square and $g(x)$ denote the least squares (Hilbert-space) projection of y on x, that is $g(x) = \operatorname{argmin}_{\tilde{g}\in\mathcal{G}} E[(y-\tilde{g}(x))^2]$. One $h(v)$ considered in this section will be $h(v) = A(g,v)$, where A is a linear function of g, and v is a subvector of z.

The simplest nonparametric example of a projection is $g(x) = E[y|x]$, where $\mathcal{G}$ is all measurable functions of x with finite mean-square. A more general example is a projection on

$$\mathcal{G} = \{\textstyle\sum_{\ell=1}^{L} \tilde{g}_\ell(\tilde{x}_\ell) + \tilde{x}_{L+1}'\eta\}, \tag{4.1}$$

with each $\tilde{x}_\ell$ a subvector of x. This is a smaller set of functions, whose consideration is motivated partly by the difficulty of estimating conditional expectations for x with many dimensions; e.g. see Stone (1985) for discussion and references. Here, where $g(x)$ is a nuisance function, important reasons to avoid high dimensional nonparametric regressions are that a projection on a larger set of functions than that to which $g(x)$ belongs will lead to higher asymptotic variances for β in some cases, as noted in Newey (1991), and will lower the rate at which remainder terms converge to zero, affecting accuracy of the asymptotic normal approximation.

The correction term is derived first for the simplest case, where $h(v) = g(x)$. Let $g(x,\theta) = \text{argmin}_{\tilde{g}\in\mathcal{G}} E_\theta[\{y - \tilde{g}(x)\}^2]$ denote the projection of y on $\mathcal{G}$ for a path. Note that for the vector of projections of elements of $M(z)$ on $\mathcal{G}$, $\delta(x) = \Pi(M(z)|\mathcal{G})$, it follows that $E[M(z)g(x,\theta)] = E[\delta(x)g(x,\theta)]$ identically in θ. Also, by $\delta(x) \in \mathcal{G}$, $E_\theta[\delta(x)g(x,\theta)] = E_\theta[\delta(x)y]$, so by the chain rule,

$$
\begin{aligned}
(4.2)\qquad & E[M(z)\partial g(x,\theta_0)/\partial\theta]|_{\theta_0} = \partial E[M(z)g(x,\theta)]/\partial\theta|_{\theta_0} = \partial E[\delta(x)g(x,\theta)]/\partial\theta|_{\theta_0} \\
& = \{\partial E_\theta[\delta(x)g(x,\theta)]/\partial\theta - \partial E_\theta[\delta(x)g(x)]/\partial\theta\}|_{\theta_0} \\
& = \partial E_\theta[\delta(x)\{y-g(x)\}]/\partial\theta|_{\theta_0} = E[\delta(x)\{y-g(x)\}S_\theta(z)]|_{\theta_0}.
\end{aligned}
$$

Equation (4.2) implies the next result.

Proposition 5: If $h(v) = g(x)$ is the projection of y on $\mathcal{G}$, then the correction term is $\alpha(z) = \Pi(M(z)|\mathcal{G})[y-g(x)]$.

A new example is an estimator for a semiparametric random effects model. Let (y_t,x_t) $(t=1,2)$, be sets of observations for two time periods, where y_t is binary, and suppose that for $x = (x_1,x_2)$, $E[y_t|x] = \Phi([x_t\beta_{10}+\rho(x)]/\sigma_{t0})$,

$\sigma_{10} = 1$, $\rho(x)$ is an unknown function, and Φ denotes the standard normal CDF. This is a binary panel data model with $y_t = 1(x_t\beta_{10} + \alpha + \varepsilon_t > 0)$ for an individual effect α, and the conditional distribution of $\alpha + \varepsilon_t$ given x is $N(\rho(x),\sigma_{t0}^2)$. This model generalizes Chamberlain's (1980) random effects model by allowing the conditional mean of α to be unknown. In contrast to Manski's (1987) semiparametric individual effects model, ε_t is allowed to be heteroskedastic over time, but the conditional distribution of $\alpha + \varepsilon_t$ is restricted to be Gaussian.

An implication of this model is that

$$(4.3) \qquad \Phi^{-1}(E[y_1|x]) = \sigma_{20}\Phi^{-1}(E[y_2|x]) + (x_1-x_2)\beta_{10}.$$

This implication can be used to construct a semiparametric minimum distance estimator by replacing the conditional expectations with nonparametric estimators $\hat{h}_t(x) = \hat{E}[y_t|x]$ and choosing $\hat{\beta}_1$ and $\hat{\sigma}_2$ from the least squares regression of $\Phi^{-1}(\hat{h}_1(x_i))$ on $x_{1i}-x_{2i}$ and $\Phi^{-1}(\hat{h}_2(x_i))$. This estimator can also be generalized to the case where the distribution of disturbances is unknown, by normalizing the scale of β and replacing Φ^{-1} by a series approximation to the unknown inverse marginal distribution functions, although further development of this estimator is beyond the scope of this paper.

To derive the influence function of the estimator of β_1 and σ_2, note that it is a semiparametric m-estimator with $\beta = (\beta_1',\sigma_2)'$, $v = x$, $m(z,\beta,h(v,\beta)) = A(x,h_2)[\Phi^{-1}(h_1(x)) - \Phi^{-1}(h_2(x))\sigma_2 - (x_1-x_2)\beta]$, and $A(x,h_2) = [x_1-x_2,\Phi^{-1}(h_2(x))]'$. Here, the correction terms are the only source of variation, since $m(z,\beta_0,h(v,\beta_0)) = 0$. Also, $D = -E[A(x,h_2)A(x,h_2)']$ and $M_j(z) = A(x,h_2)(-1)^{j-1}\phi(\Phi^{-1}(h_j(x)))^{-1}$. Then by Proposition 5, $\alpha_j(z) = A(x,h_2)(-1\sigma_2)^{j-1}\phi(\Phi^{-1}(h_j(x)))^{-1}[y_j-h_j(x)]$,

(4.4) $$\psi(z) = -D^{-1}[\alpha_1(z)+\alpha_2(z)] = -D^{-1}A(x,h_2) \cdot$$

$$\{\phi(\Phi^{-1}(h_1(x)))^{-1}[y_1-h_1(x)] - \sigma_2\phi(\Phi^{-1}(h_2(x)))^{-1}[y_2-h_2(x)]\}.$$

Proposition 5 can be generalized to linear functionals of the projection that have a particular property specified in the following assumption.

Assumption 4.1: $h(v) = A(v,g)$, $A(v,g)$ is a linear function of g, and there is $\delta(x) \in \mathcal{G}$ such that for all $\tilde{g}(x) \in \mathcal{G}$,

(4.5) $$E[M(z)A(v,\tilde{g})] = E[\delta(x)\tilde{g}(x)].$$

By the Riesz representation theorem, equation (4.5) is equivalent to assuming that the functional $E[M(z)A(v,\tilde{g})]$ is mean-square continuous in $\tilde{g}$. This condition is necessary for $\int M(z)h(v)dF(z)$ to be a $\sqrt{n}$-consistently estimable functional of $h(v)$, as discussed in Newey (1991), so that the estimation of $\hat{h}(v)$ will affect the convergence rate of $\hat{\beta}$ unless Assumption 4.1 is satisfied. Thus, for $h(v)$ a linear function of $g(x)$, Assumption 4.1 and the form of the correction term given below characterize the adjustment for mean-square projections.

Equation (4.5) leads to a straightforward form for the correction term. Noting that $h(v,\theta) = A(v,g(\theta))$, differentiation gives

(4.6) $$E[M(z)\partial h(v,\theta)/\partial\theta]|_{\theta_0} = \partial E[M(z)h(v,\theta)]/\partial\theta|_{\theta_0} = \partial E[M(z)A(v,g(\theta))]/\partial\theta|_{\theta_0}$$

$$= \partial E[\delta(x)g(x,\theta)]/\partial\theta|_{\theta_0} = E[\delta(x)\{y-g(x)\}S_\theta(z)],$$

where the last equality follows as in equation (4.2).

Proposition 6: If Assumption 4.1 is satisfied, the correction term is $\alpha(z) = \delta(x)[y-g(x)]$.

In order for this result to provide an interesting formula, it must be possible to find $\delta(x)$. In a number of cases $\delta(x)$ takes a projection form similar to that of Proposition 5. One interesting case is that where $x = (x_1,x_2)$, x_j may be a vector, $v = x_2$, and $\tilde{h}(v) = A(v,\tilde{g}) = \int_{\mathcal{A}}\tilde{g}(x_1,x_2)dx_1$. In this case, $E[M(z)\tilde{h}(v)] = E[E[M(z)|v]\tilde{h}(v)] = E[\int_{\mathcal{A}}E[M(z)|x_2]\tilde{g}(x)dx_1] = E[1(x_1\in\mathcal{A})f(x_1|x_2)^{-1}E[M(z)|x_2]\tilde{g}(x)] = E[\Pi(1(x_1\in\mathcal{A})f(x_1|x_2)^{-1}E[M(z)|x_2]|\mathcal{G})\tilde{g}(x)]$, where $f(x_1|x_2)$ is the conditional density of x_1 given x_2.

Proposition 7: If $h(v) = \int_{\mathcal{A}}g(x_1,x_2)dx_1$, x_1 *is absolutely continuous with respect to the product measure corresponding to* dx_1 *and the distribution of* x_2 *with density* $f(x_1|x_2)$, *and* $1(x_1\in\mathcal{A})f(x_1|x_2)^{-1}E[M(z)|x_2]$ *has finite second moment, then the correction term is* $\delta(x)[y-g(x)]$ *for* $\delta(x) = \Pi(1(x_1\in\mathcal{A})f(x_1|x_2)^{-1}E[M(z)|x_2]|\mathcal{G})$.

An example is average approximate consumer surplus, where x_1 is a price variable and $\hat{\beta} = \sum_{i=1}^{n}\int_a^b\hat{g}(x_1,x_{2i})dx_1/n$, which is a semiparametric m-estimator with $M(z) = 1$. By Proposition 8, the influence function for this estimator will be

$$(4.7)\qquad \psi(z) = \int_a^b g(x_1,x_2)dx_1 - \beta_0 + \Pi(1(a\le x_1\le b)f(x_1|x_2)^{-1}|\mathcal{G})[y - g(x)].$$

Results for exact consumer surplus (i.e. equivalent variation) and where the demand function is a nonlinear function of a projection (e.g. log-linear models) are analyzed in Hausman and Newey (1991).

Another case where $\delta(x)$ takes a projection form is where $h(v)$ is a derivative of a projection evaluated at some other variable v. For $x \in \mathbb{R}^r$, and a vector $\lambda = (\lambda_1,\ldots,\lambda_r)'$ of nonnegative integers, let $|\lambda| = \sum_{j=1}^{r}\lambda_j$ and denote a partial derivative by

$$(4.8)\qquad D^{\lambda}g(x) = \partial^{|\lambda|}g(x)/\partial x_1^{\lambda_1}\cdots\partial x_r^{\lambda_r}.$$

Suppose that $h(v) = D^{\lambda}g(x)$. Let $M(v) = E[M(z)|v]$ and $f_v(v)$ and $f_x(x)$ be the densities of v and x respectively, with respect to the same dominating measure. Assuming that v has a density that is differentiable to sufficient order in the components of v corresponding to nonzero components of λ, with zero derivatives on the boundary of its support, repeated integration by parts gives $E[M(z)\tilde{h}(v)] = \int M(v)D^{\lambda}\tilde{g}(v)f_v(v)dv = (-1)^{|\lambda|}\int D^{\lambda}[M(v)f_v(v)]g(v)dv = (-1)^{|\lambda|}\int D^{\lambda}[M(v)f_v(v)]|_{v=x}g(x)dx = E[(-1)^{|\lambda|}f_x(x)^{-1}D^{\lambda}[M(v)f_v(v)]_{v=x}g(x)] = E[\delta(x)g(x)]$, $\delta(x) = (-1)^{|\lambda|}\Pi(f_x(x)^{-1}D^{\lambda}[M(v)f_v(v)]_{v=x}|\mathcal{G})$.

Proposition 8: If $h(v) = D^{\lambda}g(x)|_{x=v}$, v *and* x *are absolutely continuous with respect to the same measure, which is Lebesgue measure for the components* $\bar{x}$ *of* x *corresponding to nonzero components of* λ, *the density* $f_v(v)$ *and* $E[M(z)|v]$ *are continuously differentiable to order* $|\lambda|$ *in* $\bar{x}$, *the support of* $\bar{x}$ *is a convex set with nonempty interior, and for each* $\tilde{\lambda} \leq \lambda$, $D^{\tilde{\lambda}}f_v(v)$ *is zero on the boundary of the support of* $\bar{x}$ *and* $f_x(x)^{-1}D^{\tilde{\lambda}}[M(v)f_v(v)]_{v=x}$ *has finite second moment, then the correction term is* $\delta(x)[y-g(x)]$ *for* $\delta(x) = (-1)^{|\lambda|}\Pi(f_x(x)^{-1}D^{\lambda}[M(v)f_v(v)]_{v=x}|\mathcal{G})$.

An example with no derivatives involved is Stock's (1989) nonparametric prediction estimator, where $\mathcal{G} = \{g_1(x_1) + x_2'\eta\}$, so that $g(x)$ is a partially linear projection, $v = (v_1, x_2)$ is partitioned conformably with x, and $\hat{\beta} = \sum_{i=1}^{n}[\hat{g}(v_i)-\hat{g}(x_i)]/n$. This is a semiparametric m-estimator where $\beta_0 = E[g(v)]-E[g(x)]$, $\hat{h}_1(v) = \hat{g}(v)$, $\hat{h}_2(x) = \hat{g}(x)$, and $M_1(z) = M_2(z) = 1$. From the form of the correction terms in Proposition 8, the influence function of β_0 is

$$\psi(z) = g(v)-g(x)-\beta_0 + [\Pi(f_x(x)^{-1}f_v(x)|\mathcal{G}) - 1][y - g(x)]. \tag{4.9}$$

This result differs from Stock's in the inclusion of the term $g(v)-g(x)-\beta_0$,

because Stock's result only derived the conditional distribution given the observations on x and v. The variance of the second term is the same as Stock's formula, because $\Pi(f_x(x)^{-1}f_v(x)|\mathcal{G}) = f_{x_1}(x_1)^{-1}f_{v_1}(x_1) + (x_2-E[x_2|x_1])E[Var(x_2|x_1)]^{-1}E[x_2-E[x_2|x_1]_{x_1=v_1}]$, for $f_{x_1}(x_1)$ and $f_{v_1}(v_1)$ equal to the densities of x_1 and v_1 respectively. Proposition 8 also gives the form of correction terms for the dynamic discrete choice estimators of Ahn and Manski (1989) and Hotz and Miller (1989), and the average derivative estimator of Hardle and Stoker (1989).

A correction term for density estimation can be derived under conditions similar to those for the projection. Suppose that $h(v) = A(v,f_w)$, where $A(v,f_w)$ is a linear function of the density $f_w(w)$ of a vector w, with respect to some measure. Suppose that there is $\alpha(w)$ such that $E[M(z)A(v,\tilde{f}_w)] = \int\alpha(w)\tilde{f}_w(w)dw$. Let $f_w(w|\theta)$ denote the density of w for a path. Then

$$E[M(z)\partial h(v,\theta)/\partial\theta]|_{\theta_0} = \partial E[M(z)A(v,f_w(\theta))]/\partial\theta|_{\theta_0}$$

$$= \partial E_\theta[\alpha(w)]/\partial\theta|_{\theta_0} = E[\alpha(w)S_\theta(z)].$$

Proposition 9: If $h(v) = A(v,f_w)$ for a density f_w and there is $\alpha(w)$ such that $E[M(z)A(v,f_w)] = \int\alpha(w)f_w(w)dw$ then the correction term is $\alpha(w)-E[\alpha(w)]$.

Existence of such a $\alpha(w)$ will follow from the Riesz representation theorem if $\int f_w(w)^2dw$ is finite and $E[M(z)A(v,f_w)]$ can be extended to a linear functional on the Hilbert space of square integrable (dw) functions that is continuous. Continuity of $E[M(z)A(v,f_w)]$ in f_w, in the square integrable sense, appears to be essentially necessary for the correction term to be $\sqrt{n}$-consistent, although it is difficult to give a precise result, because the

usual parameterization for checking $\sqrt{n}$-consistency is the square root of the density, rather than density itself.

A case where is it easy to compute a density correction term $\alpha(w)$ is that with $w = (y',x')$ and $h(v) = D^{\lambda}[\int a(y)f_w(w)dy]|_{x=v}$. Integration by parts gives $E[M(z)A(v,\tilde{f}_w)] = \int M(v)D^{\lambda}[\int a(y)\tilde{f}_w(w)dy]|_{x=v}f_v(v)dv = (-1)^{|\lambda|}\int D^{\lambda}[M(v)f_v(v)]|_{v=x}[\int a(y)\tilde{f}_w(w)dy]dx = \int \alpha(w)\tilde{f}_w(w)dw$, $\alpha(w) = (-1)^{|\lambda|}D^{\lambda}[M(v)f_v(v)]|_{v=x}a(y)$.

Proposition 10: If $h(v) = D^{\lambda}\int a(y)f_w(y,x)dy|_{x=v}$, v *and* x *are absolutely continuous with respect to the same measure, which is Lebesgue measure for the components* $\bar{x}$ *of* x *corresponding to nonzero components of* λ, *the density* $f_v(v)$ *and* $E[M(z)|v]$ *are continuously differentiable to order* $|\lambda|$ *in* $\bar{x}$, *the support of* $\bar{x}$ *is a convex set with nonempty interior, and for each* $\tilde{\lambda} \leq \lambda$, $D^{\tilde{\lambda}}f_v(v)$ *is zero on the boundary of the support of* $\bar{x}$ *and* $D^{\tilde{\lambda}}[M(v)f_v(v)]|_{v=x}a(y)$ *has finite second moment, then the correction term is then the correction term is* $\alpha(w) - E[\alpha(w)]$ *for* $\alpha(w) = (-1)^{|\lambda|}D^{\lambda}[M(v)f_v(v)]|_{v=x}a(y)$.

This result gives the form of the correction term for Powell, Stock, and Stoker's (1989) weighted average derivative estimator and Robinson's (1989) test statistics. Another example is Ruud's (1986) density weighted least squares estimator, which is treated in Newey and Ruud (1991).

There may be other interesting cases where the form of the correction term can be calculated. Hopefully, the ones given here illustrate the usefulness of the pathwise derivative calculation of the influence function. In the next two Sections, regularity conditions for the validity of many of these calculations are given.

5. Regularity Conditions.

This Section develops a set of regularity conditions that are sufficient for validity of the pathwise derivative formula. The regularity conditions are based on direct verification that remainder terms from the pathwise derivative are small.

The rest of the paper will focus on semiparametric generalized method of moments estimators where $\hat{h}(v)$ does not depend on parameters. Let $m(z,\beta,h)$ be a vector of functions of the data observation z, the $q \times 1$ parameter vector β and a $J \times 1$ vector h, where h represents a possible value of a vector of functions $h(v) = (h_1(v_1),\ldots,h_J(v_J))'$ and each v_j is a vector. Also, assume that the moment condition $E[m(z_i,\beta_0,h(v_i))] = 0$ is satisfied. Note that this setup allows $h(v)$ to include parameter values, by specifying that some v_j are trivial (can only take on one value). For example, some elements of $h(v)$ might be trimming parameters, as in Newey and Ruud (1991). Let $\hat{h}(v)$ denote an estimator of this vector function, $\hat{m}_n(\beta) = \sum_{i=1}^n m(z_i,\beta,\hat{h}(v_i))/n$, and $\hat{W}$ a positive semi-definite matrix. The estimator to be analyzed satisfies

$$(5.1) \qquad \hat{\beta} = \operatorname{argmin}_{\beta\in B} \hat{m}_n(\beta)'\hat{W}\hat{m}_n(\beta).$$

Although $\hat{h}$ is not allowed to depend on β in this section, the results are still useful for the general case, because they provide conditions for the important intermediate result that $\sum_{i=1}^n m(z_i,\beta_0,\hat{h}(v_i,\beta_0))/\sqrt{n}$ is asymptotically normal. This result will follow as a special case by letting $\hat{\beta} = \sum_{i=1}^n m(z_i,\beta_0,\hat{h}(v_i,\beta_0))/n$.

Because of the importance of asymptotic normality of $\sum_{i=1}^n m(z_i,\hat{h}(v_i))/\sqrt{n}$ (for $m(z,h) = m(z,\beta_0,h)$), and because this function is the source of the correction terms, it is useful to discuss this result first and organize the

discussion around a few high-level conditions. The pathwise derivative calculation is very useful in formulating these conditions, because it gives the form of a remainder term that should converge in probability to zero, implying asymptotic normality. Let $\alpha_j(z)$ be the solution to equation (3.5), $(j = 1, \ldots, J)$, and $\alpha(z) = \sum_{j=1}^{J}\alpha_j(z)$. Then, from the form of equation (3.7) one would expect that the following remainder term R_n should converge in probability to zero:

$$R_n = \sum_{i=1}^{n} m(z_i,\hat{h}(v_i))/\sqrt{n} - \sum_{i=1}^{n} u_i/\sqrt{n}, \quad u_i = m(z_i,h(v_i)) + \alpha(z_i) - E[\alpha(z)].$$

If $R_n \xrightarrow{p} 0$, then asymptotic normality of $\sum_{i=1}^{n} m(z_i,\hat{h}(v_i))/\sqrt{n}$ will follow from the central limit theorem applied to $\sum_{i=1}^{n} u_i/\sqrt{n}$.

To give conditions for R_n to be small it is helpful to decompose this remainder term. For $M(z) = \partial m(z,h)/\partial h|_{h=h(v)}$, let $M_j(z)$ denote the j^{th} column of $M(z)$ and

$$R_n^1 = \sum_{i=1}^{n}\{m(z_i,\hat{h}(v_i)) - m(z_i,h(v_i)) - M(z_i)[\hat{h}(v_i) - h(v_i)]\}/\sqrt{n}.$$

$$R_{nj}^2 = \sum_{i=1}^{n}\{M_j(z_i)[\hat{h}_j(v_i) - h_j(v_i)] - \alpha_j(z_i) + E[\alpha_j(z)]\}/\sqrt{n}.$$

Note that $R_n = R_n^1 + \sum_{j=1}^{J} R_{nj}^2$, so that $R_n \xrightarrow{p} 0$ if each of the following conditions is satisfied.

Asymptotic Linearity: $R_n^1 \xrightarrow{p} 0$.

Asymptotic Differentiability: $R_{nj}^2 \xrightarrow{p} 0, \quad (j = 1, \ldots, J)$.

Asymptotic linearity is similar to a condition formulated in Hardle and Stoker (1989), and will follow from a Taylor expansion and a sample mean-square convergence rate for $\hat{h}$ of slightly faster than $n^{-1/4}$, as discussed below. Asymptotic differentiability is a deeper, more important

condition. It can be shown to hold if $\hat{h}_j(v)$ is a kernel estimator of $\int a(y)f(y,x)dy$ using U-statistic projection results and higher-order bias reducing kernels, as in Powell, Stock, and Stoker (1989) and Robinson (1988), or if $\hat{h}_j(v)$ is a series estimator using properties of sample projections and series approximations, as in Newey (1990b) and Section 6. It is also possible to further decompose the asymptotic differentiability remainder in a way that allows application of Andrews (1990b) stochastic equicontinuity results. Let

$$R_{nj}^{21} = \sum_{i=1}^{n}\{M_j(z_i)[\hat{h}_j(v_i) - h_j(v_i)] - \int M_j(z)[\hat{h}_j(v) - h_j(v)]dF(z)\}/\sqrt{n}.$$

$$R_{nj}^{22} = \sqrt{n}\{\int M_j(z)[\hat{h}_j(v) - h_j(v)]dF(z) - \sum_{i=1}^{n}\{\alpha_j(z_i) - E[\alpha_j(z)]\}/n\}.$$

Note that $R_{nj}^{2} = R_{nj}^{21} + R_{nj}^{22}$, so that $R_{nj}^{2} \xrightarrow{p} 0$ if each of the following conditions is satisfied.

Stochastic Equicontinuity: $R_{nj}^{21} \xrightarrow{p} 0$.

Functional Convergence: $R_{nj}^{22} \xrightarrow{p} 0$.

Conditions for stochastic equicontinuity are given below. Functional convergence is specific to the form of $h(v)$. One interesting result is that if $E[M_j(z)|v] = 0$, then functional convergence holds trivially for $\alpha_j(z) = 0$. Thus, asymptotic linearity and stochastic equicontinuity are regularity conditions for Proposition 3, as further discussed below. When $\alpha_j(z)$ is not zero, functional convergence may follow from asymptotic normality of mean-square continuous linear functionals of $h(v)$, since functional convergence is only slightly stronger than asymptotic normality of $\sqrt{n}\{\int E[M_j(z)|v][\hat{h}_j(v) - h_j(v)]dF(z)$.

Some of these high level conditions will be consequences of more primitive hypotheses. The first of these limits the dependence between

observations that are far apart.

Assumption 5.1: z_i is strictly stationary and strong (α) mixing with mixing coefficients $\alpha(t) = O(t^{-\mu})$ for $\mu > 2$.

The next condition is uniform consistency of $\hat{h}$.

Assumption 5.2: For each j and the support $\mathcal{V}_j$ of v_j, $\sup_{v_j \in \mathcal{V}_j} |\hat{h}_j(v_j) - h_j(v_j)| \xrightarrow{p} 0$.

Primitive conditions for this and the other assumptions about $\hat{h}$ are given in Section 6. The following pair of hypotheses are more primitive conditions for Asymptotic Linearity. The first imposes smoothness conditions on m. For a random variable Y let $|Y|_s = (E[|Y|^s])^{1/s}$, and for any $\epsilon > 0$ let $\mathcal{H}(v,\epsilon) = \{h: \|h - h(v)\| < \epsilon\}$.

Assumption 5.3: $|m(z,\beta_0,h(v))|_{s'}$ is finite for some $s' > 2\mu/(\mu-1)$ and there is a neighborhood $\mathcal{N}$ of β_0, $\epsilon > 0$, $b_{jk}(z) \geq 1$, ℓ_{jk}, $(1 \leq j+k \leq 2)$, $\ell_{01} \geq 2$, $\ell_{10} \geq 2$ such that with probability one $m(z,\beta,h)$ is twice continuously differentiable on $\mathcal{N} \times \mathcal{H}(v,\epsilon)$, $\sup_{\beta \in \mathcal{N}, h \in \mathcal{H}(v,\epsilon)} \|\partial^{j+k} m(z,\beta,h)/\partial\beta^j \partial h^k\| \leq b_{jk}(z)$, $|b_{jk}(z)|_{\ell_{jk}} < \infty$.

The next hypothesis imposes a convergence rate on $\hat{h}$.

Assumption 5.4: i) for some $\hbar$, for each j, $\sum_{i=1}^n |\hat{h}_j(v_i) - h_j(v_i)|^2/n = o_p(n^{\hbar})$; ii) either m is linear in h or $\hbar \leq -\ell_{02}^{-}(1/2)$.

Assumption 5.4 is stated in terms of the sample L_2 norm rather than a more general norm because the literature on convergence rates of nonparametric estimators seems to give the sharpest results for this norm.

Lemma 5.1: If Assumptions 5.2 - 5.4 are satisfied then Asymptotic Linearity is satisfied.

The following condition is sufficient for stochastic equicontinuity.

Assumption 5.5 $|M(z)|_{s'} < \infty$ for $s' > 2\mu/(\mu-2)$ and for each j, there is a set $\mathcal{V}_j$ such that $E[1(v_j \in \mathcal{V}_j)\|M_j(z)\|] = 0$ and either; a) $\mathcal{V}_j$ is a singleton, or; b) $\mathcal{V}_j$ is convex with $\text{Prob}(\text{Boundary}(\mathcal{V}_j)) = 0$ and there is a positive integer $d_j > \dim(v_j)/2$ such that $h_j(v)$ is continuously differentiable, with bounded derivatives, to order d_j on $\mathcal{V}_j$ and for all $|\lambda| \leq d_j$, $\sup_{v_j \in \mathcal{V}_j} |D^\lambda \hat{h}(v_j) - D^\lambda h(v_j)| \xrightarrow{p} 0$.

Lemma 5.2: If Assumptions 5.1 and 5.5 are satisfied, then Stochastic Equicontinuity is satisfied

Although the main focus here is asymptotic distribution theory, for completeness it is appropriate to give a consistency result. The next hypothesis imposes identification and regularity conditions for consistency. Let $\rho: \mathbb{R}^+ \to \mathbb{R}^+$ be continuous at zero, with $\rho(0) = 0$.

Assumption 5.6: $E[m(z,\beta,h(v))] = 0$ has a unique solution at β_0, $\hat{W} \xrightarrow{p} W$, W is positive definite, and either a) $m(z,\beta,\hat{h}(v))$ is convex in β with probability one, for each $\beta \in \mathcal{B}$, $E[\|m(z,\beta,h(v))\|] < \infty$, there is $b(z)$ and such that $E[b(z)] < \infty$ and $\sup_{h \in \mathcal{H}(v,\epsilon)} \|m(z,\beta,h) - m(z,\beta,h(v))\| \leq b(z)\rho(\epsilon)$, or; b) $\mathcal{B}$ is compact, $m(z,\beta,h(v))$ is continuous in β, there is $b(z)$ and $\rho(\epsilon)$ continuous at zero such that $E[b(z)] < \infty$, $\sup_{\beta \in \mathcal{B}} \|m(z,\beta,h(v))\| \leq b(z)$, $\sup_{\beta \in \mathcal{B}, h \in \mathcal{H}(v,\epsilon)} \|m(z,\beta,h) - m(z,\beta,h(v))\| \leq b(z)\epsilon^p$.

Theorem 5.3: If Assumptions 5.1 - 5.2 and 5.6 are satisfied then $\hat{\beta} \xrightarrow{p} \beta_0$.

Next, regularity conditions for Proposition 3 are given. Let $m_i = m(z_i, \beta_0, h(v_i))$.

Theorem 5.4: If Assumptions 5.1-5.6 are satisfied and for each j either $E[M_j(z)|v_j] = 0$ or, more generally $h_j(v_j)$ and $\hat{h}_j(v_j)$ are elements of a set $\mathcal{H}_j$ such that $E[M_j(z)\tilde{h}_j(v_j)] = 0$ for all $\tilde{h}_j \in \mathcal{H}_j$, then for $\Omega = E[m_i m_i'] + \sum_{\ell=1}^{\infty} E[m_i m_{i+\ell}' + m_{i+\ell} m_i']$.

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V), \quad V = (D'WD)^{-1}D'W\Omega WD(D'WD)^{-1}.$$

This theorem shows that Andrews (1990a) "independence" hypothesis, that estimation of $h(v)$ does not affect the limiting distribution of $\hat{\beta}$, is a consequence of orthogonality of $M(z)$ with the set of possible $h(v)$.

The next asymptotic normality result allows for a nonzero correction term. Let $\Omega_\ell = E[u_i u_{i+\ell}']$, $\Omega = \Omega_0 + \sum_{\ell=1}^{\infty}(\Omega_\ell + \Omega_\ell')$.

Theorem 5.5: If for some $s' > 2\mu/(\mu-1)$, $|\alpha_j(z)|_{s'}$, is finite, $(j = 1, \ldots, J)$, Assumptions 5.1 - 5.4, 5.6, and Asymptotic Differentiability are satisfied, then $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V)$, $V = (D'WD)^{-1}D'W\Omega WD(D'WD)^{-1}$.

A consistent estimator $\hat{\Omega}$ of Ω is required to form a consistent estimator of the asymptotic variance of β. Such $\hat{\Omega}$ can be formed from estimates $\hat{u}_i$ of u_i. Let $\hat{\alpha}_{ji}$ denote estimates of $\alpha_j(z_i)$ and

$$\hat{m}_i = m_i(z_i, \hat{\beta}, \hat{h}(v_i)), \quad \hat{u}_i = \hat{m}_i + \sum_{j=1}^{J}[\hat{\alpha}_{ji} - \sum_{i=1}^{n}\hat{\alpha}_{ji}/n], \quad \hat{\Omega}_\ell = \sum_{i=1}^{n-\ell}\hat{u}_i\hat{u}_{i+\ell}'/n.$$

If u_i is not autocorrelated then $\hat{\Omega} = \hat{\Omega}_0$ will be an appropriate estimator. When u_i may be autocorrelated, consider a weighted autocovariance estimator like that in Newey and West (1987), with

$$\hat{\Omega} = \hat{\Omega}_0 + \sum_{\ell=1}^{L} w(\ell, L)[\hat{\Omega}_\ell + \hat{\Omega}_\ell'],$$

where $w(\ell,L)$ is a weight such that $\hat{\Omega}$ is positive semi-definite, such as $w(\ell,L) = 1 - \ell/(L+1)$. Here L can depend on the data, as is important for applications. Given this estimator of $\hat{\Omega}$, an estimator of the asymptotic covariance matrix of $\hat{\beta}$ can be formed in the usual way, as

$$\hat{V} = (\hat{D}'\hat{W}\hat{D})^{-1}\hat{D}'\hat{W}\hat{\Omega}\hat{W}\hat{D}(\hat{D}'\hat{W}\hat{D})^{-1}.$$

Theorem 5.5: Suppose that Assumptions 5.1 and 5.3 are satisfied with $\ell_{01} = \ell_{10} = 2$, *there is* $s > 4\mu/(\mu-2)$ *such that* $|m_i|_s$ *and* $|\alpha_j(z_i)|_s$ *are finite for each* j, $w(\ell,L)$ *is bounded uniformly in* ℓ *and* L *and* $\lim_{L\longrightarrow\infty} w(\ell,L) = 1$ *for each* ℓ, $\|\hat{\beta}-\beta_0\| = O_p(1/\sqrt{n})$, *there is* $\epsilon_n = o(1)$ *such that* $1/n = O(\epsilon_n^2)$, $|\hat{h}_j(v)-h_j(v)|_\infty = O_p(\epsilon_n)$, $\sum_{i=1}^{n}\|\hat{\alpha}_{ji}-\alpha_j(z_i)\|^2/n = O_p(\epsilon_n^2)$ *and either a)* $\Omega_\ell = 0$, $\ell \geq 1$, *and* $\hat{\Omega} = \hat{\Omega}_0$; *or b)* $L \xrightarrow{p} \infty$, *and* $L = o_p(\epsilon_n^{-1})$. *Then,* $\hat{V} \xrightarrow{p} V$.

As usual for minimum distance estimators, the asymptotic variance depends on W, and an optimal (asymptotic variance minimizing) choice of W is Ω^{-1} when Ω is nonsingular. The estimator $\hat{\Omega}$ can be used to form a feasible version of the optimal minimum distance estimator, by using $\hat{W} = \hat{\Omega}^{-1}$ in equation (5.1). The resulting estimator will be an optimal estimator that adjusts for the presence of first-stage, plug-in estimators in the moment functions, similarly to the estimator of Hansen (1985). For this choice of $\hat{W}$, $(\hat{D}'\hat{\Omega}^{-1}\hat{D})^{-1}$ will be a consistent estimator of the asymptotic variance of $\hat{\beta}$.

6. Series Estimation of Projection Functionals.

This Section develops regularity conditions for linear functionals of power series estimators of projections. Power series are considered because they are computationally convenient and the most complete convergence rate results seem to be available for them. Although in some contexts power series are thought to be inferior to other approximating functions, because of their "roly-poly" behavior and global sensitivity of best uniform approximations to singularities, these considerations may not be as important here, where the projection estimator is a nuisance function. Under Assumption 4.1, $\hat{\beta}$ depends on the series approximation essentially only through a weighted average, where these problems with power series seem not to be so important. An example is provided by the Monte-Carlo results of Newey (1988a), where a semiparametric power series estimator performs extremely well relative to a kernel estimator.

Here, the domain $\mathcal{G}$ of the projection will be assumed to take the form in equation (4.1). The conditions to follow will depend on the maximum across $\ell \leq L$ of the dimension of $\tilde{x}_\ell$, which will be denoted by r. A power series estimator of the projection can be obtained from a regression of y on a truncated power series with elements restricted to lie in $\mathcal{G}$, analogous to Stone's (1985) spline estimator. Let λ denote a vector of nonnegative integers as before, and let $x^\lambda = \Pi_{\ell=1}^r x_\ell^{\lambda_\ell}$. For a sequence $(\lambda(k))_{k=1}^\infty$ of distinct such vectors, a power series is

$$(6.1) \qquad p_k(x) = \begin{cases} \tilde{x}_{L+1,k}, & k = 1, \ldots, \dim(\tilde{x}_{L+1}) \\ x^{\lambda(k-s)}, & k = \dim(\tilde{x}_{L+1})+1, \ldots \end{cases},$$

It will be assumed that $\{x^{\lambda(k)}\}_{k>\dim(\tilde{x}_{L+1})}$ consists of all multivariate

powers of each $\tilde{x}_\ell$ for $\ell \leq L$, and no more, ordered so that $|\lambda(k)|$ is monotonic increasing. This assumption imposes the essential restriction that each $p_k(x)$ belongs to $\mathcal{G}$, and the spanning condition that the sequence includes all such terms.

The estimator of the projection considered here is that obtained from the least squares regression of y on $\hat{K}$ terms, where $\hat{K}$ is allowed to depend on the data. For $y \equiv (y_1,\ldots,y_n)'$, $p^K(x) = (p_1(x),\ldots,p_K(x))'$, and $p^K \equiv [p^K(x_1),\ldots,p^K(x_n)]$, the estimator of $g(x)$ is

$$(6.2) \qquad \hat{g}(x) = p^{\hat{K}}(x)'\hat{\pi}, \quad \hat{\pi} = \hat{\Sigma}^{-}p^{\hat{K}\prime}y, \quad \hat{\Sigma} = p^{\hat{K}\prime}p^{\hat{K}}/n$$

where $(\cdot)^-$ denotes a generalized inverse. Under the conditions to follow, $p^{\hat{K}\prime}p^{\hat{K}}$ will be nonsingular with probability approaching one, so that the choice of generalized inverse does not matter, asymptotically.

A data based $\hat{K}$ is essential for making operational the nonparametric properties of series estimators, allowing the estimator to adjust to conditions in particular applications. It would also be interesting to know how to best choose $\hat{K}$ in the current context, but this question is outside the scope of this paper.

For computational purposes it may be useful to replace $p^K(x)$ with nonsingular linear transformation to polynomials that are orthogonal with respect to some distribution, since these may have less of a multicollinearity problem than power series are known to have. Of course, this replacement will not affect the estimator. Also, note that the elements of each $\tilde{x}_\ell$ may be smooth, bounded transformations (e.g. the logit distribution function) of "original" variables, which may help to limit the sensitivity of the estimator to outliers. In the Monte Carlo example of Newey (1988a), such a transformation lead to reduced sensitivity to the choice of K.

The function $h(v)$ that appears in the moments will be taken to be a linear function $A(v,g)$ of the projection g, as analyzed in Section 4, and $h(v)$ will be estimated by replacing g by $\hat{g}$ in the function. By linearity of A in g, the resulting estimator takes the form

$$(6.3) \qquad \hat{h}(v) = A(v,\hat{g}) = A(v)'\hat{\pi}, \quad A(v) = (A(v,p_1),\ldots,A(v,p_{\hat{K}}))'.$$

Note that this estimator requires that $A(v,g)$ have an explicit form that does not depend on the true data-generation process.

An estimator of the correction term is required for estimation of the asymptotic variance of $\hat{\beta}$. Under Assumption 4.1, such an estimator can be constructed in a straightforward way. Let

$$(6.4) \qquad \hat{\alpha}(z) = \hat{\Psi}'\hat{\Sigma}^{-}p^{\hat{K}}(x)[y - \hat{g}(x)], \quad \hat{\Psi} = \sum_{i=1}^{n}[\partial m(z_i,\hat{\beta},\hat{h}(v_i))/\partial h]A(v_i)/n.$$

By Assumption 4.1 $\hat{\Psi}$ will be an estimator of $\int\delta(x)p^{\hat{K}}(x)dF(x)$, so that $\hat{\Psi}'\hat{\Sigma}^{-}p^{K}(x)$ is an estimator of the regression of $\delta(x)$ on $p^{\hat{K}}(x)$, which will approximate $\delta(x)$ for large $\hat{K}$ and n. Alternatively, $\hat{\alpha}(z)$ can be viewed as the estimator of the correction term obtained by treating $\hat{g}(x)$ as if it were a parametric regression, with K fixed. This procedure results in a consistent estimator of the correction term because it accounts properly for its variance, while bias from the series approximation will be small because of smoothness restrictions on $g(x)$ and $\delta(x)$ imposed below.

The following conditions are needed to apply the results of Newey (1991). Let $\tilde{X}$ denote the vector consisting of the union of all distinct variables appearing in $\tilde{x}_\ell$, $(\ell \leq L)$, and let $\tilde{\mathcal{G}} = \{\sum_{\ell=1}^{L}g_\ell(\tilde{x}_\ell) : E[g_\ell(\tilde{x}_\ell)^2] < \infty\}$. The first condition is sufficient for $\tilde{\mathcal{G}}$ to be closed.

Assumption 6.0: i) For each $\tilde{x}_\ell$, $(\ell = 1, \ldots, L)$, if $\tilde{x}$ is a subvector of $\tilde{x}_\ell$ then $\tilde{x} = \tilde{x}_{\ell'}$ for some $\ell' \leq L$; ii) There exists a constant $c > 1$ such that for each ℓ, with the partitioning $\tilde{X} = (\tilde{x}_\ell', \tilde{x}_\ell^{c\prime})'$, for all $a(q) \geq 0$, $c\int a(q)d[F(\tilde{x}_\ell)\cdot F(\tilde{x}_\ell^c)] \geq E[a(q)] \geq c^{-1}\int a(q)d[F(\tilde{x}_\ell)\cdot F(\tilde{x}_\ell^c)]$; iii) Either $\eta = 0$ (i.e. $\tilde{x}_{L+1}$ is not present) or $\tilde{x}_{L+1}$ is bounded and for the closure $\bar{\mathcal{G}}$ of $\mathcal{G}$, $E[\{\tilde{x}_{L+1}-\Pi(\tilde{x}_{L+1}|\bar{\mathcal{G}})\}\{\tilde{x}_{L+1}-\Pi(\tilde{x}_{L+1}|\bar{\mathcal{G}})\}']$ is nonsingular.

The next condition requires that the support of $\tilde{X}$ be a box and places a lower bound on its distribution.

Assumption 6.1: There are finite $\tilde{X}_{ju} > \tilde{X}_{jb}$, $\nu_j \geq 0$, $(j = 1, \ldots, \dim(\tilde{X}))$ such that the support of $\tilde{X}$ is $\prod_{j=1}^{\dim(\tilde{X})}[\tilde{X}_{ju}, \tilde{X}_{jb}]$ and the distribution of $\tilde{X}$ has absolutely continuous component with density bounded below by $C\prod_{j=1}^{r}[(\tilde{X}_{ju}-\tilde{X}_j)(\tilde{X}_j-\tilde{X}_{jb})]^{\nu}$ on the support.

The nonsingularity condition is a normalization, unless η_0 is a parameter of interest, where it is an identification assumption for η_0. Let $\varepsilon = y-g(x)$.

Assumption 6.2: $|\varepsilon|_{s'}$, is finite for $s' \geq 2$ and $E[\varepsilon^2|x]$ is bounded.

The bounded second conditional moment assumption is quite common in the literature (e.g. Stone, 1985), and simplifies the regularity conditions.

Assumption 6.3: Either a) z_i is uniform mixing with mixing coefficients $\phi(t) = O(t^{-\mu})$, $(t = 1, 2, \ldots)$, for $\mu > 2$ or; b) there exists $c(t)$ such that $|E[\varepsilon_i\varepsilon_{i+t}|q_i, q_{i+t}]| \leq c(t)$ and $\sum_{t=1}^{\infty}c(t) < \infty$.

This assumption is restrictive, but covers many cases of interest, including independent observations and dynamic nonparametric regression with $g(x_i) = E[y_i|x_i, y_{i-1}, x_{i-1}, y_{i-2}, \ldots]$. The next condition restricts the amount of variation allowed in the choice of number of terms K.

Assumption 6.4: $K = \hat{K}$ such that with probability approaching one, $\underline{K} \leq \hat{K} \leq \bar{K}$ where $\underline{K} = \underline{K}(n)$ and $\bar{K} = \bar{K}(n)$ are sequences of constants, and there is $\epsilon > 0$ such that $\underline{K}(n) \geq n^{\epsilon}$ for all n large enough.

The bounds $\underline{K}$ and $\bar{K}$ control the bias and variance, respectively, of $\hat{g}$. The next set of conditions impose smoothness assumptions used to control the bias of the estimator.

Assumption 6.5: $g_{\ell 0}(\tilde{x}_\ell)$ is continuously differentiable to order d, $(\ell \leq L)$.

Two results will be given, because the conditions are weaker and simpler in the special case where $h(v) = g(x)$, meriting its separate treatment. For any nonnegative integer d let

$$\zeta_d(K) = K^{.5+\nu+2d}. \tag{6.5}$$

The covariance matrix of $\hat{\beta}$ can be estimated by the procedure discussed in Section 5, using the estimators of $\alpha_j(z)$ given above. The asymptotic distribution results will include consistency of this variance estimator.

Theorem 6.1: Suppose that Assumptions 5.1, 5.3, 5.6, 6.0-6.5 are satisfied, and for each j, i) $s > 4\mu/(\mu-2)$, $E[\|M_j(z)\|^{s'}]$ is finite for some $s' > 4\mu/(\mu-2)$ and $E[\|M_j(z)-\delta_j(x)\|^2|x]$ is bounded; ii) $\delta_j(x)$ is continuously differentiable to order d_δ on x; iii) each of the following converge to zero: $\bar{K}^2\zeta_0(\bar{K})^4/n$, $\bar{K}^{1/2}\zeta_0(\bar{K})\underline{K}^{-d_\delta/n}$, $\bar{K}^{1/2}\zeta_0(\bar{K})\underline{K}^{-d/n}$, $\sqrt{n}\underline{K}^{-(d+d_\delta)/n}$; iv) either $m(z,\beta_0,h)$ is linear in h or $\bar{K}/\sqrt{n} + \sqrt{n}\underline{K}^{-2d/n} = o(n^{1/\ell_{02}})$; v) $\epsilon_n = n^{1/s}\bar{K}^{1/2}\zeta_0(\bar{K})(\bar{K}^{1/2}/\sqrt{n} + \underline{K}^{-d/n}) + n^{1/s}\underline{K}^{-d_\delta/n} = o(1)$ and either a) $\Omega_\ell = 0$, $\ell \geq 1$, $\hat{\Omega} = \hat{\Omega}_0$; or b) $L \xrightarrow{p} \infty$, and $L = o_p(\epsilon_n^{-1})$. Then $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V)$ and $\hat{V} \xrightarrow{p} V$.

The upper bound on the rate of growth for the number of terms in the series expansion is $n^{1/4}$ when the density of $\tilde{X}$ is bounded away from zero ($\nu = 0$), which is less than the $n^{1/2}$ rate derived in Newey (1990b). This result also requires existence of derivatives of both $g(x)$ and $\delta(x)$ up to order more than the largest dimension of an additive component, as in Newey (1990b).

To obtain asymptotic normality in the more general case, where $h(v) = A(v,g)$ is some other linear function of g, it is useful to impose a continuity condition on $A(v,g)$ as a function of g. Let $\mathcal{V}$ denote the support of v, and denote supremum Sobolev norms by

$$\|h(v)\|_d = \sup_{|\lambda|\leq d, v\in\mathcal{V}} |D^\lambda h(v)|, \quad \|g(x)\|_d = \sup_{|\lambda|\leq d, \tilde{X}} |D^\lambda \tilde{g}(\tilde{X})|,$$

Assumption 6.6: There is a constant C and an integer Δ such that $\|A(v,g)\|_0 \leq C\|g(x)\|_\Delta$.

This Assumption will imply that the bias from approximating the function $h_0(v)$ by a linear combination of $A(v)$ is bounded by the bias of approximating $g(x)$ and its derivatives to order Δ by a linear combination of $p^K(x)$. Unfortunately, for multivariate functions, a literature search has not yet revealed bias bounds for approximating a functions and derivatives by power series, except under part b) of the following condition.

Assumption 6.7: Either a) $n = 1$ or $\Delta = 0$ or; b) for each ℓ, $g_{\ell 0}(\tilde{x}_\ell)$ is continuously differentiable to all orders, and there is a constant C such that $\sup_{\tilde{X}\in\mathcal{X}} |D^\lambda g_\ell(\tilde{x}_\ell)| \leq C^{|\lambda|}$ for all λ.

Condition b) implies an approximation rate for $g(x)$ and its derivatives that is faster than $K^{-\alpha}$ for any α.

When $\hat{K}$ is random, it is useful to also have an approximation rate for

$\delta(x)$. In order that the results would apply to many cases, with more general approximation results that one anticipates will appear eventually, the rest of the results of this section will be stated in terms of

$$\epsilon_\delta(K) = \min_\pi |\delta(x)-p^K(x)'\pi|_2.$$

Let $\mathcal{K} = \mathcal{K}(n) = [\underline{K}, \bar{K}]$.

Theorem 6.2: Suppose that Assumptions 4.1, 5.1, 5.3, 5.6, 6.0-6.7 are satisfied and let $\alpha_0 = d/r$ *and* $\alpha = (d/r)-\Delta$ *under Assumption 6.7 a) and* $\alpha_0 = \alpha = +\infty$ *under Assumption 6.7 b). Also suppose that for each* j, $|m(z,\beta_0,h(v))|_s$, and $|\delta_j(x)[y_j-g_j(x)|_s$, are *finite for* $s' > 4\mu/(\mu-2)$, *and i)* $\sum_{\mathcal{K}} K\zeta_\Delta(K)^2K^{-2\alpha} \to 0$ *and* $\sqrt{n}\underline{K}^{-\alpha}o\epsilon_\delta(\underline{K}) \to 0$; *ii) Either a)* z_i *is uniform mixing or b)* $\sum_{\mathcal{K}} K^2\zeta_0(K)^4K^{-2\alpha}o \to 0$; *iii) Either a)* $m(z,h)$ *is linear in* h *and* $\epsilon_n = \bar{K}^{1/2}\zeta_\Delta(\bar{K})\underline{K}^{-\alpha} + n^{1/s}[\bar{K}\zeta_\Delta(\bar{K}) + \bar{K}^{3/2}\zeta_0(\bar{K})^2]/\sqrt{n} + [\sum_{\mathcal{K}}\epsilon_\delta(K)^2]^{1/2} \to 0$, *or b)* $\sqrt{n}\bar{K}\zeta_\Delta(\bar{K})^2[\bar{K}/n+\underline{K}^{-2\alpha}] = o(1)$ *and* $\epsilon_n = n^{1/s}\bar{K}^{3/2}\zeta_\Delta(\bar{K})[\bar{K}^{1/2}/\sqrt{n} + \underline{K}^{-\alpha}] + n^{1/s}\bar{K}^{3/2}\zeta_0(\bar{K})^2/\sqrt{n} + [\sum_{\mathcal{K}}\epsilon_\delta(K)^2]^{1/2} \to 0$; *iv) either a)* $\Omega_\ell = 0$, $\ell \geq 1$, $\hat{\Omega} = \hat{\Omega}_0$; *or b)* $L \xrightarrow{p} \infty$, *and* $L = o_p(\epsilon_n^{-1})$. *Then* $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V)$ *and* $\hat{V} \xrightarrow{p} V$.

The smallest upper bound on the number of terms allowed by this theorem is $n^{1/5}$, for $\nu = 0$ and $s = \infty$, a rate also derived in Newey (1991). This result also requires existence of derivatives of $g(x)$ of order more than 3/2 the maximum dimension of the additive components, but imposes weak smoothness restrictions on $\delta(x)$.

The requirement that the bias go to zero faster than $1/\sqrt{n}$, as needed for $\sqrt{n}$-consistency, is that $\sqrt{n}\underline{K}^{-\alpha}o\epsilon_\delta(\underline{K})$ converge to zero. This term is the product of $\sqrt{n}$, the approximation rate for $g(x)$ (i.e. the bias from estimating $g(x)$ by truncated series), and the approximation rate for $\delta(x)$.

Consequently, "undersmoothing" is not required for asymptotic normality with "plug-in" series estimators, as it is for some kernel estimators, because the bias from estimating $h(v)$ does not have to go to zero faster than $1/\sqrt{n}$. This result is a consequence of the usual orthogonality property of mean-square projections. Let $g_K(x)$ and $\delta_K(x)$ be the population regression on $p^K(x)$ of $g(x)$ (or y) and $\delta(x)$, respectively, and $h_K(v) = A(v,g_K)$ the corresponding value of $h_K(v)$. Using Assumption 4.1, the population first order bias term, analogous to that considered by Stoker (1990) for kernels, is

$$(6.6) \qquad E[M(z)\{h_K(v)-h(v)\}] = E[\delta(x)\{g_K(x)-g(x)\}]$$

$$= -E[\{\delta_K(x)-\delta(x)\}\{g_K(x)-g(x)\}],$$

so that the bias term for h is equal to product of biases terms for δ and g.

Under Assumption 6.1 with $\nu = 0$, $\hat{K}$ nonrandom, and Assumption 6.7 a), these results will also apply to uniform knot spline estimators of $\hat{g}(x)$, if the definition of $\zeta_d(K)$ is changed to $\zeta_d(K) = K^{.5+d}$. More generally, the results apply to any series estimator satisfying Assumptions 3.1-3.8 in Newey (1991), although these do not allow for Fourier series estimators.

7. Examples

This Section gives primitive regularity conditions for the validity of the examples of Section 4, and one or two additional examples. To save space, this Section has a special format, where each subsection gives an estimator an estimator of its asymptotic covariance and a result on $\sqrt{n}$-consistency and

asymptotic normality of the estimator and consistency of the estimated variance, with discussion of the results reserved until the end.

To give more specific results on the rate at which the number of terms can grow it will be useful to impose the following Assumption.

Assumption 7.1: $\underline{K}(n) = n^{\gamma}$ and $\overline{K}(n) = n^{\Gamma}$ for some $\Gamma > \gamma > 0$.

7.1 Semiparametric Random Effects for Binary Panel Data

$$(\hat{\beta}_1', \hat{\sigma}_2)' = (\sum_{i=1}^n \hat{A}_i \hat{A}_i')^{-1} \sum_{i=1}^n \hat{A}_i \Phi^{-1}(\hat{h}_1(x_i)), \quad \hat{\sigma}_1 = 1; \quad x_i = (x_{i1}', x_{i2}')',$$

$$\hat{h}_t(x) = \hat{E}[y_t | x], \quad (t = 1, 2), \quad \hat{A}_i = I_i(x_{i1}' - x_{i2}', \Phi^{-1}(\hat{h}_2(x_i)))',$$

$$I_i = 1(0 < \hat{h}_1(x_i) < 1 \text{ and } 0 < \hat{h}_2(x_i) < 1), \quad .$$

$$\hat{V} = n(\sum_{i=1}^n \hat{A}_i \hat{A}_i')^{-1} \sum_{i=1}^n \hat{u}_i \hat{u}_i' (\sum_{i=1}^n \hat{A}_i \hat{A}_i')^{-1}; \quad \hat{u}_i = \hat{A}_i \{\Phi^{-1}(\hat{h}_1(x_i)) - \hat{A}_i'(\hat{\beta}_1', \hat{\sigma}_2)'\}$$

$$+ \sum_{t=1}^2 (-1)^{t-1} \hat{\sigma}_t [\sum_{j=1}^n \phi(\Phi^{-1}(\hat{h}_t(x_j)))^{-1} p^{\hat{K}}(x_j)/n] \hat{\Sigma}^- p^{\hat{K}}(x_i) [\hat{y}_{ti} - \hat{h}_t(x_i)]\},$$

Theorem 7.1: Suppose that i) z_i are i.i.d.; ii) Assumptions 6.0 and 6.1 are satisfied for $x = (x_1, x_2)$. iii) $\rho(x)$ has continuous derivatives of up to order d on $\tilde{X}$; iv) $E[(x_1 - x_2, \rho(x))(x_1 - x_2, \rho(x))']$ is nonsingular; vi) Assumption 7.1 is satisfied with $\Gamma < 1/4$, $\gamma > \max\{2\Gamma k/d, 2\Gamma k/d_\delta, k/(d+d_\delta)\}$. Then $\sqrt{n}[(\hat{\beta}', \hat{\sigma}_2) - (\beta', \sigma_2)] \xrightarrow{d} N(0, V)$ and $\hat{V} \xrightarrow{p} V$.

7.2 Nonparametric Consumer Surplus

$$\hat{\beta} = \sum_{i=1}^n \int_{\mathcal{A}} \hat{g}(x_1, x_{2i}) dx_1 / n,$$

$$\hat{V} = \sum_{i=1}^n \hat{u}_i^2 / n, \quad \hat{u}_i = \int_{\mathcal{A}} \hat{g}(x_1, x_{2i}) dx_1 - \hat{\beta}$$

$$+ \{\sum_{j=1}^n \int_{\mathcal{A}} p^{\hat{K}}(x_1, x_{2i}) dx_1 / n\}' \hat{\Sigma}^{-1} p^{\hat{K}}(x_i) [y_i - \hat{g}(x_i)].$$

Theorem 7.2: Suppose that i) z_i *are i.i.d.; ii) Assumptions 6.0 - 6.2, and 6.5 are satisfied for* $\nu = 0$, $s > 4$. *iii)* $x = (x_1, x_2)$ *is absolutely continuous with respect to the product of the uniform density on* $\mathcal{A}$ *and the distribution of* x_2, *with bounded density* $\tilde{f}(x)$ *and either a)* $\mathcal{G} = \{\tilde{g}_1(x_1, x_2^1) + x^{2\prime}\eta\}$, *and* $f(x_1, x_2^1)^{-1} f(x_2^1)$ *and* $E[x^2|x^1]$ *are continuously differentiable to order* d_δ *on the support of* x^1 *or b)* $\Pi(\tilde{f}(x)|\mathcal{G})$ *is continuously differentiable in* $\tilde{X}$ *to order* d_δ *on the support of x. iv) Assumption 7.1 is satisfied with* $d > 3r/2$, $\Gamma < (s-2)/5s$, $\gamma > r/[2(d+d_\delta)]$, $\gamma > \Gamma r/d$. *Then* $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, V)$ *and* $\hat{V} \xrightarrow{p} V$.

7.3 Nonparametric Prediction

$$\hat{\beta} = \sum_{i=1}^{n} [\hat{g}(v_i) - \hat{g}(x_i)]/n,$$

$$\hat{V} = \sum_{i=1}^{n} \hat{u}_i^2/n, \quad \hat{u}_i = \hat{g}(v_i) - \hat{g}(x_i) - \hat{\beta}$$

$$+ \{\sum_{j=1}^{n} [p^{\hat{K}}(v_j) - p^{\hat{K}}(x_j)]/n\}' \hat{\Sigma}^{-1} p^{\hat{K}}(x_i)[y_i - \hat{g}(x_i)].$$

Theorem 7.3: Suppose that i) z_i *are i.i.d.; ii) Assumptions 6.0 - 6.2, and 6.5 are satisfied for* $\nu = 0$ *and* $s > 4$, *iii)* v *is absolutely continuous with respect to* x *with bounded density* $\tilde{f}(v)$ *and either a)* $\mathcal{G} = \{\tilde{g}_1(x^1) + x^{2\prime}\eta\}$, $v = (v^1, x^2)$, *and* $f_{x^1}(x^1)^{-1} f_{v^1}(x^1)$ *and* $E[x^2|x^1]$ *are continuously differentiable to order* d_δ *on the support of* x^1 *or b)* $\Pi(\tilde{f}(x)|\mathcal{G})$ *is continuously differentiable in* $\tilde{X}$ *to order* d_δ *on the support of x. iv) Assumption 7.1 is satisfied with* $d > 3r/2$, $\Gamma < (s-2)/5s$, $\gamma > r/[2(d+d_\delta)]$, $\gamma > \Gamma r/d$. *Then* $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, V)$ *and* $\hat{V} \xrightarrow{p} V$.

7.4 Average Derivatives

$$\hat{\beta} = \sum_{i=1}^{n} m(z_i, \hat{h}(x_i))/n, \quad \hat{h}(x) = \partial \hat{g}(x)/\partial x_1$$

$$\hat{V} = \sum_{i=1}^{n} \hat{u}_i \hat{u}_i'/n, \quad \hat{u}_i = m(z_i, \hat{h}(x_i)) - \hat{\beta}$$

$$+ [\sum_{j=1}^{n} \{\partial m(z_i, \hat{h}(x_i))/\partial h\}\{\partial p^{\hat{K}}(x_j)/\partial x_1\}'/n]\hat{\Sigma}^{-} p^{\hat{K}}(x_i)[y_i - \hat{g}(x_i)].$$

Theorem 7.4: Suppose that i) Assumption 5.1 is satisfied, and $g(x_i) = E[y_i|x_i, y_{i-1}, x_{i-1}, \ldots]$; *ii)* $\hat{K} = K(n)$ *is not random; iii) Assumptions 6.0 - 6.2, 6.5, and 6.7 b) are satisfied, there is* $s' > 4\mu/(\mu-2)$ *such that* $|\alpha_j(z)|_{s'} < \infty$, $(j = 1, \ldots, J)$, x *is absolutely continuous respect to the product of Lebesgue measure on* x_1 *and the distribution of all elements of* x *other than* x_1 *with density* $f(x)$ *that is continuously differentiable in* x_1 *on the interior of a convex support,* $\partial f(x)/\partial x_1$ *zero on the boundary of the support, and* $E[\|f(x)^{-1}\partial f(x)/\partial x_1\|^2]$ *is finite. v)* $K \geq n^{\gamma}$ *for some* $\gamma > 0$ *and* $K = O(n^{\Gamma})$ *for either a)* $m(z,h)$ *linear in* h *and* $\Gamma < (s-2)/[s(7+4v)]$, *or b)* $\Gamma < (s-2)/[s(14+4v)]$. *Then* $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, V)$ *and* $\hat{V} \xrightarrow{p} V$.

7.5 Discussion

The conditions in Section 7.4 for multidimensional average derivatives are quite restrictive, but could be relaxed if better approximation rate results were available, on approximation of derivatives and unbounded functions by power series. In particular, nonrandom K results from not having an approximation rate for $\delta(x)$, which is unbounded for average derivatives. Also, one can relax these conditions substantially for weighted average derivatives, where $\beta_0 = E[w(x)\partial g(x)/\partial x]$ and the weight function is

such that $f(x)^{-1}w(x)\partial f(x)/\partial x + \partial w(x)/\partial x$ is continuously differentiable on the support of x, including allowing for random K.

The estimators for the semiparametric random effects and for nonparametric consumer surplus seem to be new. The result for nonparametric prediction is the first result on $\sqrt{n}$-consistency of an estimator, and includes the unconditional variance in the estimation of the asymptotic variance. Series estimators for average derivatives were previously suggested by Andrews (1991), although the result here includes conditions for $\sqrt{n}$-consistency and apply to times series.

Appendix A: Proofs of Theorems

Throughout this appendix C will denote a generic positive constant that can be different in different uses and o_p will denote $o_p(1)$.

Proof of Theorem 2.1: Pathwise differentiability of $\mu(F_z)$ follows immediately from Theorem 2.1 of Van der Vaart (1991), since asymptotic linearity of $\hat{\beta}$ and the Linbergh-Levy central limit theorem imply that for any $S_\theta(z) \in \mathcal{S}$, $(\sqrt{n}(\hat{\beta}-\beta_0)', \sum_{i=1}^n S_\theta(z_i)/\sqrt{n})'$ converges in distribution to $N(0, E[(\psi(z)', S_\theta(z))'(\psi(z)', S_\theta(z))])$. Furthermore, by the the final conclusion of Lemma A.1 of Van der Vaart, it follows that for any vector b, $b'(\theta^{-1}[\mu(F_{z\theta})-\mu(F_{z0})])$ converges to $b'E[\psi(z)S_\theta(z)]$, while by pathwise differentiability it follows that $b'E[\psi(z)S_\theta(z)] = b'E[d(z)S_\theta(z)]$. Since this equality must hold for any b and path, it follow by Assumption 2.1 that $E[(\psi(z)-d(z))s(z)] = 0$ for all mean-zero $s(z)$, so that choosing $s(z)$ to be any element of $\psi(z)-d(z)$, it follows that $\psi(z) = d(z)$.

Proof of Lemma 5.1: It suffices to prove the result for scalar m. Let let $\hat{h}_i \equiv \hat{h}(v_i)$, $h_i = h_0(v_i)$, and $m(z,h) = m(z,\beta,h)$. By Assumption 5.2, $\max_{i\le n}\|\hat{h}_i-h_i\| < \epsilon$ w.p.a.1. Also, a standard result implies $\max_{i\le n}\{b_{02}(z_i)\} = O_p(n^{1/\ell_{02}})$. Thus, by an expansion,

$$(A.1) \quad \|\sum_{i=1}^n m(z_i,\hat{h}_i)-m_i-M(z_i)(\hat{h}_i-h_i)/n\| \le \sum_{i=1}^n \|\partial^2 m(z_i,\bar{h}_i)/\partial h\partial h'\|\|\hat{h}_i-h_i\|^2/n$$

$$\le \max_{i\le n}\{b(z_i)\}\sum_{i=1}^n \|\hat{h}_i-h_i\|^2/n = O_p(n^{1/\ell_{02}})o_p(n^{-(1/2)-1/\ell_{02}}) = o_p(n^{-1/2}). \quad \blacksquare$$

Proof of Lemma 5.2: Suppress the J subscript. In Andrews (1990b) notation, let $W_{aTt} = v_i$, $W_{Tt} = z_i$, $\mathcal{W}_a = \mathcal{V}$, $q = d$, $\tau = 1(v\in\mathcal{V})h(v)$, $k_a = \dim(v)$, $m(w_a,\tau) = \tau(v)$, and $g(w) = M(z)$. Note that Andrews Assumption F ii) is satisfied by hypothesis. Also by hypothesis, $\tau_0 = 1(v\in\mathcal{V})h_0(v)$ has

derivatives of up to order q on W_a, and for $|\lambda| \leq q$, $\sup_{v\in W_a}|D^\lambda\hat{\tau}(v) - D^\lambda\tau_0(v)| = o_p(1)$. Let $T = \{\tau(v) : \sup_{v\in W_a}|D^\lambda\tau(v)| \leq \max_{v\in W_a}|D^\lambda\tau_0(v)| + 1$ for all λ with $|\lambda| \leq q\}$. Then $\sup_{\tau\in T}[\sum_{|\lambda|\leq q}\int_{W_a}|D^\lambda\tau(x)|^2dx]^{1/2} < \infty$, giving Andrews Assumption F iii). Also, by construction, $m(w_a,\tau) = \tau(v)$ is zero (and hence constant) outside W_a, giving Andrews Assumption F iv). Also, $s' > 2\mu/(\mu-2)$ implies $\mu > 2s'/(s'-2)$, so that Andrews Assumptions v) and vi) are satisfied. Finally, $\hat{\tau}(v) \in T$ w.p.a.1, and $\int_V[\hat{\tau}(v)-\tau(v)]^2dv \xrightarrow{p} 0$, so that the conclusion follows by Theorem II.7 of Andrews (1990b).

Proof of Theorem 5.3: Let $\tilde{m}_n(\beta) = \sum_{i=1}^n m(z_i,\beta,h_0(v))/n$, $m(\beta) = E[m(z,\beta,h_0(v))]$, $\hat{Q}(\beta) = \hat{m}_n(\beta)'\hat{W}\hat{m}_n(\beta)$, and $Q(\beta) = m(\beta)'Wm(\beta)$. Under Assumption 5.6 a), Assumption 5.2 implies that for each β, $\|\hat{m}_n(\beta)-\tilde{m}_n(\beta)\| \xrightarrow{p} 0$, while by z_i ergodic, $\tilde{m}_n(\beta) \xrightarrow{p} m_0(\beta)$, implying $\hat{m}_n(\beta) \xrightarrow{p} m_0(\beta)$, so that $\hat{Q}(\beta) \xrightarrow{p} Q(\beta)$. Noting that $\hat{Q}(\beta)$ is convex by $\hat{m}_n(\beta)$ convex, the conclusion then follows from $Q(\beta)$ uniquely minimized at β_0, as in Anderson and Gill (1982). Under Assumption 5.6 b), $\sup_{\beta\in\mathcal{B}}\|\hat{m}_n(\beta)-\tilde{m}_n(\beta)\| \xrightarrow{p} 0$ by Assumption 5.2, while $\sup_{\beta\in\mathcal{B}}\|\tilde{m}_n(\beta)-m(\beta)\| \xrightarrow{p} 0$ follows by Andrews (1987), so that $\sup_{\beta\in\mathcal{B}}\|\hat{Q}(\beta)-Q(\beta)\| \xrightarrow{p} 0$. The conclusion now follows by the Wald argument for extremum estimators. ■

Proof of Theorem 5.4: By Lemmas 5.1 and 5.2, Asymptotic Linearity and Stochastic Equicontinuity are satisfied, and by orthogonality of $M(z)$ with $h(z)$ and $\hat{h}(z)$, $\sqrt{n}\int M(z)[\hat{h}(v)-h(v)]dF(z) = 0$. Then by the α-mixing central limit theorem of White and Domowitz (1984), $\sqrt{n}\hat{m}_n(\beta_0) = \sum_{i=1}^n m_i/\sqrt{n} + o_p(1) \xrightarrow{d} N(0,\Omega)$. The remainder of the proof then follows from a standard minimum distance argument, such as that in Newey (1988b). ■

Proof of Theorem 5.5: It follows by Lemma 5.1 that Asymptotic Linearity is

satisfied, so that by Asymptotic Differentiability, the triangle inequality, and the α-mixing central limit theorem of White and Domowitz (1984), $\sqrt{n}\hat{m}_n(\beta_0) = \sum_{i=1}^n u_i/\sqrt{n} + o_p(1) \xrightarrow{d} N(0,\Omega)$. The remainder of the proof then follows by a standard minimum distance argument, such as that in Newey (1988b). ∎

Proof of Theorem 5.6: By a mean value expansion,

$$(A.2)\quad \sum_{i=1}^n \|\hat{m}_i - m_i\|^2/n \le (\|\hat{\beta}-\beta_0\|^2 + \sup_v \|\hat{h}(v)-h(v)\|^2)\sum_{i=1}^n [b_{10}(z_i)^2 + b_{01}(z_i)^2]/n$$
$$= O_p(\epsilon_n^2).$$

Therefore, $\sum_{i=1}^n \|\hat{u}_i - u_i\|^2/n = O_p(\epsilon_n^2)$. Let $\tilde{\Omega}_\ell = \sum_{i=1}^{n-\ell} u_i u_{i+\ell}'/n$, $\tilde{\Omega} = \tilde{\Omega}_0$ in case a), and $\tilde{\Omega} = \tilde{\Omega}_0 + \sum_{\ell=1}^{\hat{L}} w(\ell,\hat{L})[\tilde{\Omega}_\ell + \tilde{\Omega}_\ell']$ in case b). Note $\|\hat{u}_i\hat{u}_{i+\ell}' - u_i u_{i+\ell}'\| \le \|\hat{u}_i - u_i\|\|\hat{u}_{i+\ell} - u_{i+\ell}\| + \|u_i\|\|\hat{u}_{i+\ell} - u_{i+\ell}\| + \|u_{i+\ell}\|\|\hat{u}_i - u_i\|$, so that for all $\ell \ge 0$, by the Cauchy-Schwartz and Markov inequalities,

$$(A.3)\quad \|\hat{\Omega}_\ell - \tilde{\Omega}_\ell\| \le \{\sum_{i=1}^{n-\ell}\|\hat{u}_i - u_i\|^2/n\}^{1/2}\{\sum_{i=1}^{n-\ell}\|\hat{u}_{i+\ell} - u_{i+\ell}\|^2/n\}^{1/2}$$
$$+ \{\sum_{i=1}^{n-\ell}\|u_i\|^2/n\}^{1/2}\{\sum_{i=1}^{n-\ell}\|\hat{u}_{i+\ell} - u_{i+\ell}\|^2/n\}^{1/2}$$
$$+ \{\sum_{i=1}^{n-\ell}\|u_{i+\ell}\|^2/n\}^{1/2}\{\sum_{i=1}^{n-\ell}\|\hat{u}_i - u_i\|^2/n\}^{1/2}$$
$$\le \sum_{i=1}^n \|\hat{u}_i - u_i\|^2/n + \{\sum_{i=1}^n \|u_i\|^2/n\}^{1/2}\{\sum_{i=1}^n \|\hat{u}_i - u_i\|^2/n\}^{1/2} = O_p(\epsilon_n).$$

In case a), $\|\hat{\Omega} - \tilde{\Omega}\| = \|\hat{\Omega}_0 - \tilde{\Omega}_0\| = O_p(\epsilon_n) = o_p(1)$, while $\|\tilde{\Omega} - \Omega\|$ follows by the law of large numbers, giving the conclusion. In case b), there is a sequence of numbers $\delta_n \to 0$ such that for $L' = \delta_n/\epsilon_n$, $\text{Prob}(L \le L') \to 1$, were L' can be chosen as an integer by the argument from Newey (1990b). Then by boundedness of $w(\ell,L)$ and eq. (A.3), with probability approaching one $\|\hat{\Omega} - \tilde{\Omega}\| \le \|\hat{\Omega}_0 - \tilde{\Omega}_0\| + C\sum_{\ell=1}^{L'}\|\tilde{\Omega}_\ell - \Omega_\ell\| \le (1+CL')O_p(\epsilon_n) = o_p(1)$. Also, $1/\sqrt{n} = O(\epsilon_n)$, so by Davidov's inequality, arguing as in Kool (1988), it follows that for $\Omega_L = \Omega_0 + \sum_{\ell=1}^{L} w(\ell,L)[\Omega_\ell + \Omega_\ell']$, with probability approaching one, $\|\tilde{\Omega} -$

$\Omega_L\| \leq \|\tilde{\Omega}_0-\Omega_0\| + C\sum_{\ell=1}^{L'}\|\tilde{\Omega}_\ell - \Omega_\ell\| = O_p(L'/\sqrt{n}) = o_p(1)$. Finally, applying the dominated convergence theorem as in Newey and West (1987), it follows by $L \xrightarrow{p} \infty$ that $\|\Omega_0+\sum_{\ell=1}^{L} w(\ell,L)[\Omega_\ell + \Omega_\ell'] - \Omega\| = o_p(1)$. ■

Proof of Theorem 6.1: To show consistency, note first that iii) implies that $\bar{K}^{1/2}\zeta_0(\bar{K})/\sqrt{n} = o(1)$ and $\bar{K}^{1/2}\zeta_0(\bar{K})\underline{K}^{-d/r} = o(1)$, so that Assumption 5.2 is satisfied by Theorem 6.1 of Newey (1991). Consistency of $\hat{\beta}$ then follow by Theorem 5.3. Let objects without subscripts denote vectors of observations. Asymptotic Linearity follows by Lemma 5.1 and iv), since by Theorem 6.1 of Newey (1991), $\|\hat{g}-g\|^2/n = O_p(\bar{K}/\sqrt{n} + \underline{K}^{-2d/r}) = o_p(n^{1/\ell_{02}-1/2})$, so Assumption 5.4 is satisfied.

Next, Asymptotic Differentiability is shown, with $\alpha_j(z)$ as derived in Section 4. For notational convenience the j subscript will be dropped and $M(z)$ treated as a scalar (for vector $M(z)$ the result follows by applying the following argument to each of its elements). Let $Q = p(p'p)^{-}p'$, $M = (M(z_1),\ldots,M(z_n))'$, and $\hat{\delta} = QM$. Then by Q idempotent,

$$(A.4) \qquad M'(\hat{g}-g) - \delta'(y-g) = \hat{\delta}'\hat{g} - M'g - \delta'y + \delta'g$$

$$= (\hat{\delta}-\delta)'(\hat{g}-g) + \delta'(y-\hat{g}) + (\hat{\delta}-\delta)'g = R_1 + R_2 + R_3.$$

By Theorem 6.1 of Newey (1991) and iii),

$$|R_1|/\sqrt{n} \leq \sqrt{n}\|\hat{\delta}-\delta\|\|\hat{g}-g\| = O_p(\sqrt{n}(\bar{K}^{1/2}/\sqrt{n} + \underline{K}^{-d_\delta/r})(\bar{K}^{1/2}/\sqrt{n} + \underline{K}^{-d/r})) = o_p(1).$$

By Lemma 8.1 of Newey (1991), there are $g_K(x) = p^K(x)'\pi_g^K$ and $\delta_K(x) = p^K(x)'\pi_\delta^K$ such that $\|g(x)-g_K(x)\|_0 \leq CK^{-d/r}$ and $\|\delta(x)-\delta_K(x)\|_0 \leq CK^{-d_\delta/r}$. Then by Q idempotent.

(A.5) $|R_2|/\sqrt{n} = |\delta'(I-Q)y|/\sqrt{n} \leq |(\delta-\delta_{\hat{K}})'(y-g)|/\sqrt{n} + |(\delta-\delta_{\hat{K}})'Q(y-g)|/\sqrt{n}$

$+ |(\delta-\delta_{\hat{K}})'(I-Q)(g-g_{\hat{K}})|/\sqrt{n} = R_{21} + R_{22} + R_{23}.$

By I-Q idempotent and $\hat{K} \geq \underline{K}$ with probability approaching one, $R_{23} \leq \|\delta-\delta_{\hat{K}}\|\|g-g_{\hat{K}}\|/\sqrt{n} = O_p(\sqrt{n}\underline{K}^{-(d+d_\delta)/r}) = o_p(1)$. By Lemma 9.8 of Newey (1991), $(y-g)'Q(y-g)/n = O_p(\bar{K}/n)$, so that $R_{22} \leq \|\delta-\delta_{\hat{K}}\|[(y-g)'Q(y-g)/n]^{1/2} = O_p(\bar{K}^{1/2}\underline{K}^{-d_\delta/r}) = o_p(1)$. By the strong mixing hypotheses, Davydov's inequality, and Assumption 6.4, for $\rho = 1-2/\mu$ and $\mathcal{K} = [\underline{K},\bar{K}]$, $R_{21}^2 \leq |(\delta-\delta_{\hat{K}})'(y-g)|^2/n = O_p(\sum_{\mathcal{K}}E[|(\delta-\delta_K)'(y-g)|^2]/n) = O_p(\sum_{\mathcal{K}}E[|(\delta-\delta_K)'(y-g)|^2]/n) \leq O_p(\sum_{\mathcal{K}}|(\delta(x)-\delta_K(x))(y-g(x))|_\rho^2) = O_p(\sum_{\mathcal{K}}|(\delta(x)-\delta_K(x))(y-g(x))|_\rho^2) = O_p(\sum_{\mathcal{K}}K^{-2d_\delta/r})$. Then since $2d_\delta/r > 1$ follows from $\bar{K}^{1/2}\zeta_0(\bar{K})\underline{K}^{-d_\delta/r}$ converging to zero, $\sum_{\mathcal{K}}K^{-2d_\delta/r} = o(1)$ follows, implying $R_{21}^2 = o_p(1)$. Note also that R_3 has the same form as R_2, with y and M interchanged, so that $R_3/\sqrt{n} = o_p(1)$ also follows. Finally, note that $\delta(x)$ is bounded, and $|\varepsilon|_s < \infty$ for $s > 2\mu/(\mu-1)$, so that $|\alpha(z)|_s < \infty$ for $s > 2\mu/(\mu-1)$, so that all of the hypotheses of Theorem 5.5 are satisfied and the first conclusion follows from its conclusion.

To prove the second conclusion, note first that by Theorem 6.1 of Newey (1991), $|\hat{g}(x)-g(x)|_\infty = O_p(\bar{K}^{1/2}\zeta_0(\bar{K})[\bar{K}^{1/2}/\sqrt{n} + \underline{K}^{-d/r}]) = O_p(\epsilon_n)$. Therefore, by Theorem 5.6, it only remains to be shown that $\sum_{i=1}^n\|\hat{\alpha}_i-\alpha(z_i)\|^2/n = O_p(\epsilon_n^2)$. It suffices to show this result for each element of α, and hence α can be assumed to be scalar without loss of generality. Let $\varepsilon_i = y_i-g(x_i)$, $\hat{\varepsilon}_i = y_i-\hat{g}(x_i)$, $\hat{M}_i = \partial m(z_i,\hat{\beta},\hat{h}(x))/\partial h$, $\hat{M} = (\hat{M}_1,\ldots,\hat{M}_n)$. Then

(A.6) $C\sum_{i=1}^n\|\hat{\alpha}_i-\alpha(z_i)\|^2/n \leq \sum_{i=1}^n\|\hat{M}'p^{\hat{K}}\hat{\Sigma}^-p_i^{\hat{K}}(\hat{\varepsilon}_i-\varepsilon_i)\|^2/n + \sum_{i=1}^n\|(\hat{M}-M)'p^{\hat{K}}\hat{\Sigma}^-p_i^{\hat{K}}\varepsilon_i\|^2/n$

$+ \sum_{i=1}^n\|(M'p^{\hat{K}}\hat{\Sigma}^-p_i^{\hat{K}}-\delta_i)\varepsilon_i\|^2/n = R_1 + R_2 + R_3.$

By $\ell_{01} \geq 2$ and Q idempotent, $\hat{M}'Q\hat{M}/n \leq \hat{M}'\hat{M}/n = O_p(1)$. Therefore, by Theorem 6.1 of Newey (1991),

$$(A.7) \qquad R_1 \leq \sup_x |\hat{g}(x)-g(x)|^2 \hat{M}'Q\hat{M}/n = O_p(\epsilon_n^2).$$

Also, by $\ell_{11} \geq 1$, $\ell_{02} \geq 1$ and a Taylor expansion, $(\hat{M}-M)'Q(\hat{M}-M)/n \leq \|\hat{M}-M\|^2/n = O_p(\epsilon_n^2 n^{-2/s})$, so

$$(A.8) \quad R_2 \leq (\max_{i\leq n}\varepsilon_i^2)\sum_{i=1}^n \|(\hat{M}-M)'p^{\hat{K}}\hat{\Sigma}^- p_i^{\hat{K}}\|^2/n = O_p(n^{2/s})(\hat{M}-M)'Q(\hat{M}-M)/n = O_p(\epsilon_n^2).$$

Finally, note that $M'p^{\hat{K}}\hat{\Sigma}^- p_i^{\hat{K}}$ is $\hat{\delta}(x_i)$ where $\hat{\delta}(x_i)$ is the series estimator of $\delta(x)$ from regressing M on $p^{\hat{K}}$. Thus, by Theorem 6.1 of Newey (1991),

$$(A.9) \qquad R_3 \leq (\max_{i\leq n}\varepsilon_i^2)\sum_{i=1}^n \|\hat{\delta}(x_i)-\delta(x_i)\|^2/n = O_p(\epsilon_n^2). \qquad \blacksquare$$

Proof of Theorem 6.2: First, consider the case where Assumption 6.7 a) is satisfied. Consistency of $\hat{\beta}$ follows as in the proof of Theorem 6.1, noting that $\alpha_0 = d/r$. Next, it follows by Theorem 6.1 of Newey (1991) and Assumption 6.6 that $\sqrt{n}\|\hat{h}(v)-h_0(v)\|_0^2 \leq C\sqrt{n}\|\hat{g}(x)-g(x)\|_\Delta^2 = O_p(\sqrt{n}\bar{K}\zeta_\Delta(\bar{K})^2[\bar{K}/n+\underline{K}^{-2\alpha}]) = o_p$, so that Asymptotic Linearity follows by the same Taylor expansion argument as used in the proof of Lemma 5.1. To show Asymptotic Differentiability, let

$$\Psi_K = \int \delta(x)p^K(x)dF(x), \quad \Sigma_K = \int p^K(x)p^K(x)'dF(x), \quad \pi_K = \Sigma_K^{-1}E[p^K(x)g(x)],$$

$$\delta_K(x) = p^K(x)'\Sigma_K^{-1}\Psi_K, \quad g_K(x) = p^K(x)'\pi_K, \quad h_K(v) = A(v,g_K) = A(v)'\pi_K,$$

$$\tilde{\Psi} = \sum_{i=1}^n M(z_i)A(v_i)'/n.$$

Note that $\hat{h}(v)$ is invariant to nonsingular linear transformations, so that without loss of generality $p_1(x)$, $p_2(x)$, ... can be assumed to be the functions in the conclusion of Lemma 8.4 of Newey (1991), for which the

smallest eigenvalue of Σ is bounded below and there is a constant C such that $\sup_{x\in\mathcal{X},\, k\leq K}|D^{\lambda}p_k(x)| \leq C\zeta_{|\lambda|}(K)$. Then, by Assumption 6.6

(A.10) $\quad \max_{k\leq K}\|A(v,p_k)\|_0 \leq \zeta_{\Delta}(K).$

It then follows by $|M(z)|_s$ finite for $s > 2\mu/(\mu-1)$ and Lemma 9.6 of Newey (1991) that

(A.11) $\quad \|\tilde{\Psi}-\Psi_{\hat{K}}\| = O_p(\bar{K}^{1/2}\zeta_{\Delta}(\bar{K})/\sqrt{n}) = o_p, \quad \|\hat{\Sigma}-\Sigma\| = O_p(\bar{K}\zeta_0(\bar{K})^2/\sqrt{n}) = o_p.$

Also, as in the proof Lemma 9.8 of Newey (1991),

(A.12) $\quad \|\hat{\Sigma}^{-1/2}p'(y-g)/\sqrt{n}\| = O_p(\bar{K}^{1/2}/\sqrt{n}).$

Next, by Lemma 8.1 of Newey (1991) there exists $\tilde{g}_K(x) = p^K(x)'\tilde{\pi}$ such that $\|g(x)-\tilde{g}_K(x)\|_{\Delta} \leq CK^{-\alpha}$. Note that $\|\pi-\tilde{\pi}\| \leq C[(\pi-\tilde{\pi})'\Sigma(\pi-\tilde{\pi})]^{1/2} = C|g_K(x)-\tilde{g}_K(x)|_2 \leq C[|g(x)-g_K(x)|_2+|g(x)-\tilde{g}_K(x)|_2] \leq C|g(x)-\tilde{g}_K(x)|_2 \leq CK^{-\alpha}$. Therefore,

(A.13) $\quad \|g(x)-g_K(x)\|_{\Delta} \leq \|g(x)-\tilde{g}_K(x)\|_{\Delta} + \|\tilde{g}_K(x)-g_K(x)\|_{\Delta} \leq C(K^{-\alpha}+\|p^K(x)\|_{\Delta}\|\tilde{\pi}-\pi\|)$

$$\leq C(K^{-\alpha}+\|p^K(x)\|_{\Delta}\|\tilde{\pi}-\pi\|) \leq CK^{1/2}\zeta_{\Delta}(K)K^{-\alpha},$$

By the definition of the least squares coefficients π and Lemma 8.1 of Newey (1991) it follows that $E[p^K(x)\{g(x)-g_K(x)\}] = 0$ and $|g(x)-g_K(x)|_2 \leq CK^{-\alpha}o$, so that under uniform mixing, for $p^K = [p^K(x_1),\ldots,p^K(x_n)]'$,

(A.14) $\quad E[\sum_{\mathcal{K}}\|p^{K\prime}(g-g_K)/\sqrt{n}\|^2] \leq C\sum_{\mathcal{K}}E[\|p^K(x)\|^2(g(x)-g_K(x))^2] \leq C\sum_{\mathcal{K}}K\zeta_0(K)^2K^{-2\alpha}o,$

which converges to zero by i). Without uniform mixing, it follows by strong mixing and boundedness of $p^K(x)$, $g(x)$, and $g_K(x)$ that

(A.15) $E[\sum_{\mathcal{K}}\|p^{K}(g-g_K)/\sqrt{n}\|^2] \leq C\sum_{\mathcal{K}}|\|p^{K}(x)\|^2(g(x)-g_K(x))^2|_\infty^2 \leq C\sum_{\mathcal{K}}K^2\zeta_0(K)^4K^{-2\alpha_0},$

which converges to zero by ii) b). Then, since $\|p'(g-g_{\hat{K}})/\sqrt{n}\|^2 \leq \sum_{\mathcal{K}}\|p^{K}(g-g_K)/\sqrt{n}\|^2$ with probability approaching one, it follows by the Markov inequality that $\|p'(g-g_{\hat{K}})/\sqrt{n}\|^2 = o_p$.

Now, it follows by a little arithmetic that

(A.16) $\sum_{i=1}^{n}M(z_i)[\hat{h}(v_i)-h(v_i)]/n = (\tilde{\Psi}-\Psi)'\hat{\Sigma}^{-1}p'(g-g_{\hat{K}})/n + \Psi'(\hat{\Sigma}^{-1}-\Sigma^{-1})p'(g-g_{\hat{K}})/n$

$+ \Psi_{\hat{K}}'\Sigma^{-1}p'(g-g_{\hat{K}})/n + (\tilde{\Psi}-\Psi_{\hat{K}})'\hat{\Sigma}^{-1}p'(y-g)/n + \Psi'(\hat{\Sigma}^{-1}-\Sigma^{-1})p'(y-g)/n$

$+ (\delta_{\hat{K}}-\delta)'(y-g)/n + \sum_{i=1}^{n}(M(z_i)[h_{\hat{K}}(v_i)-h(v_i)]-\int M(z)[h_{\hat{K}}(v)-h(v)]dF(z))/n$

$+ \int M(z)[h_{\hat{K}}(v)-h(v)]dF(z) + \delta'(y-g)/n = \sum_{j=1}^{8}R_j + \delta'(y-g)/n.$

By $\|\hat{\Sigma}-\Sigma\| = o_p$ and $|\lambda_{min}(\hat{\Sigma})-\lambda_{min}(\Sigma)| \leq \|\hat{\Sigma}-\Sigma\|$, the largest eigenvalue of $\hat{\Sigma}^{-1}$ is bounded in probability, so that $\|(\tilde{\Psi}-\Psi_{\hat{K}})'\hat{\Sigma}^{-1}\| \leq \|\tilde{\Psi}-\Psi_{\hat{K}}\|O_p(1) = o_p(1)$. Also, $\|\Psi_{\hat{K}}'\Sigma^{-1}\| \leq C[\Psi_{\hat{K}}'\Sigma^{-1}\Psi_{\hat{K}}]^{1/2} \leq CE[\delta(x)^2]$, so that $\|\Psi_{\hat{K}}'\Sigma^{-1}\| = O_p(1)$, and hence $\|\Psi_{\hat{K}}'(\hat{\Sigma}^{-1}-\Sigma^{-1})\| \leq \|\Psi_{\hat{K}}'\Sigma^{-1}\|\|(\hat{\Sigma}-\Sigma)\hat{\Sigma}^{-1}\| = o_p$. It now follows by $\sqrt{n}\|p'(g-g_{\hat{K}})/n\| = o_p$ that $\sqrt{n}R_j = o_p$, $(j = 1, 2, 3)$. Also, by $\|\hat{\Sigma}^{-1/2}p'(y-g)/\sqrt{n}\| = O_p(\bar{K})$, it follows similarly from eq. (A.11) and iii) that $\sqrt{n}R_4 = o_p$ and $\sqrt{n}R_5 = o_p$.

Next, note that by either uniform mixing or the bound on the conditional covariances, $E[\{(\delta_K-\delta)'(y-g)\}^2/n] \leq CE[\{1+\varepsilon^2\}\{\delta_K(x)-\delta(x)\}^2] \leq CE[\{1+E[\varepsilon^2|x]\}\{\delta_K(x)-\delta(x)\}^2] \leq C\epsilon_\delta(K)^2$, so that by Assumption 6.4 and iii),

(A.17) $n|R_6|^2 = O_p(\sum_{\mathcal{K}}E[\{(\delta_K-\delta)'(y-g)\}^2/n]) = O_p(\sum_{\mathcal{K}}\epsilon_\delta(K)^2) = o_p.$

Similarly, note that by eq. (A.13), $|h_K(v)-h(v)|_\infty \leq CK^{1/2}\zeta_\Delta(K)K^{-\alpha}$, so that by strong mixing, Davydov's inequality, and i) for $p = 1 - 2/\mu$,

(A.18) $$n|R_7|^2 = O_p(\sum_{\mathcal{K}}|M(z)[h_K(v)-h(v)]|_\rho^2) = O_p(\sum_{\mathcal{K}}K\zeta_\Delta(K)^2K^{-2\alpha}) = o_p.$$

Furthermore, by Assumption 4.1, $\Psi_K = E[\delta(x)p^K(x)]$, so that by $\int\delta(x)g_{\hat{K}}(x)dF(x) = \int\delta_{\hat{K}}(x)g_{\hat{K}}(x)dF(x) = \int\delta_{\hat{K}}(x)g(x)dF(x)$,

(A.19) $$\sqrt{n}|R_8| = \sqrt{n}|\int\delta(x)g_{\hat{K}}(x)dF(x) - E[\delta(x)g(x)]|$$

$$= \sqrt{n}|\int[\delta(x)-\delta_{\hat{K}}(x)][g_{\hat{K}}(x)-g(x)]dF(x)| \leq \sqrt{n}\epsilon_\delta(\hat{K})\hat{K}^{-\alpha_0} \leq \sqrt{n}\epsilon_\delta(\underline{K})\underline{K}^{-\alpha_0} \longrightarrow 0,$$

where the last inequality follows by $\epsilon_\delta(K)$ monotonically decreasing in K.

For the case where Assumption 6.7 b) is satisfied, it follows by Lemma 8.2 of Newey (1991) that all the previous arguments hold with α and α_0 replaced by any (arbitrarily large) positive number $\bar{\alpha}$. It then follows $\bar{K}$ bounded by a power of n, $\zeta_d(K)$ bounded by a power of K, and Assumption 6.4 that all terms above where $K^{-\alpha_0}$, $K^{-\alpha}$, $\underline{K}^{-\alpha_0}$, or $\underline{K}^{-\alpha}$ appear are small, so that all the terms depending on α_0 or α in the statement of the Theorem can be ignored, i.e. α_0 and α can be set to $+\infty$, again giving the first conclusion.

The second conclusion will be shown only under case a) of Assumption 6.7, because under case b) the result will follow as above. For notational convenience, suppress the j subscript on each $h_j(v)$. It follows from Theorem 6.1 of Newey (1991) that $|\hat{h}(v)-h(v)|_\infty = O_p(\bar{K}^{1/2}\zeta_\Delta(\bar{K})[\bar{K}^{1/2}/\sqrt{n}+\underline{K}^{-\alpha}]) = O_p(\epsilon_n)$. Therefore, by Theorem 5.6, it only remains to be shown that $\sum_{i=1}^n\|\hat{\alpha}_i-\alpha(z_i)\|^2/n = O_p(\epsilon_n^2)$. Let $\hat{\Psi} = \sum_{i=1}^n m_h(z_i,\hat{\beta},\hat{h}(v_i))A(v_i)'/n$ and define $\epsilon_\Psi = \bar{K}^{1/2}\zeta_\Delta(\bar{K})[\bar{K}^{1/2}/\sqrt{n} + \underline{K}^{-\alpha}]$ if m is linear in h and $\epsilon_\Psi = \bar{K}\zeta_\Delta(\bar{K})^2[\bar{K}^{1/2}/\sqrt{n} + \underline{K}^{-\alpha}]$ otherwise. By eq. (A.10) $\sup_{v\in\mathcal{V}}\|A(v)\| \leq C\bar{K}^{1/2}\zeta_\Delta(\bar{K})$, so that by an expansion,

$$(A.20)\quad \|\hat{\Psi}-\Psi_{\hat{K}}\| \le \|\tilde{\Psi}-\Psi_{\hat{K}}\| + |\|A(v)\||_\infty\{\|\hat{\beta}-\beta_0\|\sum_{i=1}^n \|m_{h\beta}(z_i,\bar{\beta},\bar{h}(v_i))\|/n$$

$$+ |\hat{h}(v)-h_0(v)|_\infty\sum_{i=1}^n \|m_{hh}(z_i,\bar{\beta},\bar{h}(v_i))\|/n\} = O_p(\epsilon_\Psi).$$

Also, it follows as in the proof of eq. (A.11) that $\|\Psi_{\hat{K}}'\Sigma^{-1}(\Sigma-\hat{\Sigma})\hat{\Sigma}^{-1}\| = O_p(\epsilon_\Sigma)$ for $\epsilon_\Sigma = \bar{K}\zeta_0(\bar{K})^2/\sqrt{n}$ and $\|\Psi_{\hat{K}}'\Sigma^{-1}\|$ is bounded, so that $\|\hat{\Psi}'\hat{\Sigma}^{-1}-\Psi_{\hat{K}}'\Sigma^{-1}\| \le \|(\hat{\Psi}-\Psi_{\hat{K}})'\hat{\Sigma}^{-1}\| + \|\Psi_{\hat{K}}'\Sigma^{-1}(\Sigma-\hat{\Sigma})\hat{\Sigma}^{-1}\| = O_p(\epsilon_\Psi+\epsilon_\Sigma) = o_p$, and $\|\hat{\Psi}'\hat{\Sigma}^{-1}\| = O_p(1)$.

Next, note that

$$(A.21)\quad \sum_{i=1}^n |\hat{\alpha}_i-\alpha_i|^2/n \le C\sum_{i=1}^n |\hat{\Psi}'\hat{\Sigma}^{-1}P_i(\hat{\varepsilon}_i-\varepsilon_i)|^2/n + C\sum_{i=1}^n |(\hat{\Psi}'\hat{\Sigma}^{-1}-\Psi_{\hat{K}}'\Sigma^{-1})P_i\varepsilon_i|^2/n$$

$$+ C\sum_{i=1}^n |[\delta_{\hat{K}}(x_i)-\delta(x_i)]\varepsilon_i|^2/n = R_1 + R_2 + R_3.$$

Therefore, by Theorem 6.1 of Newey(1991), and $d/n \ge \alpha$,

$$(A.22)\quad |R_1| \le C\|\hat{\Psi}'\hat{\Sigma}^{-1}\||\|p^K(x)\||_\infty^2\sum_{i=1}^n |\hat{g}(x_i)-g(x_i)|^2/n$$

$$= O_p(\bar{K}\zeta_0(\bar{K})^2[(\bar{K}/n) + \underline{K}^{-2d/n}]) = O_p(\epsilon_n^2).$$

Also, $\sum_{i=1}^n \|\hat{\Sigma}^{-1/2}P_i\|^2/n = \sum_{i=1}^n tr(\hat{\Sigma}^{-1/2}P_iP_i'\hat{\Sigma}^{-1/2})/n = tr(\hat{\Sigma}^{-1/2}(P'P/n)\hat{\Sigma}^{-1/2})/n$ is equal to the dimension of $\hat{\Sigma}$, less than or equal to $\bar{K}$ w.p.a.1., so

$$(A.23)\quad |R_2| \le C(\max_{i\le n}\varepsilon_i^2)(\sum_{i=1}^n \|\hat{\Sigma}^{-1/2}P_i\|^2/n)[\|(\hat{\Psi}-\Psi)'\hat{\Sigma}^{-1/2}\|^2$$

$$+ \|\Psi'\Sigma^{-1/2}\|^2\|\Sigma^{-1/2}(\Sigma-\hat{\Sigma})\hat{\Sigma}^{-1/2}\|^2] = O_p(n^{2/s}\bar{K}(\epsilon_\Sigma+\epsilon_\Psi)^2).$$

Finally, it follows as in the proof of Theorem 6.1, using $E[\varepsilon^2|x]$ bounded, that $|R_3| = O_p(\sum_{K\in\mathcal{K}}E[\{\delta_K(x)-\delta(x)\}^2]) = O_p(\epsilon_n^2)$. ∎

Proof of Theorem 7.1: The proof proceeds by verifying the hypotheses of Theorem 6.1. Assumption 5.1 holds by i). The estimator has the form of Section 6, where $m(z,\beta,\sigma_2,h) =$

$1(0<h_1<1, 0<h_2<1)(x_1'-x_2', \Phi^{-1}(h_2))'[\Phi^{-1}(h_1)-\sigma_2\Phi^{-1}(h_2)-(x_1-x_2)'\beta]$. Note that $x_t + \rho(x)$ is bounded, so that $h_t(x) = \Phi([x_t+\rho(x)]/\sigma_{t0})$ is bounded away from zero and one, and that $\Phi^{-1}(\cdot)$ is continuously differentiable on any set where its argument is bounded away from zero and one. It follows that Assumptions 5.3 and 5.6 are satisfied for case a) of Assumption 5.6 and $\ell_{jk} = \infty$, $1 \leq j+k \leq 2$. Assumptions 6.1 - 6.5 also hold, with $\nu = 0$ and $s = \infty$. Also, note that $M_j(z)$ is a bounded function of x and $\mathcal{G}$ is the set of all mean-square integrable functions of x, so that $M_j(z) = \delta_j(x)$, and i) - ii) of Theorem 6.1 are satisfied with $d_\delta = d$. Noting that $\zeta_0(K) = K^{1/2}$, it follows by vi) that Theorem 6.1 iii) and iv) are satisfied, since each of $n^{4\Gamma-1}$, $n^{\Gamma-\gamma(d/2k)}$, and $n^{.5-\gamma d/k}$ converge to zero. The first conclusion now follows by Theorem 6.1. Next, note $s = \infty$ by $y_t-h_t(x)$ bounded, so that Theorem 6.1 v) is implied by Theorem 6.1 iii), so $\epsilon_n \longrightarrow 0$ and the second conclusion also follows by Theorem 6.1. ■

Proof of Theorem 7.2: Follows similarly to the proof of Theorem 7.3 to follow, on noting that 1) Assumption 4.1 is satisfied, where $E[A(v,g)] = E[\int_{\mathcal{A}} g(x_1,x_2)dx_1] = \mathcal{L}(\mathcal{A})E[g(x)\tilde{f}(x)]$, $\mathcal{L}$ is the Lebesgue measure, and hence $\delta(x) = \mathcal{L}(\mathcal{A})\Pi(\tilde{f}(x)|\mathcal{G})$; 2) $\mathcal{L}(\mathcal{A})E[\tilde{f}(x)|x_1] = f(x_1,x_2^1)^{-1}f(x_2^1)$ in case b), where the projection has an explicit form. ■

Proof of Theorem 7.3: The proof proceeds by verifying the hypotheses of Theorem 6.2. Assumption 5.1 holds by i). The estimator has the form of Section 6, where $m(z,\beta,h) = g(v)-g(x)-\beta$, so that Assumptions 5.3 and 5.6 are satisfied for case a) of Assumption 5.6 and $\ell_{jk} = \infty$, $1 \leq j+k \leq 2$. Also, $\Delta = 0$, so that Assumption 6.7 a) is satisfied. Let $h_1(v_1) = g(v)$, $h_2(v_2) = g(x)$, so that $\delta_2(x) = 1$. To discuss $\delta_1(x)$, note first that by Lemma 8.0 of Newey (1991), $\mathcal{G}$ is closed, so that $\Pi(\tilde{f}(x)|\mathcal{G})$ exists. As shown in Section 4, Assumption 4.1 is satisfied for $\delta_1(x) = \Pi(\tilde{f}(x)|\mathcal{G})$, so that under

iii) b), it follows by Lemma 8.2 of Newey (1991) that $\epsilon_\delta(K) \leq K^{-d_\delta/r}$. Under iii) a), the projection has an explicit form $\Pi(M(z)|\mathcal{G}) = E[M(z)|x^1] + \{x^2 - E[x^2|x^1]\}(E[Var(x^2|x^1)])^{-1}E[\{x^2 - E[x^2|x^1]\}M(z)]$, and that $E[\tilde{f}(x)|x^1] = f_{x^1}(x^1)^{-1}f_{v^1}(x^1)$, so that part b) is satisfied. Noting that $\zeta_0(K) = \zeta_\Delta(K) = K^{1/2}$ and $\alpha_0 = \alpha = -d/r$, it follows by iv) that the hypotheses of Theorem 6.2 are satisfied, since each of $n^{\gamma(3-2d/r)}$, $n^{1/2-\gamma(d+d_\delta)/r}$, $n^{5\Gamma-1+(1/s)}$, and $n^{\Gamma-\gamma(d/r)}$ converge to zero. The conclusions now follow from the conclusion of Theorem 6.2. ■

Proof of Theorem 7.4: First the result will be proven when x_1 is a scalar. Assumption 5.1 holds by i). The estimator has the form of Section 6, where $m(z,\beta,h) = m(z,\partial g(x)/\partial x_1) - \beta$, so that Assumptions 5.3 and 5.6 are satisfied for case a) of Assumption 5.6 and $\ell_{jk} = \infty$, $1 \leq j+k \leq 2$. As shown in Section 4, $\delta(x) = \Pi(f(x)^{-1}\partial f(x)/\partial x_1|\mathcal{G})$, so that by the usual mean-square spanning result for polynomials, $\epsilon_\delta(K) \longrightarrow 0$ as $K \longrightarrow \infty$. Also, since Assumption 6.7 b) is satisfied, none of the conditions of Theorems 6.2 that depend on α or α_0 are binding. The conclusion now follows by Theorem 6.2, since $\Delta = 1$ and when $m(z,h)$ is linear in h and both $n^{(1/s)+\Gamma[(7/2)+\nu]-(1/2)}$ and $n^{(1/s)+\Gamma[(5/2)+2\nu]-(1/2)}$ go to zero, while otherwise $n^{(1/s)+\Gamma(2+1+2\nu+4)-(1/2)}$ converges to zero, so the conclusion follows by Theorem 6.2. ■

References

Ahn, H. and C.F. Manski (1989): "Distribution Theory for Binary Choice Analysis with Nonparametric Estimation of Expectations," mimeo, University of Wisconsin.

Ai, C. (1990): "Semiparametric Estimation of Index Distribution Models," MIT Ph.D. Thesis.

Anderson, P.K. and R.D. Gill (1982): "Cox's Regression Model for Counting Processes: A Large Sample Study," *Annals of Statistics* 10, 1100-1120.

Andrews, D.W.K. (1987): "Consistency in Nonlinear Econometric Models, A Generic Uniform Law of Large Numbers," *Econometrica* 55, 1465-1471.

Andrews, D.W.K. (1990a): "Asymptotics for Semiparametric Econometric Models: I. Estimation," mimeo, Cowles Foundation, Yale University.

Andrews, D.W.K. (1990b): "Asymptotics for Semiparametric Econometric Models: II. Stochastic Equicontinuity," mimeo, Cowles Foundation, Yale University.

Andrews, D.W.K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Models," *Econometrica* 59, 307-345.

Bickel P., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1990): "Efficient and Adaptive Inference in Semiparametric Models" monograph, forthcoming.

Boos, D.D. and R.J. Serfling (1980): "A Note on Differentials and the CLT and LIL for Statistical Functions, with Application to M-Estimates," *Annals of Statistics*, 8, 618-624.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association.* 80 580-598.

Buckley, J. and I. James (1979): "Linear Regression with Censored Data," *Biometrika*, 66, 429-436.

Chamberlain, G. (1980), "The Analysis of Covariance with Qualitative Data," *Review of Economic Studies,* 47, 225-238.

Fernholz, L.T. (1983): *von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics 19. Berlin: Springer.

Hansen, L.P. (1985): "Two-Step Generalized Method of Moments Estimators," discussion, North American Winter Meeting of the Econometric Society, Meeting, New York.

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.A. and W.K. Newey (1991): "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," working paper, MIT Department of Economics.

Hotz, J. and R. Miller (1989): "Estimation of Dynamic Discrete Choice Models," preprint, Carnegie-Mellon University.

Huber, P. (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press.

Ibragimov, I.A. and R.Z. Has'minskii (1981): *Statistical Estimation: Asymptotic Theory*, New York: Springer-Verlag.

Ichimura, H. (1987): "Estimation of Single Index Models," Ph. D. dissertation, Massachusetts Institute of Technology.

Kim, J. and D. Pollard (1989), "Cube Root Asymptotics," *Annals of Statistics,* 18, 191-219.

Klein, R.W. and R.S. Spady (1987), "An Efficient Semiparametric Estimator of the Binary Response Model," manuscript, Bell Communications Research.

Kool, H. (1988): "A Note on Consistent Estimation of Heteroskedastic and Autocorrelated Covariance Matrices," mimeo, Department of Econometrics, Free University of Amsterdam.

Koshevnik, Y.A. and B.Y. Levit (1976): "On a Non-parametric Analouge of the Information Matrix," *Theory of Probability and Applications*, 21, 738-753.

Lorentz, G.G. (1986). *Approximation of Functions*. New York: Chelsea Publishing Company.

Manski, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205-228.

Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models From Binary Panel Data," *Econometrica* 55, 357-362.

Newey, W.K. (1988a): "Adaptive Estimation of Regression Models Via Moment Restrictions," *Journal of Econometrics* 38, 301-339.

Newey, W.K. (1988b): "Asymptotic Equivalence of Closest Moments and GMM Estimators," *Econometric Theory* 4, 336-340.

Newey, W.K. (1990a): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

Newey, W.K. (1990b): "Series Estimation of Regression Functionals," mimeo, Department of Economics, Princeton University.

Newey, W.K. (1991): "Consistency and Asymptotic Normality of Nonparametric Projection Estimators," mimeo, MIT Department of Economics.

Newey, W.K. and K.D. West (1987): "A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 55, 703-708.

Newey, W.K. and P.A. Ruud (1991): "Density Weighted Least Squares Estimation," working paper, MIT Department of Economics.

Pfanzagl, J., and Wefelmeyer (1982): *Contributions to a General Asymptotic Statistical Theory*, New York: Springer-Verlag.

Powell, J.L., J.H. Stock, and T.M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica,* 57, 1403-1430.

Reeds, J.A. (1976): *On the Definition of Von Mises Functionals*, Ph.D. Thesis, Harvard University.

Ritov, Y. and P.J. Bickel (1987): "Achieving Information Bounds in Non and Semiparametric Models," Technical Report No. 116, Department of Statistics, University of California, Berkeley.

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

Robinson, P. (1989): "Hypothesis Testing in Semiparametric and Nonparametric Models for Economic Time Series," *Review of Economic Studies*, 56, 511-534.

Ruud, P.A. (1986): "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157-187.

Stock, J.H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association,* 84, 567-575.

Stoker, T.M. (1990): "Smoothing Bias in Derivative Estimation," MIT, Sloan School of Management, Working Paper 8185-90-EFA.

Stone, C.J. (1985): "Additive regression and other nonparametric models." *Annals of Statistics.* 13 689-705.

Stone, C.J. (1990): "L_2 Rate of Convergence for Interaction Spline Regression," Technical Report No. 268, Berkeley Statistics Department.

Van der Vaart, A. (1991): "On Differentiable Functionals," *Annals of Statistics,* 19, 178-204.

Von Mises (1947): "On the Asymptotic Distributions of Differentiable Statistical Functionals," *Annals of Mathematical Statistics*, 18, 309-348.

White, H. and I. Domowitz (1984): "Nonlinear Regression with Dependent Observations," *Econometrica* 52, 143-161.