The Asymptotic Variance of Semiparametric Estimators

Author(s): Whitney K. Newey

Source: *Econometrica*, Nov., 1994, Vol. 62, No. 6 (Nov., 1994), pp. 1349-1382

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/2951752

Econometrica, Vol. 62, No. 6 (November, 1994), 1349–1382

THE ASYMPTOTIC VARIANCE OF SEMIPARAMETRIC ESTIMATORS

BY WHITNEY K. NEWEY[1]

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 invariance of the limiting distribution with respect to the nonparametric estimator, conditions for nonparametric estimation to have no effect on the asymptotic distribution, and the form of a correction term for the presence of nonparametric projection and density estimators. Examples discussed are quasi-maximum likelihood estimation of index models, panel probit with semiparametric individual effects, average derivatives, and inverse density weighted least squares.

The paper also develops a set of regularity conditions for the validity of the asymptotic variance formula. Primitive regularity conditions are derived for $\sqrt{n}$-consistency and asymptotic normality for functions of series estimators of projections. Specific examples are polynomial estimators of average derivative and semiparametric panel probit models.

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 that depend on nonparametric estimators of functions. Despite the complicated nature of such estimators, the formula is often straightforward to derive, requiring only some calculus. Although the formula is not based on primitive conditions, it should be useful for semiparametric estimators, just as analogous formulae are for parametric estimators. It gives the form of remainder terms, which facilitates specification of primitive conditions. It can also be used to make asymptotic efficiency comparisons and to find an efficient estimator in some class.

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). It uses a semiparametric efficiency bound calculation like that of Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), and Van der Vaart (1991). The innovation of this paper is to calculate this bound for the limit of the estimator under general misspecification, i.e. for the parameter that is estimated nonparametrically. The resulting asymptotic variance formula allows for explicit dependence on conditional expectations or densities, unlike the Gateaux derivative formula for von-Mises estimators. Some of the examples build on previous work on semiparametric

[1] Helpful comments were provided by a co-editor, the referees, M. Arellano, J. Hausman, R. Klein, P. C. B. Phillips, J. Powell, J. Robins, and T. Stoker. Financial support was provided by the NSF, the Sloan Foundation, and BellCore.

estimation, including Bickel, Klaassen, Ritov, and Wellner (1993), Hardle and Stoker (1989), Ichimura (1993), Klein and Spady (1993), Powell, Stock, and Stoker (1989), and others cited below.

The usefulness of the asymptotic variance formula is illustrated in several ways. A number of propositions are derived. One proposition shows that the method of estimating a function (e.g. kernel or polynomial regression) does not affect the asymptotic variance of the estimator. Also, sufficient conditions are given for function estimates to have no effect on the asymptotic variance. Correction terms for the presence of function estimates are also derived. Specific results are given for the case of conditional expectations, or other mean square projections, and for densities.

Regularity conditions for $\sqrt{n}$-consistency and asymptotic normality are formulated. The general discussion is organized around a few "high-level" assumptions. Primitive conditions are given for series estimators of conditional expectations and other projections.

The paper analyzes several example estimators, that are described in Sections 3. A new example is a minimum distance estimator of a panel probit model that allows an individual effect to be correlated with the regressors in a nonparametric way. Other examples are quasi-maximum likelihood estimation of index models, average derivative estimation, and inverse density weighted least squares. The general formula is used to derive the asymptotic variance of each of these estimators. Also, primitive conditions are given for $\sqrt{n}$-consistency and asymptotic normality of series estimators for the panel probit and average derivative examples.

Section 2 gives the formula for the asymptotic variance. Sections 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 conditions for $\sqrt{n}$-consistency and asymptotic normality when an estimator depends on a series estimator of a conditional expectation or other projection. Section 7 gives primitive conditions for the 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, as the variance bound for the *functional* that it nonparametrically estimates. In other words, the formula is the variance bound for the functional that is the limit of the estimator under general misspecification.

To describe the formula, let $z_1, \ldots, z_n$ be i.i.d. data, with (true) distribution F_0 of z_i, and let $\hat{\beta} = \beta_n(z_1, \ldots, z_n)$ denote a $q \times 1$ vector of estimators. Suppose

$\hat{\beta}$ can be associated with a family of distributions and a functional as in

$$(2.1)\qquad \hat{\beta} \to \begin{cases} \mathcal{F} = \{F\}; & \text{general family of distributions of } z, \\ \mu: \mathcal{F} \to \mathbb{R}^q; & \text{if } z_i \text{ has distribution } F \text{ then plim}\left(\hat{\beta}\right) = \mu(F). \end{cases}$$

The word "general" is taken to mean that $\mathcal{F}$ is unrestricted, except for regularity conditions, and allows for general misspecification. This condition will be made more precise below. This equation also specifies that $\mu(F)$ is the limit of $\hat{\beta}$ when z_i has distribution F. Thus, $\mu(F)$ traces out the limits of $\hat{\beta}$ as F varies within the general family $\mathcal{F}$.

The variance formula for $\hat{\beta}$ is the semiparametric bound for estimation of $\mu(F)$, calculated as in Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), and others. Let $\{F_\theta: F_\theta \in \mathcal{F}\}$ denote a one-dimensional subfamily of $\mathcal{F}$, i.e. a path in $\mathcal{F}$, that is equal to the true distribution F_0 when $\theta = 0$. Suppose that F_θ has a density dF_θ and a corresponding score $S(z) = \partial \ln(dF_\theta)/\partial\theta$, where derivatives with respect to θ are evaluated at $\theta = 0$ unless otherwise indicated. Suppose that the set of scores can approximate in mean square any mean zero, finite variance function of z. This is the precise "generality" of $\mathcal{F}$ that is needed for the formula. Let $E[\cdot]$ denote the expectation at the true distribution F_0. The *pathwise derivative* of $\mu(F)$ is a $q \times 1$ vector $d(z)$ with $E[d(z)] = 0$ and $E[\|d(z)\|^2] < \infty$ such that for *every* path,

$$(2.2)\qquad \partial\mu(F_\theta)/\partial\theta = E\left[d(z)S(z)\right].$$

The variance bound for estimation of $\mu(F)$ is $\mathrm{Var}(d(z))$. Thus, the asymptotic variance formula suggested here is the variance of the pathwise derivative of the functional $\mu(F)$ that is estimated under general misspecification.

An example may help fix ideas. The parameter $\beta_0 = \int f_0(z)^2\,dz$, where $f_0(z)$ is the density function of z_i, is important in several contexts, e.g. as discussed in Prakasa Rao (1983). One estimator is $\tilde{\beta} = \sum_{i=1}^n \hat{f}(z_i)/n$, for a nonparametric density estimator $\hat{f}(z)$ of z_i. Suppose z is symmetrically distributed around zero. Then one might hope to improve efficiency by using the antithetic estimate $\hat{f}(-z)$ of the density to, say, form $\hat{\beta} = \sum_{i=1}^n [\hat{f}(z_i) + \hat{f}(-z_i)]/2$. The asymptotic variance can be found by calculating the limit of $\hat{\beta}$ under general misspecification, where z need not be symmetric about zero, and the pathwise derivative of this limit. Let $E_F[\cdot]$ denote the expectation at a distribution F and let $E_\theta[\cdot] = E_{F_\theta}[\cdot]$ for a path F_θ. By an appropriate uniform law of large numbers the limit of $\hat{\beta}$ is $\mu(F) = \int [f(z) + f(-z)]f(z)\,dz/2$. Assuming that differentiation inside the integral is allowed, $\partial\mu(F_\theta)/\partial\theta = \int [\partial f_\theta(z)/\partial\theta]f_0(z)\,dz + \{\int [\partial f_\theta(-z)/\partial\theta]f_0(z)\,dz + \int [\partial f_\theta(z)/\partial\theta]f_0(-z)\,dz\}/2 = E[\{f_0(z) + f_0(-z)\}S(z)] = E[d(z)S(z)]$, for $d(z) = 2\{f_0(z) - \beta_0\}$. Thus, in this example the asymptotic variance formula is $\mathrm{Var}(2f_0(z))$, which is the well known asymptotic variance of $\tilde{\beta}$, so no efficiency improvement results.

The pathwise derivative generalizes the Gateaux derivative formula for von-Mises estimators. The pathwise derivative formula works for estimators that are explicit functions of densities or expectations, where the domain of $\mu(F)$

may only include continuous distributions. The Gateaux derivative formula only applies when the domain of $\mu(F)$ also includes discrete distributions.

A precise justification for the asymptotic variance formula 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, when z_i has distribution F_0,

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)=\sum_{i=1}^{n}\psi(z_i)/\sqrt{n}+0_p(1), \tag{2.3}$$

$$E[\psi(z)]=0, \qquad \operatorname{Var}(\psi(z)) \text{ finite}.$$

This condition is satisfied by many semiparametric estimators, under sufficient regularity conditions. Asymptotic linearity and the central limit theorem imply $\hat{\beta}$ is asymptotically normal with variance $\operatorname{Var}(\psi(z))$.

To state a result some additional regularity conditions are needed. Following Van der Vaart (1991), define the path $\{F_\theta: \theta\in(-\varepsilon,\varepsilon)\subset\mathbb{R}, \varepsilon>0, F_\theta\in\mathscr{F}\}$, to be regular if each distribution is absolutely continuous with respect to the same dominating measure and $S(z)$ satisfies the mean-square derivative condition

$$\int\left[\theta^{-1}\left(dF_\theta^{1/2}-dF_0^{1/2}\right)-\frac{1}{2}S(z)\,dF_0^{1/2}\right]^2 dz\to 0, \quad \text{as} \quad \theta\to 0.$$

Define $\hat{\beta}$ to be a regular estimator of $\mu(F)$ if for any regular path and $\theta_n=0(1/\sqrt{n})$, 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 $\{\theta_n\}_{n=1}^{\infty}$. The following result is a precise statement of the pathwise derivative asymptotic variance formula.

THEOREM 2.1: *Suppose that (i) the set of scores for regular paths is linear; (ii) for any $\varepsilon>0$ and measureable $s(z)$ with $E[s(z)]=0$ and $E[s(z)^2]<\infty$ there is a regular path with score $S(z)$ satisfying $E[|s(z)-S(z)|^2]<\varepsilon$; (iii) $\hat{\beta}$ is asymptotically linear and regular. Then there is $d(z)$ such that equation* (2.2) *is satisfied and $\psi(z)=d(z)$.*

Condition (ii), that the scores can approximate any mean zero function, is the precise version of the "generality" property of $\mathscr{F}$. It means that set of scores associated with the family $\mathscr{F}$ is essentially unrestricted (except for the usual mean zero property). Also, regularity of $\hat{\beta}$ is the precise condition that specifies that $\hat{\beta}$ is a nonparametric estimator of $\mu(F)$. Regularity implies that the mean of the limiting distribution is zero for all local alternatives, a local unbiasedness condition.

The thing that seems to be novel here is the idea of calculating the bound for the functional $\mu(F)$ that is nonparametrically estimated by $\hat{\beta}$. The main innovation of this theorem is that it is based on the fundamental $\mu(F)$ that is estimated under general misspecification. The fact that asymptotic linearity and regularity imply pathwise differentiability follows by Van der Vaart (1991,

Theorem 2.1), and the fact that condition (ii) 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 (1990).

Another way to describe equation (2.2), using the Riesz representation theorem, is that $\partial\mu(F_\theta)/\partial\theta$ is a linear mapping on the set of scores for regular paths that is mean square continuous. If this mapping is linear but not mean square continuous then no such $d(z)$ will exist. Consequently, as shown by Van der Vaart (1991), there will be no $\sqrt{n}$-consistent, regular estimator of $\mu(F)$. For example, the value of a density function at a point does not have a mean-square continuous derivative, and neither does the limit of Manski's (1975) maximum score estimator. The pathwise derivative cannot be used to find 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 pathwise derivative can also be used to calculate the influence function when observations are dependent and stationary. If the limit of $\hat{\beta}$ under general misspecification is determined by the marginal distribution of a single observation z, then the pathwise derivative of the limit should equal the influence function. Because dependence allows flexibility in the specification of z, which could include several more primitive observations, this result applies to many estimators with dependent observations. In this case, allowing for general misspecification would overturn restrictions implied by stationarity, such as equality of marginal distributions for each observation. Regularity conditions for a dependent observations version of Theorem 2.1 are difficult to formulate, but the validity of the pathwise derivative formula is apparent for the more primitive conditions discussed in Section 5.

The point of Theorem 2.1 is to give a justification for the pathwise derivative formula, rather than an approach to showing asymptotic normality. A better approach is to solve equation (2.2) for the pathwise derivative, as a candidate for the influence function, and then formulate regularity conditions for the remainder $\sqrt{n}(\hat{\theta}-\theta_0)-\sum_{i=1}^n\psi(z_i)/\sqrt{n}$ to be small. The formula is a very important part of this approach, because it provides the form of the remainder. This approach, with formal calculation followed by regularity conditions, is similar to that used in parametric asymptotic theory (e.g. for Edgeworth expansions).

The rest of the paper includes both pathwise derivative calculations and regularity conditions. In Sections 3 and 4, pathwise derivative calculations are used to derive some propositions about semiparametric estimators. These results are labeled as "propositions" rather than "theorems" because they do not include regularity conditions. Instead, these propositions give solutions to equation (2.2) obtained using the chain rule of calculus, differentiation under integrals, integration, and $\partial\int a(z)\,dF_\theta/\partial\theta = E[a(z)S(z)]$ for $a(z)$ with finite mean square, without specifying conditions under which these calculus rules can be applied. Section 5–7 give regularity conditions for the validity of the resulting formulae.

3. SEMIPARAMETRIC *M*-ESTIMATORS

The rest of the paper will focus on a class of semiparametric m-estimators, obtained from moment conditions that can depend on estimated functions. Let h denote a function, that can depend on the parameters β and the data z. The arguments of the function h are suppressed for notational convenience. Let $m(z,\beta,h)$ be a vector of functions with the same dimension as β. Here $m(z,\beta,h)$ can depend on the entire function h, rather than just its value at particular points, so $m(z,\beta,h)$ is a vector of functionals. Suppose that $E[m(z,\beta_0,h_0)]=0$ for the true values β_0 and h_0. Let $\hat{h}$ denote an estimator of h. A semiparametric m-estimator is one that solves a moment equation of the form

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

The general idea here is that $\hat{\beta}$ is obtained by a procedure that "plugs-in" an estimated function $\hat{h}$.

An early and important example is the Buckley and James (1979) estimator for censored regression. Other examples are Robinson's (1988) semiparametric regression estimator, Powell, Stock, and Stoker's (1989) weighted average derivative estimator, and Ahn and Manski's (1993) estimator for dynamic discrete choice models. Additional examples are useful for illustrating the results of this paper.

EXAMPLE 1: *Quasi-maximum Likelihood for a Conditional Mean Index*. The conditional mean index model is one where $E[y|x]=\tau(v(x,\beta_0))$ for a known function $v(x,\beta)$ and an unknown function $\tau(\cdot)$. Let $\hat{h}(x,\beta)$ be a nonparametric estimator of $E[y|v(x,\beta)]$, such as a kernel estimator. An estimator of β_0 suggested by Ichimura (1993) minimizes $\sum_{i=1}^{n}[y_i-\hat{h}(x_i,\beta)]^2$. When y_i is binary Klein and Spady (1993) have suggested maximizing $\sum_{i=1}^{n}\{y_i\ln[\hat{h}(x_i,\beta)]+(1-y_i)\ln[1-\hat{h}(x_i,\beta)]\}$. A generalization of these estimators is a quasi-maximum likelihood estimator (QMLE) for an exponential family. The QMLE can have higher efficiency than Ichimura's least squares estimator when heteroskedasticity is present, because it is asymptotically equivalent to weighted least squares. Also, the estimator will be efficient when the true distribution has the exponential form. To describe the estimator, let $l(u,\nu)=\exp(A(\nu)+B(u)+C(\nu)u)$ be a linear exponential density, with mean ν. Consider an estimator $\hat{\beta}$ that maximizes $\sum_{i=1}^{n}\ln l(y_i,\hat{h}(x_i,\beta))$. The first order conditions for this estimator make it a special case of equation (3.1), with

$$(3.2)\qquad m(z,\beta,h)=\left[A_\nu(h(x,\beta))+C_\nu(h(x,\beta))\,y\right]\partial h(x,\beta)/\partial\beta.$$

EXAMPLE 2: *Panel Probit with Semiparametric Individual Effects*. 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 $x=(x_1',x_2')'$, $E[y_t|x]=\Phi([x_t'\gamma_0+\rho(x)]/(\sigma_0)^{t-1})$, $\rho(x)$ is an unknown function, and Φ denotes the standard normal CDF. This is a binary

panel model with $y_t = 1(x_t\gamma_0 + \alpha + \varepsilon_t > 0)$ for an individual effect α, and the conditional distribution of $\alpha + \varepsilon_t$ given x is $N(\rho(x),(\sigma_0)^{t-1})$. 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

$$\Phi^{-1}(E[y_1|x]) = \sigma_0\Phi^{-1}(E[y_2|x]) + (x_1 - x_2)'\gamma_0. \tag{3.3}$$

This implication can be used to construct a semiparametric minimum distance estimator by replacing the conditional expectations with nonparametric estimators $\hat{h}_t(x) = \hat{E}[y_t|x]$, such as a series estimator, and choosing $\hat{\gamma}$ and $\hat{\sigma}$ from the least squares regression of $\Phi^{-1}(\hat{h}_1(x_t))$ on $x_{1t} - x_{2i}$ and $\Phi^{-1}(\hat{h}_2(x_t))$. This estimator has the form in equation (3.1), with $\beta = (\gamma', \sigma)'$ and

$$\begin{aligned} m(z,\beta,h) = &\left[x_1 - x_2, \Phi^{-1}(h_2(x))\right]' \\ &\times\left\{\phi^{-1}(h_1(x))\text{-}\sigma\Phi^{-1}(h_2(x)) - (x_1 - x_2)'\gamma\right\}. \end{aligned} \tag{3.4}$$

EXAMPLE 3: *Average Derivatives*. The average derivative of a conditional expectation $h_0(x) = E[y|x]$ is $E[\partial h_0(x)/\partial x]$. As discussed in Stoker (1986), this functional is useful for estimating scaled coefficients in conditional mean index models with $v(x,\beta_0) = x'\gamma_0$, where $E[\partial h_0(x)/\partial x] = E[\tau_v(x'\gamma_0)]\gamma_0$ and $\tau_v(v) = d\tau(v)/dv$. Also, average derivatives are useful summary measures that have some economic applications, as in Hardle, Hildenbrand, and Jerison (1991). They can be estimated by replacing $h_0(x)$ by a differentiable nonparametric estimator $\hat{h}(x)$ and by replacing the expectation by a sample average to form $n^{-1}\sum_{i=1}^n \partial\hat{h}(x_i)/\partial x$. This estimator is a special case of equation (3.1) with

$$m(z,\beta,h) = \partial h(x)/\partial x - \beta. \tag{3.5}$$

This example illustrates the correction terms when derivatives are present.

EXAMPLE 4: *Inverse Density Weighted Least Squares*. Let $w(x) = r((x - \zeta)'\Omega(x - \zeta))$ be an elliptically symmetric density function, where ζ is a vector and Ω a positive definite matrix. Also, let $\hat{h}(x_t)$ be an estimator of the density of x, such as a kernel estimator. As shown by Ruud (1986), the weighted least squares estimator $\hat{\beta} = [\sum_{i=1}^n w(x_i)\hat{h}(x_i)^{-1}x_i x_i']^{-1}\sum_{i=1}^n w(x_i)\hat{h}(x_i)^{-1}x_i y_i$ will be consistent up to scale, for the coefficients γ_0 of an index model $E[y|x] = \tau(x'\gamma_0)$. This example is a special case of equation (1.1) with

$$m(z,\beta,h) = w(x)h(x)^{-1}x(y - x'\beta). \tag{3.6}$$

This estimator will be used to illustrate the correction term for density estimates. Asymptotic normality and $\sqrt{n}$-consistency for $\hat{h}(x)$ a kernel estimator, are shown in Newey and Ruud (1991).

It is possible, at the level of generality of equation (3.1), to derive a number of propositions. These results will have many applications, such as the examples

given above. To use the pathwise derivative formula in this derivation, it is necessary to identify the functional that is nonparametrically estimated by $\hat{\beta}$. Let $h(F)$ denote the limit of $\hat{h}$ when z has distribution F. By the usual method of moments reasoning, the limit $\mu(F)$ of $\hat{\beta}$ for a general F should be the solution to

$$(3.7) \qquad E_F[m(z,\mu,h(F))] = 0.$$

That is, equation (3.1) sets $\hat{\beta}$ so that sample moments are zero, and the sample moments have a limit of $E_F[m(z,\beta,h(F))]$ (by the law of large numbers and $h(F)$ equal to the limit of $\hat{h}$), 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 $h(F)$, and not on the particular form of the estimator $\hat{h}$. Thus, different nonparametric estimators of the same functions should result in the same asymptotic variance. For example, this reasoning explains the asymptotic equivalence of various derivative estimators found by Stoker (1991), as well as the asymptotic equivalence between series and kernel average derivative estimators found in Section 7.

PROPOSITION 1: *The asymptotic variance of semiparametric estimators depends only on the function that is nonparametrically estimated, and not on the type of estimator.*

As discussed in Section 2, this result and the others in Section 3 and 4 are based on pathwise derivative calculations, rather than primitive regularity conditions of Section 5.

To obtain more results, it is useful to be more specific about the form of the pathwise derivative. For a path $\{F_\theta\}$, let $h(\theta) = h(F_\theta)$. Here, $\mu(F_\theta)$ will satisfy the population moment equation

$$(3.8) \qquad E_\theta[m(z,\mu,h(\theta))] = 0.$$

Let $m(z,h) = m(z,\beta_0,h)$. Differentiation under the integral gives

$$\partial E_\theta[m(z,h_0)]/\partial\theta = \int m(z,h_0)[\partial\, dF_\theta/\partial\theta]\, dz = E[m(z,h_0)S(z)].$$

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

$$\partial E_\theta[m(z,h(\theta))]/\partial\theta = E[m(z,h_0)S(z)] + \partial E[m(z,h(\theta))]/\partial\theta.$$

Assuming $M \equiv \partial E[m(z,\beta,h_0)]/\partial\beta|_{\beta_0}$ is nonsingular, by the implicit function theorem

$$\partial\mu(F_\theta)/\partial\theta = -M^{-1}\{E[m(z,h_0)S(z)] + \partial E[m(z,h(\theta))]/\partial\theta\}.$$

The first term is already in an outer product form, so that the pathwise derivative can be found by putting the second term in singular form. Suppose

there is a $\alpha(z)$ such that $E[\alpha(z)] = 0$ and

$$(3.9) \qquad \partial E\big[m(z, h(\theta))\big]/\partial\theta = E\big[\alpha(z)S(z)\big].$$

Then, moving $-M^{-1}$ inside the expectation, it follows that the pathwise derivative is $d(z) = -M^{-1}\{m(z, h_0) + \alpha(z)\}$, so that by Theorem 2.1 the influence function of $\hat{\beta}$ is

$$(3.10) \qquad \psi(z) = -M^{-1}\{m(z, \beta_0, h_0) + \alpha(z)\}.$$

This influence function has an interesting structure. The leading term $-M^{-1}m(z, \beta_0, h_0)$ is the usual Huber (1967) formula for the influence function of an m-estimator with moment functions $m(z, \beta, h_0)$, i.e. the formula that would be obtained if estimation of h were ignored. Thus, the solution to equation (3.9) is an adjustment term for the estimation of h_0. Solving equation (3.9) is therefore the essential step in discovering how the estimation of h affects the asymptotic variance. This solution can be interpreted either as the pathwise derivative of the functional $-M^{-1}E[m(z\,\beta_0, h(F))]$ or as the influence function of $-M^{-1}\int m(z, \beta_0, \hat{h})\, dF_0(z)$.

If h contains more than one component, e.g. both a density and a conditional expectation, then the adjustment term can be calculated as the sum of terms for individual components. To be specific, suppose $h = (h_1, \ldots, h_J)$. The chain rule gives $\partial E[m(z, h_1(\theta), \ldots, h_J(\theta))]/\partial\theta = \sum_{j=1}^{J} \partial E[m(z, h_1(\theta_0), \ldots, h_j(\theta), \ldots, h_J(\theta_0))]/\partial\theta$. Then if for each j there is $\alpha_j(z)$ with $\partial E[m(z, h_1(\theta_0), \ldots, h_j(\theta), \ldots, h_J(\theta_0))]/\partial\theta = E[\alpha_j(z)S(z)]$, it follows that equation (3.9) is satisfied with $\alpha(z) = \sum_{j=1}^{J} \alpha_j(z)$. This property is useful, because the various results to follow can be combined to derive the adjustment term when $\hat{h}$ consists of several types of estimators, as the sum of separate terms.

It is useful to know when the adjustment term is zero. In such cases, it should not be necessary to account for the presence of $\hat{h}$, i.e. $\hat{h}$ can be treated as if it were equal to h_0, 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 $\hat{h}$ has a limit that maximizes the population value of the same function. To be specific, suppose that there is a function $q(z, \beta, h)$ and a set of functions $\mathscr{H}(\beta)$, possibly depending on β but not on the distribution F of z, such that

$$(3.11) \qquad m(z, \beta, h) = \partial q(z, \beta, h)/\partial\beta,$$

$$h(F) = \operatorname{argmax}_{\tilde{h} \in \mathscr{H}(\beta)} E_F\big[q(z, \beta, \tilde{h})\big].$$

Note here that since q depends on β this parameter must also be one of the arguments of h, although this argument is still suppressed in the notation. The interpretation of equation (3.11) is that $m(z, \beta, h)$ are the first order conditions for a maximum of the function q and that $h(F)$ maximizes the expected value of the same function, i.e. that $h(F)$ has been "concentrated out." Then for any

parametric model F_θ, since $h(\theta) = h(F_\theta)$, it follows that $E[q(z, \beta, h(\theta))]$ is maximized at $\theta = 0$. The first order conditions for this maximization are $\partial E[q(z, \beta, h(\theta))]/\partial\theta = 0$, *identically* in β. Differentiating again with respect to β,

$$(3.12)\quad 0 = \partial^2 E[q(z, \beta, h(\theta))]/\partial\theta\partial\beta = \partial E[\partial q(z, \beta, h(\theta))/\partial\beta]/\partial\theta$$
$$= \partial E[m(z, \beta, h(\theta))]/\partial\theta.$$

Evaluating this equation at β_0, it follows that $\alpha(z) = 0$ will solve equation (3.9), and hence the adjustment term is zero. Summarizing, we have the following propositions.

PROPOSITION 2: *If equation* (3.10) *is satisfied, then the estimation of h can be ignored in calculating the asymptotic variance, i.e. it is the same as if* $\hat{h} = h_0$.

EXAMPLE 1 CONTINUED: It is straightforward to show that equation (3.11) is satisfied in this example, so that estimation of h_0 has no effect on the asymptotic variance. Let $\mathscr{H}(\beta)$ be the set of functions of $v(x, \beta)$. Then, e.g. as in Gourieroux, Monfort, and Trognon (1984), $h(x, \beta, F) = E_F[y|v(x, \beta)]$ maximizes $E_F[\ln l(y, \tilde{h}(x, \beta))]$ over $\tilde{h} \in \mathscr{H}(\beta)$. It then follows from Proposition 10 that the influence function will be $-M^{-1}m(z, \beta_0, h_0)$. This influence function can be given a more explicit expression. Let $h(x, \beta_1, \beta_2) = E[\tau(v(x, \beta_0) - v(x, \beta_1) + v(x, \beta_2))|v(x, \beta_2)]$. Then by the chain rule $h_\beta = \partial h_0(x, \beta)/\partial\beta = \partial h(x, \beta_1, \beta_0)/\partial\beta_1 + \partial h(x, \beta_0, \beta_2)/\partial\beta_2 = \tau_v(v)\{v_\beta - E[v_\beta|v]\}$ for $v = v(x, \beta_0)$ and $v_\beta = \partial v(x, \beta_0)/\partial\beta$. Also, $E[A_\nu(\tau) + C_\nu(\tau)y|x] = 0$ bystandard exponential family results, for $\tau = \tau(v(x, \beta_0))$, so that

$$(3.13)\quad \psi(z) = -M^{-1}[A_\nu(\tau) + C_\nu(\tau)y]h_\beta,$$
$$M = E[\{A_{\nu\nu}(\tau) + C_{vv}(\tau)y\}h_\beta h_\beta'].$$

A result related to Proposition 2 is efficiency of semiparametric maximum likelihood estimation. Consider a semiparametric model (density) $f(z|\beta, h)$, where β is a Euclidean parameter and h is an unknown function. Suppose that h_0 does not depend on β, and that $S_\beta = \partial \ln f(z|\beta, h_0)/\partial\beta$ is the score for β. Let $\hat{h}(\beta)$ be an estimator that is allowed to depend on β. Suppose that under a distribution F that is general (i.e. it need not have a density of the form $f(z|\beta, h)$), $\hat{h}(\beta)$ has a limit $h(\beta, F) = \operatorname{argmax}_{h \in \mathscr{H}(\beta)} E_F[\ln f(z|\beta, h)]$ that maximizes the expected log-likelihood. Consider an estimator $\hat{\beta} = \operatorname{argmax}_\beta \sum_{i=1}^n \ln f(z_i|\beta, \hat{h}(\beta))$. Since the nonparametric component $\hat{h}$ maximizes the limiting log-likelihood, $\hat{\beta}$ is a semiparametric version of a "concentrated" or "profile" maximum likelihood estimator, in the limit. This is a semiparametric m-estimator with $m(z, \beta, h) = \partial \ln f(z|\beta, h(\beta))/\partial\beta \equiv \partial q(z, \beta, h)/\partial\beta$, so Proposition 2 implies that estimation of $h(\beta, F)$ has no effect on the asymptotic variance. The influence function of $\hat{\beta}$ will then be

$$(3.14)\quad \psi(z) = M^{-1}\partial \ln f(z|\beta, h_0(\beta))/\partial\beta,$$
$$h_0(\beta) = \operatorname{argmax}_{h \in \mathscr{H}(\beta)} E[\ln f(z|\beta, h)].$$

Furthermore, it can be shown that the variance of $\psi(z)$ is the semiparametric variance bound, so that β is efficient. By the chain rule,

$$(3.15)\quad m \equiv m(z,\beta_0,h_0) = \partial \ln f(z|\beta, h_0(\beta))/\partial\beta$$
$$= \partial \ln f(z|\beta,h_0)/\partial\beta + \partial \ln f(z|\beta_0, h_0(\beta))/\partial\beta.$$

The first term is S_β, the score for β. The second term is in the set of scores for parametric models for h, referred to as the "tangent set." Furthermore, for any parametric model $\{h(\theta)\}$ for h with corresponding score $S_h = \partial \ln f(z|\beta_0, h(\theta))/\partial\theta$, by differentiation of the identity

$$0 \equiv \int m(z,\beta_0,h(\theta))\, f(z|\beta_0,h(\theta))\, dz$$

with respect to θ,

$$(3.16)\quad 0 = E[mS_h'] + \partial E\big[m(z,\beta_0,h(\theta))\big]/\partial\theta = E[mS_h'],$$

where $\partial E[m(z,\beta_0,h(\theta))]/\partial\theta = 0$ holds by equation (3.12). Hence, m is orthogonal to the tangent set. It follows that $\partial \ln f(z|\beta_0,h_0(\beta))/\partial\beta$ is the projection of S_β on the tangent set, so that m is the residual from the projection of S_β on the tangent set. Efficiency of $\hat\beta$ then follows from well known results on semiparametric models (e.g. Newey (1990)), that characterize the optimal moment function as this projection residual. This is a semiparametric version of the R. A. Fisher calculation of efficiency of parametric maximum likelihood although regularity conditions for efficiency are more complicated in the semiparametric case. In relation to Severini and Wong (1992), this result shows that choosing $\hat h(\beta)$ so that its limit maximizes the expected log-likelihood under general misspecification gives a "least favorable curve."

EXAMPLE 1 CONTINUED: Consider a model that has an exponential conditional density with a mean that depends only on $v(x,\beta_0)$. Then the QMLE of Example 1 is a special case of semiparametric maximum likelihood, with nonparametric component h equal to a function of $v(x,\beta)$. Also, $E_F[y|v(x,\beta)]$ maximizes the expected log-likelihood. Efficiency of the QMLE then follows by the reasoning in the previous paragraph.

There is another, more direct condition for estimation of the nuisance function to not affect the asymptotic variance. To formulate this condition, suppose that $m(z,h)$ depends on h only through its value $h(v)$ at a subvector v of z, i.e. $m(z,h) = m(z,h(v))$ where the last function depends on a real vector argument in $m(z,h)$. Each of the examples has this property. Let $h(v,\theta)$ denote the limiting value of $\hat h(v,\beta_0)$ for a path. For $D(z) = \partial m(z,\beta_0,h)/\partial h|_{h=h_0(v)}$, differentiation gives

$$(3.17)\quad \partial E\big[m(z,h(\theta))\big]/\partial\theta = E\big[D(z)\partial h(v,\theta)/\partial\theta\big] = \partial E\big[D(z)h(v,\theta)\big]/\partial\theta.$$

If this derivative is zero for all $h(v,\theta)$, then $\alpha(z) = 0$ will solve equation (3.9), and the adjustment term is zero. One simple condition for this is that

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

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

This condition can be checked by straightforward calculation, unlike Proposition 2, which requires finding $q(z, \beta, h)$ satisfying equation (3.11).

4. FUNCTIONS OF MEAN-SQUARE PROJECTIONS AND DENSITIES

In this section, the form of the correction term is derived when h is a conditional expectation or other mean-square projection, such as additive or partially linear regressions, and where h is a density. It is useful to separate this derivation into two parts. The first part is a linearization of the function $m(z, h) = m(z, \beta_0, h)$. Assume that there is a function $D(z, h)$ such that

$$(4.1) \qquad \partial E[m(z, h(\theta))]/\partial\theta = \partial E[D(z, h(\theta))]/\partial\theta, \quad D(z, h) \text{ is linear in } h.$$

For example, when $m(z, h)$ depends on h only through its value $h(v)$, then by equation (3.17),

$$(4.2) \qquad D(z, h) = D(z)h(v).$$

More generally, $D(z, h)$ will be a functional derivative of $m(z, h)$ with respect to h. It can typically be calculated by formal differentiation, similarly to equation (3.17). Regularity conditions for this calculation are discussed in Section 5.

The second part of the derivation is to develop an integral representation for $E[D(z, h)]$. The most convenient representation is different for projections and densities, and so will be described separately. To describe the result when h is a mean-square projection, let y be a random variable with finite second moment and x an $r \times 1$ vector. Let $\mathscr{G}$ 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 $\mathscr{G}$ (for the inner product $E[yg(x)]$), that is $g(x) = \text{argmin}_{\tilde{g} \in \mathscr{G}} E[\{y - \tilde{g}(x)\}^2]$. Then the first step nonparametric estimator will be a projection when $h = g$.

The simplest nonparametric example of a projection is $g(x) = E[y|x]$, where $\mathscr{G}$ is all measurable functions of x with finite mean-square, i.e. $\mathscr{G} = \{\tilde{g}: E[\tilde{g}(x)^2] < \infty\}$. A more general example is a projection on

$$(4.3) \qquad \mathscr{G} = \left\{ \sum_{l=1}^{L} \tilde{g}_l(\tilde{x}_l) + \tilde{x}'_{L+1}\eta \right\},$$

where each $\tilde{x}_l$ is 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.

To derive the form of the correction term for a projection, suppose that there is $\delta(x) \in \mathscr{G}$ such that

$$(4.4) \qquad E[D(z,\tilde{g})] = E[\delta(x)\tilde{g}(x)] \quad \text{for all } \tilde{g} \in \mathscr{G},\ E\left[\|\delta(x)\|^2\right] < \infty.$$

By the Riesz representation theorem, such a $\delta(x)$ will exist if the functional $E[D(z,\tilde{g})]$ is mean-square continuous in $\tilde{g}$. Let $g(x,\theta) = \operatorname{argmin}_{\tilde{g}\in\mathscr{G}} E_\theta[\{y - \tilde{g}(x)\}^2]$ denote the projection of y on $\mathscr{G}$ for a path. By $\delta(x) \in \mathscr{G}$, $E_\theta[\delta(x)g(x,\theta)] = E_\theta[\delta(x)y]$. Then by the chain rule,

$$\begin{aligned}(4.5) \qquad \partial E[D(z,h(\theta))]/\partial\theta &= \partial E[\delta(x)g(x,\theta)]/\partial\theta \\ &= \partial E_\theta[\delta(x)g(x,\theta)]/\partial\theta - \partial E_\theta[\delta(x)g_0(x)]/\partial\theta \\ &= \partial E_\theta[\delta(x)\{y - g_0(x)\}]/\partial\theta \\ &= E[\delta(x)\{y - g_0(x)\}S(z)].\end{aligned}$$

Then by equation (4.1), it follows that equation (3.9) is satisfied with $\alpha(z) = \delta(x)[y - g_0(x)]$, giving the next result.

PROPOSITION 4: *If equations* (4.1) *and* (4.4) *are satisfied, then the correction term is* $\alpha(z) = \delta(x)[y - g_0(x)]$.

By the Riesz representation theorem, the outer product condition in equation (4.4) is necessary for mean-square continuity of $E[D(z,g)]$. Furthermore, using arguments like those of Van der Vaart (1991), it will follow that mean-square continuity is necessary for existence of a $\sqrt{n}$-consistent estimator. Thus, estimation of $g_0(x)$ will affect the convergence rate of $\hat{\beta}$ unless equation (4.4) is satisfied, so that this equation and Proposition 4 characterize the form of the correction term for mean-square projections.

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. One interesting case is where $m(z,g)$ depends on the value of a derivative of g at an observed variable. 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

$$(4.6) \qquad \partial^\lambda g(x) = \partial^{|\lambda|} g(x)/\partial x_1^{\lambda_1} \circ\circ\circ \partial x_r^{\lambda_r}.$$

Suppose that $m(z,g) = m(z,\partial^\lambda g(v))$, so that $D(z,g) = D(z)\partial^\lambda g(v)$ for $D(z) = \partial m(z,h)/\partial h|_{h=\partial^\lambda g_0(v)}$ as in equation (3.17). Let $\bar{D}(v) = E[D(z)|v]$ and $f_v(v)$ and $f_0(x)$ be the densities of v and x respectively, with respect to the same

dominating measure. Assuming $f_v(v)$ 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, then repeated integration by parts gives $E[D(z, \tilde{g})] = \int \bar{D}(v)\partial^\lambda \tilde{g}(v) f_v(v)\, dv = (-1)^{|\lambda|}\int \partial^\lambda[\bar{D}(v)f_v(v)]\tilde{g}(v)\, dv = (-1)^{|\lambda|}\int \partial^\lambda[\bar{D}(v)f_v(v)]|_{v=x}\tilde{g}(x)\, dx = E[\delta(x)\tilde{g}(x)]$, where Π is the projection and

$$\delta(x) = (-1)^{|\lambda|}\Pi\left(f_0(x)^{-1}\partial^\lambda\left[\bar{D}(v)f_v(v)\right]_{v=x}|\mathscr{S}\right). \tag{4.7}$$

PROPOSITION 5: *If $h(v) = \partial^\lambda g(x)|_{x=v}$, 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 $\bar{D}(v)$ are continuously differentiable to order $|\lambda|$ in $\bar{x}$, the support of $\bar{x}$ is a convex set with nonempty interior, and for each $\tilde{\lambda} \leqslant \lambda$, $\partial^{\tilde{\lambda}} f_v(v)$ is zero on the boundary of the support of $\bar{x}$ and $f_x(x)^{-1}\partial^{\tilde{\lambda}}[\bar{D}(v)f_v(v)]_{v=x}$ has finite second moment, then the correction term is $\delta(x)[y - g_0(x)]$ for $\delta(x)$ in equation* (4.7).

EXAMPLE 2 CONTINUED: Proposition 5 can easily be applied to derive the influence function of the panel probit estimator. In this example $m(z, \beta_0, h_0) = 0$, so that the adjustment term is the only source of variation. As described in Section 3, one can derive the adjustment terms separately for each $g_t(x) = h_t(x)$ $(t = 1, 2)$, and then take the adjustment to be the sum of the separate components. Let $J(x) = [x_1 - x_2, \Phi^{-1}(g_{20}(x))]'$. Then $\partial m(z, \beta_0, g_1, g_2)/\partial g_t|_{g_{10}(x), g_{20}(x)} \equiv D_t(z) = J(x)(-\sigma_0)^{t-1}\phi(\Phi^{-1}(g_{t0}(x)))^{-1}$, so by Proposition 5, the influence function is

$$\psi(z) = -M^{-1}[\alpha_1(z) + \alpha_2(z)], \quad M = -E[J(x)J(x)'], \tag{4.8}$$

$$\alpha_t(z) = J(x)(-\sigma_0)^{t-1}\phi\left(\Phi^{-1}(g_{t0}(x))\right)^{-1}[y_t - g_{t0}(x)].$$

The asymptotic variance of $\hat{\beta}$ is thus $M^{-1}\mathrm{Var}(\alpha_1(z) + \alpha_2(z))M^{-1}$.

EXAMPLE 3 CONTINUED: A generalization of the average derivative of a conditional expectation is the average derivative of a projection on $\mathscr{S}$, where $m(z, \beta, h) = \partial g(x)/\partial x - \beta$ and $h = g$. This generalization would be useful when the average derivative is used as a summary measure and a projection is used to avoid the difficulty of estimating a high-dimensional function. Proposition 5 applies to this example, with $v = x$, and $D(v) = 1$, and $\delta(x) = -\Pi(f_x(x)^{-1}\partial f_x(x)/\partial x|\mathscr{S})$. Hence, the conclusion of Proposition 5 gives an influence function

$$\psi(z) = \partial g_0(x)/\partial x - \beta_0 - \Pi\left(f_x(x)^{-1}\partial f_x(x)/\partial x|\mathscr{S}\right)[y - g_0(x)]. \tag{4.9}$$

This influence function is like that of Hardle and Stoker (1989), except that $f_x(x)^{-1}\partial f_x(x)/\partial x$ has been replaced by its projection on the set of functions $\mathscr{S}$. When $g_0(x) = E[y|x]$ and $\mathrm{Var}(y|x)$ is constant the projection derivative esti-

mator can be more efficient than the expectation estimator. In this case the first and last terms in equation (4.9) are orthogonal and the variance of the last term is proportional to the second moment of the projection of the score, which is smaller than the second moment of the score itself.

A correction term for density estimation can be derived when $h = f$ and $f(x)$ is a density function for x, with respect to some measure. Suppose that there is $\delta(x)$ such that for every possible density function $\tilde{f}$,

$$(4.10) \qquad E\left[D(z,\tilde{f})\right] = \int \delta(x)\tilde{f}(x)\,dx,$$

so that $\partial E[D(z, f(\theta))]/\partial\theta = \partial E_\theta[\delta(x)]/\partial\theta = E[\delta(x)S(z)]$. Then by equation (4.1), it follows that equation (3.9) is satisfied with $\alpha(z) = \delta(x) - E[\delta(x)]$, giving this result:

PROPOSITION 6: *If equations* (4.1) *and* (4.10) *are satisfied then the correction term is* $\alpha(z) = \delta(x) - E[\delta(x)]$.

Existence of such a $\delta(x)$ will follow from the Riesz representation theorem if $\int f(x)^2\,dx$ is finite and $E[D(z,\tilde{f})]$ can be extended to a linear functional on the Hilbert space of square integrable (dx) functions that is continuous. Continuity of $E[D(z,f)]$ in this square integrable sense appears to be important for $\sqrt{n}$-consistency, although it is beyond the scope of this paper to develop arguments to that effect.

The density correction term can be computed when $m(z,h)$ depends on a derivative of a density evaluated at a variable v. Suppose that $m(z,f) = m(z,\partial^\lambda f(v))$, and let $D(z) = \partial m(z,h)/\partial h|_{h=\partial^\lambda f_0(v)}$ and $\bar{D}(v) = E[D(z)|v]$. Integration by parts gives $E[D(z,\tilde{f})] = \int \bar{D}(v)\partial^\lambda \tilde{f}(v) f_v(v)\,dv = (-1)^{|\lambda|}\int \partial^\lambda[\bar{D}(v)F_v(v)]|_{v=x}\tilde{f}(x)\,dx = \int \delta(x)\tilde{f}(x)\,dx$, for

$$(4.11) \qquad \delta(x) = (-1)^{|\lambda|}\partial^\lambda\left[\bar{D}(v)f_v(v)\right]|_{v=x}.$$

The correction term is $\alpha(z) = \delta(x) - E[\delta(x)]$ for this $\delta(x)$, although to save space no formal proposition is given.

EXAMPLE 4 CONTINUED: The correction term just derived can easily be applied to density weighted least squares. Note that $v = x$, and for a scalar f, $\partial f^{-1}w(x)[y - x'\beta_0]/\partial f = -f^{-2}w(x)x[y - x'\beta_0]$, so that $\bar{D}(v) = -f_0(x)^{-2}w(x)x\{E[y|x] - x'\beta_0\}$. Therefore, $\delta(x) = -f_0(x)^{-1}w(x)x\{E[y|x] - x'\beta_0\}$. Noting that $E[\delta(x)] = E[m(z,\beta_0,h_0)] = 0$, the influence function for $\hat{\beta}$ is

$$(4.12) \qquad \psi(z) = -M^{-1}h_0(x)^{-1}w(x)x\{y - E[y|x]\}, \quad M = \int w(x)xx'\,dx.$$

Proposition 4 and equation (4.11) can be combined to derive the correction terms for density weighted conditional expectations in Powell, Stock, and Stoker (1989) and Robinson (1989), where estimators of $E[y|x]f_0(x)$ or its derivatives

are present. The correction term can be derived by adding separate terms for the conditional expectation and density in the way described in Section 3.

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

5. REGULARITY CONDITIONS

The rest of the paper will give regularity conditions for semiparametric generalized method of moments estimators. To describe the estimator, let $m(z,\beta,h)$ be a vector of functions as previously discussed, except that $m(z,\beta,h)$ may have more elements than β. Suppose that the moment condition $E[m(z_i,\beta_0,h_0)]=0$ is satisfied. Let $\hat{h}$ denote an estimator of the true function h_0, $\hat{m}_n(\beta)=\sum_{i=1}^n m(z_i,\beta,\hat{h})/n$, and $\hat{W}$ a positive semi-definite matrix. A two-step GMM estimator with a nonparametric first step is

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

Conditions for $\sqrt{n}$-consistency and asymptotic normality of this estimator are given in the rest of the paper.

Several intermediate results are useful for showing asymptotic normality, including uniform convergence in probability of $\hat{m}_n(\beta)$ and $\partial\hat{m}_n(\beta)/\partial\beta$ and asymptotic normality of $\sqrt{n}\,\hat{m}_n(\beta_0)$. Suppose that $\hat{\beta}$ is consistent and for any $\bar{\beta}\xrightarrow{p}\beta_0$,

$$\sqrt{n}\,\hat{m}_n(\beta_0)\xrightarrow{d}N(0,\Omega),\qquad \hat{W}\xrightarrow{p}W,$$
$$\partial\hat{m}_n(\bar{\beta})/\partial\beta\xrightarrow{p}M=E[\partial m(z,\beta_0,h_0)/\partial\beta].$$

Then by the usual expansion of $\hat{m}_n(\hat{\beta})$ around β_0, in the first-order conditions for $\hat{\beta}$, it will follow that

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)\xrightarrow{d}N(0,V),\quad V=(M'WM)^{-1}M'W\Omega WM(M'WM)^{-1}. \tag{5.2}$$

In developing regularity conditions for this result the paper first focuses on asymptotic normality of $\sqrt{n}\,\hat{m}_n(\beta_0)$. This condition is the most interesting one, because it is the channel through which estimation of h affects the asymptotic distribution of $\hat{\beta}$.

It is useful to organize the discussion around a few high level conditions. Let $\|h\|$ denote a norm for the function h, such as a Sobolev norm (supremum norm for a function and its derivatives), and let $m(z,h)=m(z,\beta_0,h)$.

ASSUMPTION 5.1 (Linearization): *(i) There is a function $D(z,h)$ that is linear in h such that for all h with $\|h-h_0\|$ small enough,*

$$\|m(z,h)-m(z,h_0)-D(z,h-h_0)\|\leq b(z)\|h-h_0\|^2;$$

(ii) $E[b(z)]\sqrt{n}\,\|\hat{h}-h_0\|^2\xrightarrow{p}0$.

This condition requires that the remainder term from a linearization be small. It is the precise regularity condition corresponding to equation (4.1) implying the previous linearization when $b(z)\|h(\theta)-h_0\|^2 \leq b(z)\|\theta\|^2$, as further discussed below. The remainder term in this condition is analogous to $m(z,h)-m(z,h_0)-[\partial m(z,h_0)/\partial h](h-h_0)$ in the parametric case, where h is a Euclidean vector, with $D(z,h-h_0)$ taking the place of $[\partial m(z,h_0)/\partial h](h-h_0)$. In the general case, part (i) implies Frechet differentiability of $m(z,h)$ at h_0, with $D(z,h-h_0)$ being the derivative. Also, part (ii) either requires that $m(z,h)$ be linear in h (i.e. $b(z)=0$), or that the convergence rate of $\hat{h}$ be faster than $n^{-1/4}$, in terms of the norm $\|h\|$. Stone's (1982) bounds on convergence rates mean that this rate will only be achievable when h is sufficiently smooth, being differentiable to a certain order that increases with the dimension of the argument of h.

Assumption 5.1 is often straightforward to verify when $\|h\|$ is a Sobolev norm. Frechet differentiability with respect to such a strong norm is often easy to show, while the $n^{-1/4}$ convergence can be obtained using known uniform convergence rates. For example, Newey (1994a) gives uniform convergence rates for series estimators, that are used in Section 6 to verify Assumption 5.1. Also, for kernel estimators, uniform convergence rates of Bierens (1987), Newey (1994b), and others can be used to develop primitive conditions for Assumption 5.1.

Let F_0 denote the true distribution of z.

ASSUMPTION 5.2 (Stochastic Equicontinuity):

$$\sum_{i-1}^{n}\left[D\left(z_i,\hat{h}-h_0\right)-\int D\left(z,\hat{h}-h_0\right)dF_0\right]\Big/\sqrt{n}\xrightarrow{p}0.$$

This condition is analogous to the requirement for parametric two-step estimators that $\sqrt{n}\{n^{-1}\sum_{i=1}^{n}\partial m(z_i,h_0)/\partial h - E[\partial m(z,h_0)/\partial h]\}(\hat{h}-h_0)$ converge to zero. Andrews (1994) has recently given quite general sufficient conditions for stochastic equicontinuity, that may be applied. Alternatively, stochastic equicontinuity can be shown by direct calculation. For kernel estimators Assumption 5.2 will follow from the well known U-statistic projection theorem and a "small bias" condition as discussed in Newey and McFadden (1994). For series estimators Assumption 5.2 can be shown as in Section 6.

The following condition is important for $\sqrt{n}$-consistency.

ASSUMPTION 5.3 (Mean-square Continuity): *(i) There is $\alpha(z)$ and a measure $\hat{F}$ such that $E[\alpha(z)]=0$, $E[\|\alpha(z)\|^2]<\infty$, and for all $\|\hat{h}-h_0\|$ small enough, $\int D(z,\hat{h}-h_0)dF_0=\int\alpha(z)\,d\hat{F}$. (ii) For the empirical distribution $\tilde{F}(z)=n^{-1}\sum_{i=1}^{n}1(z_i\leq z)$, $\sqrt{n}\left[\int\alpha(z)\,d\hat{F}-\int\alpha(z)\,d\tilde{F}\right]\xrightarrow{p}0$.*

Assumptions 5.1 and 5.2 involve "second-order" terms. Thus both of these conditions are "regularity conditions," meaning that they should be satisfied if $m(z,\beta_0,h)$ is sufficiently smooth and $\hat{h}$ sufficiently well behaved. The terms in

Assumptions 5.3 are "first-order" terms. These conditions are the ones that allow $\sum_{i=1}^n m(z_i, \hat{h})/\sqrt{n}$ to be asymptotically normal, even though $\hat{h}$ may converge at a slower rate. Condition (i) imposes a representation of $\int D(z, \hat{h} - h_0)\, dF_0$ as an integral with respect to an estimated measure, i.e. an average over some estimated distribution. Condition (ii) is an asymptotic equivalence requirement for the estimator $\hat{F}$ and the empirical distribution function $\tilde{F}$. It requires that $\hat{F}$ be nonparametric in the sense that the integral over $\hat{F}$ is asymptotically equivalent to the integral over the nonparametric estimator $\hat{F}$.

One could simplify Assumption 5.3 by just requiring that there be an $\alpha(z)$ with $\sqrt{n}\int D(z, \hat{h} - h_0)\, dF_0 - \sum_{i=1}^n \alpha(z_i)/\sqrt{n} \xrightarrow{p} 0$. However, the form given above leads to a useful link with the pathwise derivative formula, that can be used to calculate $\alpha(z)$. Under certain conditions, the $\alpha(z)$ of Assumption 5.3 must be the pathwise derivative of $m(h) = E[m(zh)]$, i.e. the correction term for estimation of h that was described earlier in the paper. Hence, the pathwise derivative calculation can be used to find a function that is a candidate for Assumption 5.3. Let $h(F)$ denote the limit of $\hat{h}$ when F is the true distribution and suppose that $\int D(z, h(F) - h_0)\, dF_0 = \int \alpha(z)\, dF$ when $\|h(F) - h_0\|$ is sufficiently small. Consider a parametric submodel F_θ with score $S(z)$ and $h(\theta) = h(F_\theta)$ and suppose that $\|h(\theta) - h_0\| \leq C\|\theta - \theta_0\|$ and $\int \|\alpha(z)\|^2\, dF_\theta$ is bounded in a neighborhood of θ_0. Then by Lemma 7.1 of Ibragimov and Hasminskii (1981), $\partial \int D(z, h(\theta) - h_0)\, dF_0/\partial\theta = \partial \int \alpha(z)\, dF_\theta/\partial\theta = E[\alpha(z)S(z)]$, and under Assumption 5.1, $\|m(h(\theta)) - m(h_0) - \int D(z, h_\theta - h_0)\, dF_0\|/\|\theta - \theta_0\| \leq E[b(z)]\|h(\theta) - h_0\|^2/\|\theta - \theta_0\| \to 0$. Hence, $E[m(z, h(\theta))]$ is differentiable in θ and the derivative satisfies equation (3.9) for $\alpha(z)$ from Assumption 5.3.

Estimation of h will have no effect on the asymptotic variance when $\alpha(z) = 0$. A simple and powerful condition for $\alpha(z) = 0$ is then $E[D(z, h - h_0)] = 0$ for all h close enough to h_0. For example, consider the case of Proposition 2 of Section 3, where $m(z, h) = m(z, h(v))$ depends on h only through its value at v. There $D(z, h) = D(z)h(v)$ for $D(z) = \partial m(z, h_0(v))/\partial h$, so that $E[D(z, h - h_0)] = E[E[D(z)|v]\{h(v) - h_0(v)\}] = 0$ if $E[D(z)|v] = 0$. Here the pathwise derivative condition of Proposition 2 implies that Assumption 5.3 is satisfied with $\alpha(z) = 0$.

Asymptotic normality of $\sqrt{n}\,\hat{m}_n(\beta_0)$ is an immediate consequence of Assumptions 5.1–5.3, the triangle inequality, and the central limit theorem.

LEMMA 5.1: *If Assumptions* 5.1–5.3 *are satisfied then for*

$$\Omega = E\left[\{m(z, h_0) + \alpha(z)\}\{m(z, h_0) + \alpha(z)\}'\right],$$

$$\sqrt{n}\,\hat{m}_n(\beta_0) = \sum_{i=1}^n \left[m(z_i, h_0) + \alpha(z_i)\right]/\sqrt{n} + o_p(1) \xrightarrow{d} N(0, \Omega).$$

Assumptions 5.1–5.3 thus provide a set of sufficient conditions for the important intermediate result that $\sqrt{n}\,\hat{m}_n(\beta_0)$ is asymptotically normal.

The equality in the conclusion of Lemma 5.1 is an immediate consequence of Assumptions 5.1–5.3 (and the triangle inequality). Thus, this part of the conclusion should continue to hold with dependent observations, because indepen-

dence is not required in Assumptions 5.1–5.3. Of course, with dependence the asymptotic variance may have a different form. A general form is the asymptotic variance of $\sum_{i=1}^{n}[m(z_i,h_0)+\alpha(z_i)]/\sqrt{n}$, that would include covariance terms between $m(z,h_0)+\alpha(z)$ for different observations. Also, it may be more difficult to show that Assumptions 5.1–5.3 hold when the observations are dependent. Nevertheless, the absence of an independence requirement in these conditions indicates that the form of the correction term $\alpha(z)$ for first step estimation will be the same under dependence as under independence.

The conclusion of Lemma 5.1 can be combined with consistency and uniform convergence results to obtain asymptotic normality of $\hat{\beta}$. The following condition is useful for uniform convergence.

ASSUMPTION 5.4: *There are* $\varepsilon, \|h\|, b(z), \tilde{b}(z) > 0$ *such that* (*i*) *for all* $\beta \in \mathscr{B}$, $m(z,\beta,h_0)$ *is continuous at* β *with probability one,* $\|m(z,\beta,h_0)\| \leq b(z)$; (*ii*) $\|m(z,\beta,h)-m(z,\beta,h_0)\| \leq \tilde{b}(z)(\|h-h_0\|)^{\varepsilon}$.

Part (i) of this condition is a standard assumption for uniform convergence, while part (ii) guarantees that the remainder from estimation of h is uniformly small. Together these conditions will lead to $\sup_{\beta\in\mathscr{B}}\|\hat{m}_n(\beta)-E[m(z,\beta,h_0)]\| \xrightarrow{p} 0$. Identification is also important for consistency, as in the next condition.

ASSUMPTION 5.5: $\hat{W} \xrightarrow{p} W$, *W is positive semidefinite,* $WE[m(z,\beta,h_0)]=0$ *has a unique solution on* $\mathscr{B}$ *at* β_0, *and* $\mathscr{B}$ *is compact.*

It is straightforward to relax the compactness assumption on $\mathscr{B}$ when the moment function $m(z,\beta,h)$ is linear in β. To save space this case is not considered separately. Consistency of $\hat{\beta}$ follows from Assumptions 5.4 and 5.5.

LEMMA 5.2: *If Assumptions* 5.4 *and* 5.5 *are satisfied and* $\|\hat{h}-h_0\| \xrightarrow{p} 0$, *then* $\hat{\beta} \xrightarrow{p} \beta_0$.

It is also important for asymptotic normality to have uniform convergence of the Jacobian with a full rank limit. The next condition is useful in this regard.

ASSUMPTION 5.6: (*i*) $\beta \in$ *interior*$(\mathscr{B})$; (*ii*) *there is* $\|h\|$, $\varepsilon > 0$, *and a neighborhood* $\mathscr{N}$ *of* β_0 *such that for all* $\|h-h_0\|<\varepsilon$, $m(z,\beta,h)$ *is differentiable in* β *on* $\mathscr{N}$; (*iii*) $M'WM$ *is nonsingular for* $M=E[\partial m(z,\beta_0,h_0)/\partial\beta]$; (*iv*) $E[\|m(z,\beta_0,h_0)\|^2]<\infty$; (*v*) *Assumption* 5.4 *is satisfied with* $m(z,\beta,h)$ *there equal to each row of* $\partial m(z,\beta,h)/\partial\beta$.

Asymptotic normality now follows by combining previous conditions.

LEMMA 5.3: *If Assumptions* 5.1–5.6 *are satisfied and* $\|\hat{h}-h_0\| \xrightarrow{p} 0$, *then for V from equation* (5.2), $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V)$.

A consistent estimator of the asymptotic variance can be formed by replacing, in equation (5.2), W with $\hat{W}$, M with $\hat{M} = n^{-1}\sum_{i=1}^{n} \partial m(z_i, \hat{\beta}, \hat{h})/\partial\beta$, and Ω with a consistent estimator $\hat{\Omega}$. This replacement gives

$$(5.3) \qquad \hat{V} = (\hat{M}'\hat{W}\hat{M})^{-1}\hat{M}'\hat{W}\hat{\Omega}\hat{W}\hat{M}(\hat{M}'\hat{W}\hat{M})^{-1},$$

$$\hat{\Omega} = \sum_{i=1}^{n} \left[m(z_i, \hat{\beta}, \hat{h}) + \hat{\alpha}(z_i)\right]\left[m(z_i, \hat{\beta}, \hat{h}) + \hat{\alpha}(z_i)\right]'/n,$$

where $\hat{\alpha}(z)$ is an estimator of $\alpha(z)$. The only potentially difficult part of this procedure is forming $\hat{\alpha}(z)$. Often $\hat{\alpha}(z)$ can be constructed by finding a formula for $\alpha(z)$ and then "plugging in" estimates for unknown parameters and functions. Also, for some $\hat{h}$ it is possible to construct $\hat{\alpha}(z)$ without knowing the form of $\alpha(z)$. Such a construction for kernel estimators is given in Newey (1944b) and for series estimators is discussed below. For specific estimators it is possible to give primitive conditions for $\hat{\alpha}(z)$ to lead to consistent variance matrix, but at the level of generality of this section it seems impossible to do more than give the high-level condition of the next result.

LEMMA 5.4: *If* (*i*) $\sum_{i=1}^{n}\|\hat{\alpha}(z_i) - \alpha(z_i)\|^2/n \xrightarrow{p} 0$; (*ii*) $\hat{\beta} \xrightarrow{p} \beta_0$ *and there is* $\|h\|$ *such that* $\|\hat{h} - h_0\| \xrightarrow{p} 0$, $\|m(z, \beta, h) - m(z, \beta_0, h_0)\| \leq b(z)(\|\beta - \beta_0\| + \|h - h_0\|)$ *and* $E[b(z)^2] < \infty$; *and* (*iii*) *Assumption* 5.6 *is satisfied, 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 estimators, similar to that of Hansen (1985). For this choice of $\hat{W}$, $(\hat{M}'\hat{\Omega}^{-1}\hat{M})^{-1}$ will be a consistent estimator of the asymptotic variance of $\hat{\beta}$.

6. SERIES ESTIMATION OF PROJECTION FUNCTIONALS

One type of nonparametric first step that is useful in many applications is a series estimator of a population least squares projection. A series estimator is obtained from a sample regression on a finite dimensional subspace, with dimension that is allowed to grow with the sample size. Let $\mathcal{G}$ be a linear set of functions of x and $g_0(x)$ the mean-square projection of y on $\mathcal{G}$, as considered in Section 4. Let $p^K(x) = (p_{1K}(x), \ldots, p_{KK}(x))'$ be a vector of functions, each of which is an element of $\mathcal{G}$. Denote the data observations by y_i and x_i $(i = 1, 2, \ldots)$, and let $y \equiv (y_1, \ldots, y_n)'$ and $p^K \equiv [p^K(x_1), \ldots, p^K(x_n)]$, for sample size n. A series estimator of $g_0(x)$ is

$$(6.1) \qquad \hat{g}(x) = p^K(x)'\hat{\pi}, \quad \hat{\pi} = (p^{K\prime}p^K)^{-}p^{K\prime}y,$$

where $(\cdot)^{-}$ denotes a generalized inverse, and K subscripts for $\hat{g}(x)$ and $\hat{\pi}$

have been suppressed for notational convenience. Examples of series estimators are polynomial and spline regressions. Polynomial regression is based on functions $p_{kK}(x)$ that are products of powers of components of x. Spline regression is based on functions that are powers over some range and constant elsewhere, with switch points, referred to as "knots," that vary with k, e.g. see Powell (1981).

A data based $\hat{K}$ helps operationalize the nonparametric nature of series estimators, by 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, where $\hat{\beta}$ is the main object of interest, but this question is outside the scope of this paper.

To avoid multicollinearity problems, it may be useful to replace $p^K(x)$ with nonsingular linear transformations. For example, power series can be replaced by polynomials that are orthogonal with respect to some known distribution, and splines can be replaced by B-splines. Of course, this replacement will not affect the estimator. Also, note that the elements of each x 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.

An estimator of the correction term $\alpha(z)$ is needed for variance estimation. For series an estimator can be based on only the series and the moment function, without having to know the form of $\alpha(z)$. Let

$$\hat{\alpha}(z) = \hat{\psi}'\hat{\Sigma}^{-}p^{\hat{K}}(x)[y - \hat{g}(x)], \tag{6.2}$$

$$\hat{\psi} = \partial\left[n^{-1}\sum_{i=1}^{n} m\left(z_i, \hat{\beta}, p^{\hat{K}\prime}\pi\right)\right]\Big/\partial\pi|_{\pi=\hat{\pi}}.$$

One interpretation of $\hat{\alpha}(z)$ is that $\hat{\psi}'\hat{\Sigma}^{-}p^{\hat{K}}(x)$ is an estimator of $\delta(x)$, so that $\hat{\alpha}(z)$ estimates $\alpha(z) = \delta(x)[y - g_0(x)]$. Under Assumption 5.1 and $\hat{g}$ close to g_0, $\hat{\psi}$ should be approximately $n^{-1}\Sigma_{i=1}^{n}[D(z_i, p_{1\hat{K}}), \ldots, D(z_i, p_{\hat{K}\hat{K}})]'$, which estimates $\int \delta(x)p^{\hat{K}}(x)\,dF_0$. Thus, $\hat{\Psi}'\hat{\Sigma}^{-}p^{\hat{K}}(x)$ is an estimator of the regression of $\delta(x)$ on $p^{\hat{K}}(x)$, which should approximate $\delta(x)$ for large $\hat{K}$ and n. Alternatively, $\hat{\alpha}(z)$ can be viewed as a standard parametric correction for estimation of $\hat{\pi}$, with K fixed at $\hat{K}$. The reason that a standard parametric correction works as K grows is that it accounts properly for variability, and bias for the series approximation will be small when $g(x)$ and/or $\delta(x)$ have a sufficient number of bounded derivatives, as discussed below. Also, estimation of K does not affect the variance, because the asymptotic variance is the same for all K sequences in certain ranges.

To specify regularity conditions for series estimators, let $u = y - g_0(x)$, $u_i = y_i - g_0(x_i)$. The first assumption is common in the literature on series estimation, e.g. Stone (1985), and dropping it would complicate the results.

ASSUMPTION 6.1: *$E[\|u_i\|^2|x_i]$ is bounded.*

Next is a normalization condition for the second moment matrix of the series terms and an assumption that the vectors of functions are nested within a range of $\hat{K}$ values.

ASSUMPTION 6.2: (i) *The smallest eigenvalue of* $E[p^K(x)p^K(x)']$ *is bounded away from zero uniformly in* K; (ii) *there are* $\underline{K}(n)$ *and* $\bar{K}(n)$ *such that* $\underline{K}(n) \leq \hat{K} \leq \bar{K}(n)$ *with probability approaching one,* $p^{\bar{K}}(x)$ *is a subvector of* $p^{K+1}(x)$ *for all* K *with* $\underline{K}(n) \leq K < K+1 \leq \bar{K}(n)$; (iii) *for each* K *there is nonzero* $\bar{\pi}$ *such that* $\bar{\pi}'p^K(x)$ *is a nonzero constant on the support of* x.

Part (i) is a normalization that places restrictions on the magnitude of the series terms. It is often necessary to work with a nonsingular linear transformation of $p^K(x)$ in order to obtain part (i). Such transformations do not affect the estimator but are useful in specifying regularity conditions. Polynomials that are orthonormal with respect to some weight function provide a convenient transformation for power series, with part (i) holding if the density of the x's is bounded below by the weight. B-splines provide a useful transformation for splines. Bounds on the transformed functions are also important for the regularity conditions, as further discussed below.

Assumptions 6.1 and 6.2 are useful for controlling the variance of a series estimator. The bias will be determined by the degree of approximation of the true function by a linear combination. A supremum Sobolev norm is useful in specifying a degree of approximation. For a vector of functions $f(x)$, let λ and $\partial^\lambda f(x)$ be as described in Section 4, and let

$$|f(x)|_d = \max_{|\lambda| \leq d} \max_{x \in \mathscr{X}} |\partial^\lambda f(x)|,$$

where $\mathscr{X}$ is the support of x and with $|f(x)|_d$ equal to infinity if $\partial^\lambda f(x)$ does not exist for some $|\lambda| \leq d$. The next condition specifies an approximation rate.

ASSUMPTION 6.3: *For each nonnegative integer* d, *if* $|g_0(x)|_d$ *is finite then there are constants* C, $\alpha_d > 0$ *such that for all* K *there is* π *with* $|g_0(x) - p^K(x)'\pi|_d \leq CK^{-\alpha}d$.

More primitive conditions for Assumption 6.3 follow from known approximation rate results for series. For example, for power series, splines, and $d = 0$ Lorentz (1986) and Schumaker (1981) give $\alpha = s/r$, where s is the number of continuous derivatives of $g_0(x)$ that exist and r is the dimension of x.

Assumptions 6.1–6.3 are useful for the conditions of Section 5, in particular for convergence rates when the norm $\|\cdot\|$ of Assumption 5.1 is a Sobolev norm. The rates, taken from Newey (1994a), depend on the magnitude of the series terms and their derivatives, so it is useful to have some notation for these. Let

$$\zeta_k(K) = \sup_{|\lambda| = d,\, x \in \mathscr{X}} \|\partial^\lambda p^K(x)\|.$$

For orthonormal polynomials the Euclidean norm and the supremum can be interchanged without affecting the bound, but not for B-splines (where all but a bounded number of terms are zero at each x). For this reason bounds depending on $|p_{kK}|_d$ will also appear in the regularity conditions given below.

Specific approximating functions have corresponding bounds $\zeta_d(K)$ and $|p_{kK}|_d$. For orthonormal polynomials with respect to a uniform weight, $\zeta_d(K)$ is bounded above by CK^{1+2d} for a constant C and $|p_{kK}|_d$ by $CK^{(1/2)+2d}$, e.g. see Abramowitz and Stegum (1972). For B-splines both $\zeta_d(K)$ and $|p_{kK}|_d$ are bounded above by $CK^{.5+d}$, e.g. see Powell (1981).

Asymptotic normality and consistency of the variance estimator can be shown by specifying more primitive conditions for the assumptions of Section 5. The next condition is a more primitive version of the linearization condition. For notational convenience, suppress the n argument of $\bar{K}(n)$ and $\underline{K}(n)$. Also, unless otherwise specified, in each of the conditions to follow let the $\alpha = \alpha_d$ for d in the corresponding condition and α_d from Assumption 6.3.

ASSUMPTION 6.4: (i) *There are* $\varepsilon, d, b(z) > 0$ *and* $D(z, g; \beta, \tilde{g})$ *that is linear in* g *such that for all* $\|\beta - \beta_0\| < \varepsilon$ *and* $|\tilde{g} - g_0|_d < \varepsilon$, $\|m(z, \beta, g) - m(z, \beta, \tilde{g}) - D(z, g - \tilde{g}; \beta, \tilde{g})\| \leq b(z)(|g - g_0|_d)^2$; (ii) $E[b(z)]\zeta_d(\bar{K})[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}] \to 0$ *and* $E[b(z)]\sqrt{n}\,\zeta_d(\bar{K})^2[\bar{K}/n + \underline{K}^{-2\alpha}] \to 0$.

Part (i) is a stronger version of Assumption 5.1(i). This strengthened condition is useful for showing consistency of the asymptotic variance estimator. Part (ii) implies that $\hat{g}$ is consistent in the Sobolev norm $|g|_d$, and that the linearization remainder is small, using Newey's (1994a) result that $|\hat{g} - g_0|_d = O_p(\zeta_p(\zeta_d(\bar{K})[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}]))$.

The next condition is a more primitive assumption for stochastic equicontinuity.

ASSUMPTION 6.5: *There is* $b(z)$, $d > 0$ *such that* $E[b(z)^2] < \infty$, $\|D(z, g; \beta_0, g_0)\| \leq b(z)|g|_d$, $\Sigma_{\underline{K} \leq K \leq \bar{K}} K^{-2\alpha} \to 0$, *and* $(\Sigma_{k=1}^{\bar{K}} |p_{k\bar{K}}|_d^2)^{1/2}[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}] \to 0$.

The rate conditions here are very close to being consistency and small bias conditions in a norm for which $D(z, g; \beta_0, g_0)$ is continuous in g.

The next condition corresponds to mean square differentiability.

ASSUMPTION 6.6: (i) *There is* $\delta(x) \in \mathscr{G}$ *such that* $E[D(z, g; \beta_0, g_0)] = E[\delta(x)g(x)]$ *for all* $g \in \mathscr{G}$; (ii) *for each* K *there are* π_K *and* ξ_K *such that*

$$n \cdot E\left[\left\|\delta(x) - \xi_{\underline{K}} p^{\underline{K}}(x)\right\|^2\right] \cdot E\left[\left\|g_0(x) - \pi_{\underline{K}} p^{\underline{K}}(x)\right\|^2\right] \to 0,$$

$$\zeta_0(\bar{K})^4\left(\sum_{\underline{K} \leq K \leq \bar{K}} K\right)\Big/ n \to 0,$$

$$\sum_{\underline{K} \leq K \leq \bar{K}} \zeta_0(K)^2 E\left[\left\|g_0(x) - \pi_K p^K(x)\right\|^2\right] \to 0, \quad \text{and}$$

$$\sum_{\underline{K} \leq K \leq \bar{K}} E\left[\left\|\delta(x) - \xi_K p^K(x)\right\|^2\right] \to 0.$$

The first part of this assumption is mean-square continuity of the expected derivative $E[D(z, g; \beta_0, g_0)]$. As discussed in Section 4, this condition is important for $\sqrt{n}$-consistency of the correction term. Most of the other parts are asymptotic bias bounds.

It is interesting to note that the approximation error for $g_0(x)$ is *not* required to go to zero faster than $1/\sqrt{n}$. Instead, the product of the approximation errors for $\delta(x)$ and $g_0(x)$ must go to zero faster than $1/\sqrt{n}$. This feature is a consequence of orthogonality of least squares projection errors, which means that the bias term $E[\delta(x)\{g_0(x) - g_K(x)\}]$ is actually equal to $E[\{\delta(x) - \delta_K(x)\}\{g_0(x) - g_K(x)\}]$, where g_K and δ_K are the respective projections of g_0 and δ on p^K. This feature is different from kernel first step estimators, which require approximation error to go to zero faster than $1/\sqrt{n}$, a feature referred to as "undersmoothing" (the bias-squared is going to zero faster than the variance shrinks at a rate less than $1/n$). Thus, if the approximation rate for $\delta(x)$ is sufficiently rapid, undersmoothing is not required for series estimators to achieve $\sqrt{n}$-consistency.

The next condition is important for variance consistency.

ASSUMPTION 6.7: *There is* $b(z)$ *such that for all* $\tilde{g} \in \mathcal{G}$, $\|D(z, \tilde{g}; \beta, g) - D(z, \tilde{g}; \beta_0, g_0)\| \leq b(z)(\|\beta - \beta_0\| + |g - g_0|_d)|\tilde{g}|_d$, $\zeta_0(\bar{K})[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}] \to 0$, *and*

$$E[b(z)]\Big(\sum_{k \leq \bar{K}} |p_{k\bar{K}}|_d\Big)\bar{K}^{1/2}\zeta_d(\bar{K})\Big[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}\Big] \to 0.$$

THEOREM 6.1: *If Assumptions* 5.4–5.6 *are satisfied with* $g = h$ *and* $\|h\| = |g|_d$, *and Assumptions* 6.1–6.6 *are satisfied, then for* $\alpha(z) = \delta(x)u$ *and* V *in equation* (5.2),

$$\sqrt{n}\,(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V).$$

If in addition Assumption 6.7 *is satisfied, then for* $\hat{\alpha}(z)$ *in equation* (6.2), *and* $\hat{V}$ *in equation* (5.3), $\hat{V} \xrightarrow{p} V$.

Primitive conditions for asymptotic normality and consistent variance estimation can be obtained by specifying conditions for the hypothesis of this result. In Section 7, such conditions are given for first step power series estimators for two examples.

7. POWER SERIES ESTIMATORS FOR SEMIPARAMETRIC INDIVIDUAL EFFECTS AND AVERAGE DERIVATIVES

In this section primitive conditions are given for asymptotic normality in the panel data and average derivative examples introduced in Section 2, when the first step is a polynomial regression estimator. To describe the first step, let λ denote a vector of nonnegative integers, as in Section 4. For a vector τ with the

same dimension as x let $\tau^{\lambda} = \prod_{l=1}^{r} \tau_l^{\lambda_l}$. Let $\tau(x)$ be a one-to-one function of x that on any compact set is differentiable to all orders with bounded derivatives and has $\det(\partial\tau(x)/\partial x)$ bounded away from zero. For a sequence $(\lambda(k))_{k=1}^{\infty}$ of distinct such vectors, a power series approximation corresponds to

$$(7.1) \qquad p_{kK}(x) = \tau(x)^{\lambda(k)} \qquad (k = 1, 2, \ldots).$$

In the results to follow the natural ordering condition that $|\lambda(k)|$ is increasing in k will be assumed to be satisfied. Given the choice of approximating function in equation (7.1), series estimators can be formed as described in Section 6.

The transformation $\tau(x)$ is permitted to attenuate outlier effects using bounded functions. For example, $\tau(x) = (l(x_1), \ldots, l(x_r))$ for the logit transform $l(u) = 1/(1 + e^u)$ is a bounded transformation that may help lower the influence of large values of the regressors. In addition, it may improve computation to replace each $\tau(x)^{\lambda(k)}$ with the product of orthogonal polynomials of order corresponding to components of $\lambda(k)$, with respect to some weight function on the range of $\tau(x)$. This replacement can help alleviate severe multicollinearity of power series.

To apply previous results to power series it is important to be able to check the conditions of Section 6. An assumption that is useful in this respect is that the x is continuously distributed, has support equal to a Cartesian product of compact intervals, with density that is bounded below by a positive multiple of a product of univariate beta densities. Then Assumption 6.2 will hold for the nonsingular transformation that replaces each power term by a product of polynomials that are orthonormal with respect to the beta density. Furthermore, bounds on the series terms will follow from standard results for orthonormal polynomials. For example, if the density is assumed to be bounded away from zero on the support, then Assumption 6.2 will be satisfied for polynomials that are orthonormal with respect to the uniform weight, and these will satisfy (e.g. see Lemma A.15 of Newey (1994a)),

$$(7.2) \qquad \zeta_d(K) \leqslant K^{1+2d}, \qquad |p_{kK}|_d \leqslant K^{1/2+2d}.$$

This particular case is useful for the panel data example.

EXAMPLE 2 CONTINUED: A probit panel data estimator can be obtained by solving equation (3.1) for the $m(z, \beta, h)$ in equation (3.4), with $\hat{h}_t(x) = \hat{g}_t(x)$ equal to a power series estimator based on equation (7.1). Specifically, let $\hat{g}_{ti} = p^K(x_i)'\hat{\pi}_t$ for $\hat{\pi}_t = (p^{K\prime}p^K)^- p^{K\prime}(y_{t1}, \ldots, y_{tn})'$, let $\hat{f}_i = (x'_{1i} - x'_{2i}, \Phi^{-1}(\hat{g}_{2i}))'$, and let $\hat{\beta} = (\hat{\gamma}', \hat{\sigma})$ solve

$$n^{-1} \sum_{i=1}^{n} \hat{f}_i \left[\Phi^{-1}(\hat{g}_{1i}) - \sigma \Phi^{-1}(\hat{g}_{2i}) - (x_{1i} - x_{2i})'\gamma \right] = 0.$$

It is possible that $\hat{g}_{ti}$ may not be between zero and one for some t and i, so that the inverses in this estimating equation may not exist, although under the conditions given below the probability of this event goes to zero as the sample

size grows. In practice it may be desirable to constrain these predicted probabilities away from zero and one.

An estimator of the asymptotic variance could be constructed as described in Section 6. In this example there is a simpler alternative, because the derivative of $m(z,\beta,g)$ with respect to the value of g is a function of only x (and hence is already in $\mathscr{G}$, rather than having to be projected on $\mathscr{G}$). Let

$$(7.3)\qquad \hat{\delta}_{ti} = (-\sigma)^{t-1}\hat{J}_i\left[1/\phi\left(\Phi^{-1}(\hat{g}_{ti})\right)\right], \quad \hat{\alpha}_{ti} = \hat{\delta}_{ti}(y_{ti} - \hat{g}_{ti}) \qquad (t = 1,2),$$

$$\hat{V} = \hat{M}^{-1}\left[n^{-1}\sum_{i=1}^{n}(\hat{\alpha}_{1i} + \hat{\alpha}_{2i})(\hat{\alpha}_{1i} + \hat{\alpha}_{2i})'\right]\hat{M}^{-1}, \qquad \hat{M} = -n^{-1}\sum_{i=1}^{n}\hat{J}_i\hat{J}_i'.$$

The following result shows asymptotic normality of $\hat{\beta}$ and consistency of $\hat{V}$. Let r denote the dimension of x_t.

THEOREM 7.1: *If equation* (3.3) *is satisfied,* $E[J(x)J(x)']$ *is nonsingular for* $J(x) = (x_1' - x_2', x_2'\gamma_0 + \rho(x))'$, x *is continuously distributed with support equal to a Cartesian product of compact intervals, the density of* x *is bounded away from zero on its support,* $\rho(x)$ *is continuously differentiable of order* s, $\bar{K}^6/n \to 0$, $n\bar{K}^4\underline{K}^{-2s/r} \to 0$. *Then for* $\alpha_t(z)$ *in equation* (4.8), $\psi(z) = \alpha_1(z) + \alpha_2(z)$, *and* $V = E[\psi(z)\psi(z)']$,

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0,V), \qquad \hat{V} \xrightarrow{p} V.$$

A necessary condition for the growth rate condition on $\bar{K}$ and $\underline{K}$ is that $s > 5r$. Also, when $\rho(x)$ is smoother (has higher order derivatives), the lower bound $\underline{K}$ can be allowed to go to infinity slower.

EXAMPLE 3 CONTINUED: As previously discussed, an average derivative estimator for a projection on a restricted set of functions, such as additive ones, may be of interest. This estimator can be formed by differentiating the predicted value of a regression on the functions in (7.1) that are elements of a restricted set. Specifically, an additive-interactive projection can be estimated by eliminating the terms in equation (7.1) that are not included in the set of additive interactive functions. For example, an additive projection can be estimated by restricting the series in equation (7.1) to consist only of the univariate terms.

The average derivative can then be estimated by differentiating the predicted value from the series regression, as in

$$(7.4)\qquad \hat{\beta} = n^{-1}\sum_{i=1}^{n}\left[\partial p^{\hat{K}}(x_i)/\partial x\right]'\hat{\pi} = \hat{\Psi}'\hat{\pi},$$

$$\hat{\pi} = \left(p^{\hat{K}}, p^{\hat{K}}\right)^{-}p^{\hat{K}\prime}y, \quad \hat{\Psi} = n^{-1}\sum_{i=1}^{n}\partial p^{\hat{K}}(x_i)/\partial x.$$

The asymptotic variance can be estimated in the general way discussed in

Section 6, as

$$\hat{V} = n^{-1} \sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i',$$

$$\hat{\psi}_i = \left[\partial p^{\hat{K}}(x_i)/\partial x\right]' \hat{\pi} - \hat{\beta} + \hat{\psi}' \left(p^{\hat{K}\prime} p^{\hat{K}}/n\right)^{-} p^{\hat{K}}(x_i)' \left[y_i - p^{\hat{K}}(x_i)' \hat{\pi}\right].$$

The asymptotic theory for the average derivative estimator is more difficult than for the panel data estimator. A necessary condition for $\sqrt{n}$-consistency is that the density of x must go to zero on the boundary of the support, as discussed in Newey and Stoker (1993). Also, the score in the adjustment term is often unbounded and the theory requires approximation rates for a function and its derivatives (i.e. Sobolev approximation rates). Because of these complications it is helpful to restrict K to be nonrandom and impose a strong smoothness condition on the projection. These conditions are imposed in the following result.

THEOREM 7.2: *Suppose that $E[u^2|x]$ is bounded, x is continuously distributed with support $x_{j=1}^{r}[x_{lj}, x_{uj}]$, the density of x is bounded below by $C\prod_{j=1}^{r}(x - x_{lj})^{\delta}(x_{uj} - x)^{\delta}$, $g_0(x)$ is continuously differentiable for all orders and there is a constant C such that $|\partial^{\lambda} g_0(x)| \leq C^{|\lambda|}$, $f(x)$ is continuously differentiable, $\partial f(x)/\partial x$ is zero on the boundary of the support, and $E[\|f(x)^{-1}\partial f(x)/\partial x\|^2]$ is finite, $\overline{K} = \underline{K} = K(n)$ such that $K(n) \geq n^{\varepsilon}$ for some $\varepsilon > 0$ and $K(n)^{7+2\delta}/n \to 0$. Then for $\psi(z)$ in equation (4.9) and $V = E[\psi(z)\psi(z)']$, $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$ and $\hat{V} \xrightarrow{p} V$.*

For the case where $g_0(x) = E[y|x]$, the asymptotic variance V is the same as that of Hardle and Stoker (1989), confirming the conclusion of Proposition 1 that a series estimator of the average derivative should have the same asymptotic variance as a kernel estimator.

Dept. of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02139, U.S.A.

Manuscript received August, 1989; final revision received March, 1994.

APPENDIX: PROOFS OF THEOREMS

Throughout this Appendix C will denote the 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 Linberg-Levy central limit theorem imply that for any $S_\theta(z) \in \mathcal{S}$, $(\sqrt{n}(\hat{\beta} - \beta_0)', \Sigma_{i=1}^{n} S_\theta(z_i)/\sqrt{n})'$ converges in distribution to $N(0, E[(\psi(z)', S_\theta(z))'(\psi(z)', S_\theta(z))])$. Furthermore, by the 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: Follows immediately from the triangle inequality and the Lindberg-Levy central limit theorem. *Q.E.D.*

PROOF OF LEMMA 5.2: Let $\tilde{m}_n(\beta) = \sum_{i=1}^n m(z_i, \beta, h_0)/n$, $m(\beta) = E[m(z, \beta, h_0)]$, $\hat{Q}(\beta) = \hat{m}_n(\beta)'\hat{W}\hat{m}_n(\beta)$, and $Q(\beta) = m(\beta)'Wm(\beta)$. By Assumption 5.4, $\sup_{\beta \in \mathscr{B}} \|\hat{m}_n(\beta) - \tilde{m}_n(\beta)\| \xrightarrow{p} 0$, while by Andrews (1987), $\sup_{\beta \in \mathscr{B}} \|\tilde{m}_n(\beta) - m(\beta)\| \xrightarrow{p} 0$. It then is straightforward to show that $\sup_{\beta \in \mathscr{B}} \|\hat{Q}(\beta) - Q(\beta)\| \xrightarrow{p} 0$. The conclusion now follows by the Wald argument for consistency of extremum estimators. *Q.E.D.*

PROOF OF LEMMA 5.3: It follows by an argument like that of Lemma 5.2 that $\sup_{\beta \in \mathscr{B}} \|\partial \hat{m}_n(\beta)/\partial \beta - E[\partial m(z, \beta, h_0)/\partial \beta]\| \xrightarrow{p} 0$ for a neighborhood $\mathscr{B}$ of β_0. The conclusion then follows from the conclusion of Lemma 5.1 by a standard minimum distance argument, as in Newey and McFadden (1994). *Q.E.D.*

PROOF OF LEMMA 5.4: Using the argument of, e.g. Powell, Stock, and Stoker (1989), it will suffice for $\hat{\Omega} \xrightarrow{p} \Omega$ that $\sum_{i=1}^n \|m(z_i, \hat{\beta}, \hat{h}) - m(z_i, \beta_0, h_0)\|^2/n \xrightarrow{p} 0$ and $\sum_{i=1}^n \|\hat{\alpha}(z_i) - \alpha(z_i)\|^2/n \xrightarrow{p} 0$. The second of these holds by assumption, and the first follows from the conditions of the Lemma. The result then follows by consistency of $\hat{M}$ and $\hat{W}$ and nonsingularity of $M'WM$. *Q.E.D.*

PROOF OF THEOREM 6.1: For simplicity the result will be proven for the case where $g(x)$ is a scalar. First, the hypotheses of Lemma 5.1 will be verified. Let $\Sigma_K = E[p^K(x)p^K(x)']$, $\Sigma = \Sigma_{\hat{K}}$, $\hat{\Sigma} = \sum_{i=1}^n p^{\hat{K}}(x_i)p^{\hat{K}}(x_i)'/n$. By Lemma A.9 of Newey (1994a) and Assumption 6.6,

$$\text{(A.1)} \qquad \|\hat{\Sigma} - \Sigma\| = O_p\left(\zeta_0(\bar{K})^2/\sqrt{n}\right) = o_p(1).$$

Also, by Assumptions 6.1–6.3, $\zeta_0(\bar{K})^4/n \to 0$, and Theorem 3.2 of Newey (1994a),

$$\text{(A.2)} \qquad |\hat{g} - g_0|_d = O_p\left(\zeta_d(\bar{K})\left[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}\right]\right).$$

Therefore, Assumption 6.4 implies Assumption 5.1 for $h = g$ and $D(z, h) = D(z, g; \beta, h)$.

Next, let $D_i^K = (D(z_i, p_{1K}), \ldots, D(z_i, p_{KK}))'$. Let π_K be such that for $g_K(x) = p^K(x)'\pi_K$, $|g_0 - g_K|_d \leq CK^{-\alpha}$ for the d and α of Assumption 6.5. Then by the Markov inequality, for $\Delta(z, g) = D(z, g) - E[D(z, g)]$ and $\mathscr{K} = \{\underline{K}, \underline{K} + 1, \ldots, \bar{K}\}$, with probability approaching one (w.p.a.1),

$$\text{(A.3)} \qquad \left\| \sum_{i=1}^n \Delta(z_i, g_{\hat{K}} - g_0)/\sqrt{n} \right\|^2 \leq \sum_{K \in \mathscr{K}} \left\| \sum_{i=1}^n \Delta(z_i, g_K - g_0)/\sqrt{n} \right\|^2$$
$$= O_p\left(\sum_{k \in \mathscr{K}} E\left[\|\Delta(z_i, g_K - g_0)\|^2 \right] \right)$$
$$= O_p\left(\sum_{K \in \mathscr{K}} K^{-2\alpha} \right) = o_p(1).$$

Let $\Delta_i^K = D_i^K - E[D_i^K]$. By Assumption 6.2, Δ_i^K is a subvector of Δ_i^{K+1} for all K with $\underline{K}(n) \leq K < K + 1 \leq \bar{K}(n)$. Then by the Markov inequality and Assumption 6.5, $\|\sum_{i=1}^n \Delta_i^{\hat{K}}\sqrt{n}\| \leq \|\sum_{i=1}^n \Delta_i^{\bar{K}}/\sqrt{n}\| = O_p((\sum_{k=1}^{\bar{K}} |p_{kK}|_d^2)^{1/2})$, so by linearity of $D(z, g)$ and Lemma A.8 of Newey (1994a),

$$\text{(A.4)} \qquad \left\| \sum_{i=1}^n \Delta(z_i, \hat{g} - g_K)/\sqrt{n} \right\| = \left\| \left(\sum_{i=1}^n \Delta_i^{\hat{K}}/\sqrt{n} \right)'(\hat{\pi} - \pi_{\hat{K}}) \right\|$$
$$\leq \left\| \sum_{i=1}^n \Delta_i^{\hat{K}}/\sqrt{n} \right\| \|\hat{\pi} - \pi_{\hat{K}}\| \leq \left\| \sum_{i=1}^n \Delta_i^{\bar{K}}/\sqrt{n} \right\| \|\hat{\pi} - \pi_{\hat{K}}\|$$
$$= O_p\left(\left(\sum_{k=1}^{\bar{K}} |p_{kK}|_d^2 \right)^{1/2} \left[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha} \right] \right) = o_p(1).$$

It then follows by the triangle inequality that Assumption 5.2 is satisfied.

Next, partition z as $z = (\tilde{z}, x)$, and for a random variable $a(z)$ let

$$\text{(A.5)} \qquad \int a(z)\, d\hat{F} = \int p^K(x)'\hat{\Sigma}^- \sum_{i=1}^n p^K(x_i) a(\tilde{z}_i, x)\, dF_0.$$

Note that $p^K(x)'\hat{\Sigma}^- \Sigma_{i=1}^n p^K(x_i) = 1$ by existence of a linear combination of $p^K(x)$ that is constant. Then by Assumption 6.6, for $\alpha(z) \| = \delta(x)[y - g_0(x)]$,

$$\text{(A.6)} \qquad E[D(z, \hat{g} - g_0)] = \int \delta(x)[\hat{g}(x) - g_0(x)]\, dF_0$$

$$= \int p^K(x)\hat{\Sigma}^- \sum_{i=1}^n p^K(x_i)\delta(x)[y_i - g_0(x)]\, dF_0$$

$$= \int \alpha(z)\, d\tilde{F},$$

giving the first part of Assumption 5.3. To show the second part, let $g_K(x) = p^K(x)'\Sigma^{-1}E[p^K(x)g(x)]$, $\psi_K = E[\delta(x)p^K(x)]$, $\Psi = \Psi_{\hat{K}}$, and $\delta_K(x) = p^K(x)'\Sigma^{-1}\Psi$. By orthogonality of least squares projections and Assumption 6.6,

$$\text{(A.7)} \qquad \left\| \sqrt{n} \int \delta(x)[g_{\hat{K}}(x) - g_0(x)]\, dF_0 \right\|^2$$

$$= \left\| \sqrt{n} \int [\delta_{\hat{K}}(x) - \delta(x)][g_{\hat{K}}(x) - g_0(x)]\, dF_0 \right\|^2$$

$$\leq n \int \| \delta_{\hat{K}}(x) - \delta(x) \|^2\, dF_0 \cdot \int \| g_{\hat{K}}(x) - g_0(x) \|^2\, dF_0$$

$$\leq nE\left[\| \delta_{\underline{K}}(x) - \delta(x) \|^2 \right] \cdot E\left[\| g_{\underline{K}}(x) - g_0(x) \|^2 \right] \to 0.$$

Also, note that $\int \delta(x)[\hat{g}(x) - g_{\hat{K}}(x)]\, dF_0 = \Psi'(\hat{\pi} - \pi_{\hat{K}}) = \Psi'\hat{\Sigma}^{-1}\Sigma_{i=1}^n p^{\hat{K}}(x_i)[y_i - g_{\hat{K}}(x_i)]/n$. Note that for each K, $E[p^K(x)\{y - g_K(x)\}] = 0$. Therefore,

$$\text{(A.8)} \qquad E\left[1(\hat{K} \in \mathscr{K}) \left\| \Sigma^{-1} \sum_{i=1}^n p^{\hat{K}}(x_i)[y_i - g_{\hat{K}}(x_i)]/\sqrt{n} \right\|^2 \right]$$

$$\leq E\left[\sum_{\mathscr{K}} \left\| \Sigma_K^{-1/2} \sum_{i=1}^n p^K(x_i)[y_i - g_K(x_i)]/\sqrt{n} \right\|^2 \right]$$

$$= \sum_{\mathscr{K}} E\left[p^K(x)'\Sigma_K^{-1} p^K(x)\{y - g_K(x)\}^2 \right]$$

$$\leq C \sum_{\mathscr{K}} E\left[p^K(x)'\Sigma_K^{-1} p^K(x) \right] + \zeta_0(K)^2 E\left[\{g(x) - g_K(x)\}^2 \right]$$

$$= \sum_{\mathscr{K}} \left\{ K + \zeta_0(K)^2 E\left[(g_0(x) - g_K(x))^2 \right] \right\}.$$

Note that $\|\Psi'\Sigma^{-1}\|^2 \leq C\Psi'\Sigma^{-1}\Psi \leq CE[\|\delta(x)\|^2]$, and that $\|\Psi'\hat{\Sigma}^{-1} - \Psi'\Sigma^{-1}\| \leq \|\Psi'\Sigma^{-1}\|\,\|(\hat{\Sigma} - \Sigma)\hat{\Sigma}^{-1}\| \xrightarrow{p} 0$, *so that* $\|\Psi'\hat{\Sigma}^{-1}\| = O_p(1)$. Then by Assumption 6.6,

$$\text{(A.9)} \qquad \Psi'(\hat{\Sigma}^{-1} - \Sigma^{-1}) \sum_{i=1}^{n} p^{\hat{K}}(x_i)[y_i - g_{\hat{K}}(x_i)]/\sqrt{n}$$

$$= \Psi'\hat{\Sigma}^{-1} \sum_{i=1}^{n} p^{\hat{K}}(x_i)[y_i - g_{\hat{K}}(x_i)]/\sqrt{n}$$

$$= O_p\left(\left[\zeta_0(\bar{K})^4\left\{\Sigma_K\left\{K + \zeta_0(K)^2 E\left[\left(g_0(x) - g_K(x)\right)^2\right]\right\}/n\right]^{1/2}\right) = o_p(1).$$

Also, for $\delta_K(x) = \Psi_K'\Sigma_K^{-1}p^K(x)$,

$$\text{(A.10)} \qquad E\left[\sum_{\mathscr{K}}\left\|\sum_{i=1}^{n}\{\delta_K(x_i)[y_i - g_K(x_i)] - \delta(x_i)[y_i - g_0(x_i)]\}/\sqrt{n}\right\|^2\right]$$

$$\leq \sum_{\mathscr{K}} E\left[\left\|\{\delta_K(x)[y - g_K(x)] - \delta(x)[y - g_0(x)]\right\|^2\right]$$

$$\leq C\sum_{\mathscr{K}} E\left[\|\delta_K(x) - \delta(x)\|^2 u^2\right] + E\left[\|\delta_K(x)[g_0(x) - g_K(x)]\|^2\right]$$

$$\leq C\sum_{\mathscr{K}} E\left[\|\delta_K(x) - \delta(x)\|^2\right] + E\left[\left\|\Psi_K'\Sigma_K^{-1}p^K(x)\{g_0(x) - g_K(x)\}\right\|^2\right]$$

$$\leq o(1) + \sum_{\mathscr{K}}\left\|\Psi_K'\Sigma_K^{-1}\Psi_K\right\|^2 E\left[p^K(x)'\Sigma_K^{-1}p^K(x)\left(g_0(x) - g_K(x)\right)^2\right]$$

$$\leq o(1) + \sum_{\mathscr{K}}\zeta_0(K)^2 E\left[\left(g_0(x) - g_K(x)\right)^2\right].$$

It then follows that $\|\Sigma_{i=1}^{n}\{\delta_{\hat{K}}(x_i)[y_i - g_{\hat{K}}(x_i)] - \delta(x_i)[y_i - g_0(x_i)]\}/\sqrt{n} \xrightarrow{p} 0$. The second part of Assumption 5.3 then follows by the triangle inequality and eqs. (A.7), (A.9), and (A.10).

Thus, each of Assumptions 5.1–5.3 are satisfied. Also, Assumptions 5.4–5.6 hold by hypothesis, and by Assumption 6.4, $\|\hat{h} - h_0\| = |\hat{g} - g_0|_d = O_p(\zeta_d(\bar{K})[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}]) = o_p(1)$. Therefore, the first conclusion follows by Lemma 5.3.

To show the second conclusion, use Assumption 6.7 and $\hat{\beta} \xrightarrow{p} \beta_0$ to obtain

$$\text{(A.11)} \qquad \sum_{i=1}^{n}\left\|m\left(z_i, \hat{\beta}, \hat{h}\right) - m(z_i, \beta_0, h_0)\right\|^2/n \leq \left(\left\|\hat{\beta} - \beta_0\right\|^2 + |\hat{g} - g_0|_d^2\right)\sum_{i=1}^{n} b(z_i)^2/n \xrightarrow{p} 0.$$

Also, let $\hat{\Psi}$ be as specified in the text, $\tilde{\Psi}_K = \Sigma_{i=1}^{n} D_i^K/n$, and $\Psi_K = E[D_i^K]$, where D_i^K was defined above. By Assumption 6.7, w.p.a.1,

$$\|\hat{\Psi} - \hat{\Psi}_{\bar{K}}\| \leq \left[\sum_{i=1}^{n} b(z_i)/n\right]\left(\sum_{k \leq \bar{K}} |p_{k\bar{K}}|_d\right)\left(\|\hat{\beta} - \beta_0\| + |\hat{g} - g_0|_d\right)$$

$$= O_p\left(E[b(z)]\left(\sum_{k \leq \bar{K}} |p_{k\bar{K}}|_d\right)\zeta_d(\bar{K})\left[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}\right]\right) = o_p(1).$$

Also, by the Markov inequality and Assumption 6.5,

$$\|\tilde{\Psi}_{\hat{K}} - \Psi\|^2 \leq \|\tilde{\Psi}_{\bar{K}} - \Psi_{\bar{K}}\|^2 = O_p\left(\left(\sum_{k \leq \bar{K}} |p_{k\bar{K}}|_d^2\right)/n\right) = o_p(1).$$

Then by the triangle inequality and $\bar{K} \to \infty$, the equalities in eq. (A.11) continue to hold with Ψ replacing $\tilde{\Psi}_{\hat{K}}$. Also, as discussed above, $\|\hat{\Sigma} - \Sigma\| = O_p(\zeta_0(\bar{K})^2/\sqrt{n})$. Furthermore, by the smallest

eigenvalue of Σ_K bounded below, $\|\Psi_K'\Sigma_K^{-1}\|^2 \leq C\text{tr}(\Psi_K'\Sigma_K^{-1}\Psi_K) \leq CE[\|\delta(x)\|^2]$, by $\Psi_K = E[\delta(x)p^K(x)']$. It then follows by the largest eigenvalue of $\hat{\Sigma}^{-1}$ bounded in probability that

$$\text{(A.12)} \quad \left\|\hat{\Psi}'\hat{\Sigma}^{-1} - \Psi_{\bar{K}}'\Sigma_{\bar{K}}^{-1}\right\| \leq \left\|\left(\hat{\Psi} - \Psi_{\hat{K}}\right)'\hat{\Sigma}^{-1}\right\| + \left\|\Psi_{\hat{K}}'\left(\hat{\Sigma}^{-1} - \Sigma_{\bar{K}}^{-1}\right)\right\|$$

$$\leq O_p(1)\|\hat{\Psi} - \Psi_{\hat{K}}\| + \left\|\Psi_{\hat{K}}'\Sigma_{\bar{K}}^{-1}\left(\hat{\Sigma} - \Sigma_{\hat{K}}\right)\hat{\Sigma}^{-1}\right\|$$

$$\leq O_p(1)\left(\|\hat{\Psi} - \Psi_{\hat{K}}\| + \|\Psi_{\hat{K}}'\Sigma_{\bar{K}}^{-1}\|\,\|\hat{\Sigma} - \Sigma_{\hat{K}}\|\right) = o_p(1).$$

Therefore, $\|\hat{\Psi}'\hat{\Sigma}^{-1}\| = O_p(1)$. Also, by Lemma A.8 of Newey (1994a) and Assumption 6.7,

$$\text{(A.13)} \quad \sum_{i=1}^n \left\|\hat{\Psi}'\hat{\Sigma}^{-1}p^{\hat{K}}(x_i)(y_i - \hat{g}(x_i)) - \hat{\Psi}'\hat{\Sigma}^{-1}p^{\hat{K}}(x_i)u_i\right\|^2/n$$

$$\leq \|\hat{\Psi}'\hat{\Sigma}^{-1}\|^2 \sum_{i=1}^n \|p^{\hat{K}}(x_i)\|^2\|\hat{g}(x_i) - g_0(x_i)\|^2/n$$

$$\leq O_p(1)\zeta_0(\bar{K})^2 \sum_{i=1}^n \|\hat{g}(x_i) - g_0(x_i)\|^2/n$$

$$= O_p\left(\zeta_0(\bar{K})^2\left[\bar{K}/n + \underline{K}^{-2\alpha}\right]\right) = o_p(1).$$

Also, by Assumption 6.6,

$$\text{(A.14)} \quad \sum_{i=1}^n \left\|\hat{\Psi}'(\hat{\Sigma}^{-1} - \Sigma^{-1})p^{\hat{K}}(x_i)u_i\right\|^2/n$$

$$\leq \|\hat{\Psi}'\hat{\Sigma}^{-1}\|^2\|\hat{\Sigma} - \Sigma_{\hat{K}}\|^2 \sum_{i=1}^n \left\|\Sigma_{\bar{K}}^{-1}p^{\hat{K}}(x_i)u_i\right\|^2/n$$

$$= O_p\left(\left[\zeta_0(\bar{K})^4/n\right]E\left[\left\|\Sigma_K^{-1/2}p^{\bar{K}}(x_i)u_i\right\|^2\right]\right) = O_p\left(\zeta_0(\bar{K})^4\bar{K}/n\right) = O_p(1).$$

Similarly,

$$\text{(A.15)} \quad \sum_{i=1}^n \left\|\left(\hat{\Psi} - \Psi_{\hat{K}}\right)'\Sigma_{\bar{K}}^{-1}p^{\hat{K}}(x_i)u_i\right\|^2/n$$

$$= O_p\left(\left(\sum_{k \leq \bar{K}} |p_{k\bar{K}}|_d\right)^2\left\{E[b(z)]^2\zeta_k(\bar{K})^2\left[\bar{K}/n + \underline{K}^{-2\alpha}\right] + n^{-1}\right\}\bar{K}\right) = o_p(1).$$

Furthermore, for $\delta_K(x) = \Psi_K'\Sigma_K^{-1}p^K(x)$ as above, by Assumption 6.6, w.p.a.1

$$\text{(A.16)} \quad \sum_{i=1}^n \left\|\left[\Psi_{\hat{K}}'\Sigma_{\bar{K}}^{-1}p^{\hat{K}}(x_i) - \delta(x_i)\right]u_i\right\|^2/n = \sum_{i=1}^n \|\delta_{\hat{K}}(x_i) - \delta(x_i)\|^2\|u_i\|^2/n$$

$$\leq \sum_{\mathscr{K}}\sum_{i=1}^n \|\delta_K(x_i) - \delta(x_i)\|^2\|u_i\|^2/n$$

$$= O_p\left(\sum_{\mathscr{K}} E\left[\|\delta_K(x_i) - \delta(x_i)\|^2\right]\right) = o_p(1).$$

Then $\sum_{i=1}^n \|\hat{\alpha}(z_i) - \alpha(z_i)\|^2/n \xrightarrow{p} 0$ follows by equations (A.13)–(A.16) and the triangle inequality. The second conclusion then follows by Lemma 5.4. *Q.E.D.*

PROOF OF THEOREM 7.1: The proof proceeds first by showing consistency of $\hat{\beta}$ and then by verifying the hypotheses of Theorem 6.1. Note that $\Phi(\cdot)$ is differentiable to all orders and that the derivatives are bounded, so that $h_{t0}(x)$ $(t = 1, 2)$ are continuously differentiable to order s. By

choosing $p^K(x)$ to be products of polynomials that are orthonormal with respect to a uniform weight on the support of x, it follows by equation (7.1) and Lemma A.15 of Newey (1994a) that Assumption 6.2 is satisfied, and that $\zeta_0(K) = K$ and $|p_{kK}|_0 = K^{1/2}$. Also, by Lorentz (1986, Theorem 8), Assumption 6.3 is satisfied for $d = 0$ and $\alpha = s/2r$. Then by Theorem 3.1 of Newey (1994a),

$$(A.17) \qquad |\hat{g} - g_0|_0 = O_p\left(\bar{K}\left[\left(\bar{K}/n\right)^{1/2} + \underline{K}^{-\alpha}\right]\right) = o_p(1).$$

Also, note that $x_t + \rho(x)$ is bounded, so that $g_{t0}(x) = \Phi([x_t + \rho(x)]/\sigma_{t0})$ is bounded away from zero and one, and hence so is $g_t(x)$ for $|g_t - g_{t0}|_0$ small enough. Furthermore, $\Phi^{-1}(\cdot)$ is continuously differentiable to all orders on any set where its argument is bounded away from zero and one. It then follows by a standard expansion that $n^{-1}\sum_{i=1}^n \partial m(z_i, \beta, \hat{h})/\partial\beta$, which does not depend on β, converges in probability to $M = E[J(x)J(x)']$, which is nonsingular. It follows similarly that $n^{-1}\sum_{i=1}^n \{m(z_i, \beta, \hat{h}) - [\partial m(z_i, \beta, \hat{h})/\partial\beta]\beta\}$ converges in probability. Then $\hat{\beta} \xrightarrow{p} \beta_0$ follows by the usual least squares arguments. For $\nu(a) = \Phi^{-1}(a)$, let $\beta = (\gamma', \sigma)'$ and $m(z, \beta, h) = 1(0 < g_1(x) < 1, 0 < g_2(x) < 1)(x_1' - x_2', \nu(g_2(x)))'[\nu(g_1(x)) - \sigma\nu(g_2(x)) - (x_1 - x_2)'\gamma]$. Set $d = 0$. By x bounded, Assumption 5.4 is satisfied with $b(z)$ and $\bar{b}(z)$ equal to constants. Let $\mathscr{B}$ be any compact set containing β_0 in its interior. Assumption 5.5, with $\hat{W} = W = I$ follows by $E[J(x)J(x)']$ nonsingular. Assumption 5.6 follows similarly. Assumptions 6.1–6.3 were checked above. For Assumption 6.4, let

$$(A.18) \qquad D(z, g; \beta, \tilde{g}) = \left(x_1' - x_2', \nu(g_2(x))\right)'\left[\nu_a(\tilde{g}_1(x))g_1(x) - \sigma\nu_a(\tilde{g}_2(x))g_2(x)\right]$$
$$+ \left(O, \nu_a(\tilde{g}_2(x))\right)'\left[\nu(\tilde{g}_1(x)) - \sigma\nu(\tilde{g}_2(x)) - (x_1 - x_2)'\gamma\right]g_2(x).$$

It follows by a standard mean value expansion, differentiating with respect to possible values for $g_t(x)$, that Assumption 6.4(i) is satisfied, with $b(z)$ equal to a positive constant. Assumption 6.4(ii) follows by the rate conditions, since $b(z)$ is positive and $\sqrt{n}\,\zeta_0(\bar{K})^2[\bar{K}/n + \underline{K}^{-2\alpha}] = \bar{K}^3/\sqrt{n} + \sqrt{n}\bar{K}^2\underline{K}^{-2\alpha} \to 0$. Also, Assumption 6.5 is satisfied, with $b(z)$ some positive constant and $d = 1$, since $\alpha = s/2r > 5/2$ is implied by the rate conditions, and $(\sum_{k=1}^{\bar{K}}|p_{kK}|_d^2)[\bar{K}/n + \underline{K}^{-2\alpha}] \leqslant \bar{K}^2[(\bar{K}/n) + \underline{K}^{-2\alpha}] \to 0$.

To check Assumption 6.6, note that by equation (A.18), $E[D(z, g; g_0, \beta_0)] = E[\delta(x)g(x)]$ for $\delta(x) = (\delta_1(x), \delta_2(x))'$ and $\delta_t(x) = J(x)(-\sigma_0)^{t-1}\nu_a(g_{t0}(x))$. It is easy to argue that $\delta_t(x)$ is continuously differentiable of order s on the support, so that by Lorentz (1986, Theorem 8), there exists π_K and ξ_K with $E[\|\delta(x) - \xi_K p^K(x)\|^2] = O(K^{-2\alpha})$ and $E[\|g_0(x) - \pi_K p^K(x)\|^2] = O(K^{-2\alpha})$. Assumption 6 then follows by $n\underline{K}^{-4\alpha} \to 0$, $\zeta_0(\bar{K})^4\bar{K}^2/n = O(\bar{K}^6/n) \to 0$, $\zeta_0(K)^2K^{-2\alpha} \leqslant K^{2-2\alpha} \leqslant K^{-3}$ (using $\alpha > 5/2$), and $K^{-2\alpha} \leqslant K^{-5}$. The first conclusion then follows from the first conclusion of Theorem 6.1. The second conclusion follows from equation (A.17) and arguments analogous to the consistency proof. *Q.E.D.*

PROOF OF THEOREM 7.2: Assumptions 5.4–5.6 are satisfied for $\|h\| = |g|_1$. Assumption 6.1 is satisfied by hypothesis. By Lemma A.15 of Newey (1994a), Assumption 6.2 is satisfied by $p_{kK}(x)$ equal to products of polynomials that are orthonormal with respect to $\prod_{j=1}^r[(x - x_{lj})(x_{uj} - x)]^{\lambda}$, with $\zeta_d(K) = K^{1+\lambda+2d}$ and $|p_{kK}|_d = K^{.5+\lambda+2d}$. By Lemma A.13 of Newey (1994a), Assumption 6.3 is satisfied with α_d equal to any positive constant. Next, $m(z, \beta, h) = \partial g(x)/\partial x - \beta$ is linear in g, so that Assumption 6.4 is satisfied with $b(z) = 0$ and $D(z, g; \beta, \tilde{g}) = \partial g(x)/\partial x$. Also, by the rate conditions on $\bar{K}$ and $\underline{K}$, for any $a, b > 0$ there is an α large enough that $n^a\bar{K}^b\underline{K}^{-\alpha} \to 0$. For Assumption 6.5, note that $\|D(z, g; \beta, \tilde{g})\| \leqslant |g|_1$, and that there is an α large enough that $\sum_{K=\underline{K}}^{\bar{K}} K^{-2\alpha} \to 0$ and $(\sum_{k=1}^{\bar{K}}(|p_{kK}|_1)^2)^{1/2}[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}] = \bar{K}^{3+\lambda}[(\bar{K}/n)^{1/2} + \underline{K}^{-\alpha}] = (\bar{K}^{7+2\lambda}/n)^{1/2} + o(1) \to 0$. For Assumption 6.6, the integration by parts argument in Section 4 gives $E[D(z, g; \beta_0, g_0)] = E[\partial g(x)/\partial x] = E[\{-f(x)^{-1}\partial f(x)/\partial x\}g(x)]$. Also, by Lemma A.3 of Newey (1994a), the set $\mathscr{S}$ of additive interactive functions whose individual components have finite mean-square is closed in mean square, so that $E[D(z, g; \beta_0, g_0)] = E[\delta(x)g(x)]$ for $\delta(x) = -\prod(f(x)^{-1}\partial f(x)/\partial f(x)/\partial x|\mathscr{S})$. By $\delta(x)$ having finite mean square and standard polynomial approximation results, there exists ξ_K such that $E[\|\delta(x) - \xi_K p^K(x)\|^2] \to$ as $K \to \infty$. The remainder of Assumption 6.6 then follows by arguments like those above. Also, Assumption 6.7 is satisfied with $b(z) = 0$. Therefore, the conclusion follows from Theorem 6.1. *Q.E.D.*

REFERENCES

ABRAMOWITZ, M., AND I. A. STEGUM, EDS. (1972): *Handbook of Mathematical Functions*. Washington, D.C.: Commerce Department.

AHN, H., AND C. MANSKI (1993): "Distribution Theory for the Analysis of Binary Choice Under Uncertainty with Nonparametric Estimators in Expectations," *Journal of Econometrics*, 56, 291–321.

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

——— (1994): "Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity," *Econometrica*, 62, 43–72.

BICKEL, P., C. A. J. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: John Hopkins University Press.

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

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.

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 Statistic 19. Berlin: Springer.

GOURIEROUX, C., A. MONFORT, AND A. TROGNON (1984): "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

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

HARDLE, W., W. HILDENBRAND, AND M. JERISON (1991): "Empirical Evidence on the Law of Demand," *Econometrica*, 59, 1525–1594.

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.

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. HASMINSKII (1981): *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag.

ICHIMURA, H. (1993): "Estimation of Single Index Models," forthcoming, *Journal of Econometrics*, 58, 71–120.

KIM, J., AND D. POLLARD (1989): "Cube Root Asymptotics," *Annals of Statistics*, 18, 191–219.

KLEIN, R. W., AND R. S. SPADY (1993): "An Efficient Semiparametric Estimator of the Binary Response Model," *Econometrica*, 61, 387–422.

KOSHEVNIK, Y. A., AND B. Y. LEVIT (1976): "On a Non-parametric Analogue 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.

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

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

——— (1994a): "Convergence Rates for Series Estimators," forthcoming in *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C. R. Rao*.

——— (1994b): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, forthcoming.

NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation," *Handbook of Econometrics*, Vol. 4, forthcoming.

NEWEY, W. K., AND P. A. RUUD (1991): "Density Weighted Least Squares Estimation," Working Paper, MIT Department of Economics.

NEWEY, W. K., AND T. M. STOKER (1993): "Efficiency of Average Derivative Estimators and Index Models," *Econometrica*, 61, 1199–1223.

PFANZAGL, J., AND WEFELMEYER (1982): *Contributions to a General Asymptotic Statistical Theory*. New York: Springer-Verlag.

POWELL, M. J. D. (1981): *Approximation Theory and Methods*. Cambridge, England: Cambridge University Press.

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

PRAKASA RAO, B. L. S. (1983): *Nonparametric Functional Estimation*. New York: Academic Press.

REEDS, J. A. (1976): "On the Definition of von Mises Functionals," Ph.D. Thesis, Harvard University.

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

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

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

SCHUMAKER, L. L. (1981): *Spline Functions: Basic Theory*. New York: Wiley.

SEVERINI, T. A., AND W. H. WONG (1992): "Profile Likelihood and Conditionally Parametric Models," *Annals of Statistics*, 20, 1768–1802.

STOKER, T. M. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

——— (1991): "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods*, ed. by W. A. Barnett, J. L. Powell, and G. Tauchen. Cambridge: Cambridge University Press.

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

——— (1985): "Additive Regression and other Nonparametric Models," *Annals of Statistics*, 13, 689–705.

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.