SERIES ESTIMATION
OF REGRESSION FUNCTIONALS

Whitney K. Newey
Princeton University
Bell Communications Research

Econometric Research Program
Research Memorandum No. 348

June 1989

Financial support was provided by the National Science Foundation and the Sloan Foundation. R. Klein and J. Powell provided helpful comments.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, NJ 08544, USA

Series Estimation of Regression Functionals

Whitney K. Newey

ABSTRACT

There are a number of important models where estimators of parameters of interest depend on predicted values for a conditional expectation. Examples include the residual variance, microeconometric expectations models, generalized least squares corrections for heteroskedasticity of unknown form, efficient instrumental variables estimation of nonlinear models, and estimation of Euclidean parameters of additive semiparametric regression models. The purpose of this paper is to analyze the general asymptotic properties of sample averages of functions of a nonparametric regression, and show how these results are useful in these examples. The specific nonparametric regression method considered here is series estimation, e.g. polynomial regression. New results on expectations models, heteroskedasticity corrected generalized least squares, and additive semiparametric regression models are given.

Keywords: Nonparametric Regression, Series Estimators, Regression Functionals, Residual Variance, Microeconometric Expectations, Generalized Least Squares, Best Instrumental Variables, Additive Semiparametric Regression.

1. Introduction

There are a number of important models where estimators of parameters of interest depend on predicted values for a conditional expectation. A statistical example is a functional of the residuals of a nonparametric regression, such as the residual variance. An econometric example is microeconometric expectations models, such as those of Manski (1988). Examples also include generalized least squares corrections for heteroskedasticity of unknown form (HGLS), as in Carroll (1982) and Robinson (1987), efficient instrumental variable estimators of nonlinear models, as in Newey (1989a), and estimation of the Euclidean parameters of the additive semiparametric regression (ASR) model of Engle, Granger, Rice, and Weiss (1984), as in N. Heckman (1986), Rice (1986), Schick (1986), Robinson (1988), Chamberlain (1986), and Andrews (1988). The purpose of this paper is to analyze the general asymptotic properties of sample averages of functions of a nonparametric regression, and show how these results are useful in these examples. The specific nonparametric regression method considered here is series estimation, e.g. polynomial regression. New results on specific estimation methods include asymptotic distribution theory for estimators of microeconometric expectations models, a series based HGLS estimate and a proof of its asymptotic efficiency, and a generalization of Chamberlain's (1986) series estimator for ASR to nonlinear models and a proof of its asymptotic normality.

Related work on estimators of parameters of interest that depend on series estimates of conditional expectations includes Chamberlain (1986), Andrews (1988), and Newey (1987, 1988, 1989a). The estimators considered here are of analogous form, although the regularity conditions are different than those of Newey (1987, 1988) and Andrews (1988) in an important way. Because the estimators here depend only on the regression function, and not, say, on its derivatives, and because this regression is estimated from the same observations as the parameters of interest, it is possible to do without any assumption on the magnitude of the series terms or the smallest eigenvalue of their second moment matrix. Because of this feature, few conditions have to be imposed on the distribution of the regressors and on the form of the series. For example, the regressor distribution could be discrete with infinite support. In this respect, the results here are like those of Chamberlain (1986) and Newey (1989a), although the assumptions and conclusions here are stronger than those of Chamberlain (1986). Here $\sqrt{n}$-consistency is shown, unlike Chamberlain's (1986) ASR results.

A fundamental theoretical result of this paper is the limiting distribution of a sample average that depends on nonparametric regression estimates. Intuitively, the form of this distribution should not depend on the form of the nonparametric regression estimator, e.g. series rather than kernels. Therefore, it is of some interest to discuss this result in a wider context than series estimates. Such a discussion is given in Section 2, where the efficient influence function is derived, in the sense of Koshevnik and Levitt (1976) and Pfanzagl (1982), for the expectation of a function that depends on unknown conditional expectation. If the distribution of the data is unrestricted, then the efficient influence function gives the limiting distribution of any estimator satisfying certain regularity conditions; e.g. see Bickel, Klaassen, Ritov, and Wellner (1989) or Newey (1989b) for exposition.

The remainder of the paper focuses on series estimates of conditional expectations. Section 3 discusses series nonparametric regression and some of its asymptotic properties. Section 4 presents general consistency and asymptotic normality Theorems for sample averages of functions of the predicted regression and parameters of interest, verifying the conjecture of Section 2. Section 5 applies these results to derive asymptotic distribution theory for the examples. Section 6 discusses possible extensions.

2. The Efficient Influence Function.

The fundamental theoretical results of this paper concern the limiting behavior of an object of the form

(2.1) $$A_n(\beta) = \sum_{i=1}^n a^n(z_i,\beta,\hat{g}(x_i,\beta))/n,$$

where z_i is a data observation from an i.i.d. sequence $(z_1,\ldots,z_n)$, β a vector of parameters with true value β_0, and $\hat{g}(x_i,\beta)$ is a nonparametric estimate of the conditional expectation $g_0(x,\beta) = E[w(z,\beta,\eta_0)|x]$ for a vector of functions $w(z,\beta,\eta)$. The estimates $\hat{g}(x_i,\beta)$ are assumed to be calculated from observations $w(z_i,\beta,\hat{\eta})$, x_i, $(i = 1,\ldots,n)$, where $\hat{\eta}$ is an estimator of η_0. The function $a^n(z,\beta,g)$ depends on the sample size n in order to allow for the imposition of smoothness by trimming, which will be useful in some of the examples. For notational simplicity, $a^n(z,\beta,g)$ will be taken to be a scalar in much of Sections 2, 3, and 4; results for the vector case can be obtained by applying the scalar results to each component.

It is easy to formulate a conjecture concerning the limiting value of $A_n(\beta)$. As long as $\hat{g}(x,\beta)$ is consistent for $g_0(x,\beta)$ in an appropriate sense, the law of large numbers would suggest that $A_n(\beta)-\bar{A}_n(\beta) = o_p(1)$, where $\bar{A}_n(\beta) = E[a^n(z,\beta,g_0(x,\beta))]$. However, finding the limiting distribution of $\sqrt{n}[A_n(\beta_0)-\bar{A}_n(\beta_0)]$ is more difficult, because estimation of $g_0(x,\beta)$ might affect this distribution.

One conjecture for the limiting distribution makes use of the semiparametric asymptotic variance bound for a functional $\mu_0 = E[a(z,g_0(x))]$, where $g_0(x) = g_0(x,\beta_0)$ and the presence of η is ignored. When this functional is pathwise differentiable in the sense of Pfanzagl (1982), and a certain linearity condition is satisfied, then there exists a function $d(z)$ such that the bound is $E[d(z)^2]$. In the efficiency literature d is referred to as the *efficient influence function.*

There is an important result concerning the relationship of the asymptotic distribution of an estimator $\hat{\mu}$ of μ_0 and the efficient influence function. Suppose that $\hat{\mu} = \sum_{i=1}^{n} a(z_i, \hat{g}(x_i))/n$ satisfies $\sqrt{n}(\hat{\mu}-\mu_0) = \sum_{i=1}^{n}\psi(z_i)/\sqrt{n} + o_p(1)$ for some function $\psi(z)$ with $E[\psi] = 0$ and $E[\psi^2] < \infty$. Then by the central limit theorem $\hat{\mu}$ will be asymptotically normal with variance $E[\psi^2]$. If the distribution of the data is unrestricted and $\hat{\mu}$ is locally regular, then $\psi(z) = d(z)$; e.g. see Newey (1989b) for exposition. Therefore, if the assumption of an unrestricted distribution is appropriate, one would expect that the asymptotic variance of $\hat{\mu}$ is $E[d^2]$.

The efficient influence function can be calculated by a projection argument. Let θ be the parameters of a smooth parametric submodel, which is a model that passes through the truth (the true distribution corresponds to some θ_0), and satisfies certain regularity conditions (including mean-square differentiability of the square-root of the likelihood). In such a submodel the parameter of interest μ, which here is the functional $E[a(z,g_0(x))]$, will depend on θ, i.e. for each value of θ it will take on a corresponding value $\mu(\theta)$. Let S_θ be the score for a parametric submodel at the truth, which can be thought of as the derivative of the log-likelihood for a single observation. (Here and henceforth the z argument may be suppressed for notational convenience.) A pathwise differentiable functional is one where there exists $d(z)$ such that for all regular parametric submodels $\mu(\theta)$ is differentiable at θ_0 and $\partial\mu(\theta_0)/\partial\theta = E[dS_\theta]$. Define the tangent set $\mathcal{S}$ to be the mean square closure of the union of random vectors of the form $c'S_\theta$, where c is a conformable constant vector and S_θ is the score for a regular parametric submodel. That is

$$\mathcal{S} = \{\Delta : E[\Delta^2] < \infty,\ \exists\ c_j,\ S_{\theta j} \text{ with } \lim_{j\longrightarrow\infty} E[(\Delta - c_j'S_{\theta j})^2] = 0\}$$

The efficient influence function is the projection d of d on $\mathcal{T}$ in the Hilbert space of random vectors with inner product $\langle\Delta|\tilde{\Delta}\rangle = E[\Delta\tilde{\Delta}]$, which exists as long as $\mathcal{T}$ is linear.

When the distribution of z is unrestricted, then one would expect that $\mathcal{T} = \{\Delta : E[\Delta^2] < \infty,\ E[\Delta] = 0\}$; the only restriction on the form of Δ should be the usual mean zero property of scores. When the tangent set takes this form the efficient influence function is simply $d = \text{d} - E[\text{d}]$.

To apply this calculation to the model at hand, it is necessary to find d for the functional $E[a(z,g_0(x))]$. To do so, note that for a parametric submodel, $g(x)$ will depend on θ, as $g(x,\theta) = E_\theta[w|x]$, where $E_\theta[\circ]$ denotes the expectation for the distribution indexed by θ. Thus, $\mu(\theta) = E_\theta[a(z,g(x,\theta))]$. Assuming the order of differentiation and integration can be interchanged (e.g. see Lemma 7.2 of Ibragimov and Hasminskii (1981)), $g(x,\theta)$ will be differentiable with derivative $\partial g(x,\theta_0)/\partial\theta = E[wS_\theta(w|x)'|x]$, where $S_\theta(w|x)$ is the score for the conditional density of w given x. Let $G(x) = \partial E[a(z,g)|x]/\partial g|_{g=g_0(x)}$, where g denotes a vector of real numbers, being a possible value of $g_0(x)$. It follows by the chain rule that

$$(2.2)\qquad \partial\mu(\theta_0)/\partial\theta = \{\ \partial E_\theta[a(z,g_0(x))]/\partial\theta + \partial E[a(z,g_0(x,\theta))]/\partial\theta\ \}|_{\theta=\theta_0},$$

$$= E[a(z,g_0(x))S_\theta] + E[G(x)'\partial g(x,\theta_0)/\partial\theta] = E[a_0S_\theta'] + E[G(x)'wS_\theta(w|x)],$$

where $a_0(z) = a(z,g_0(x))$. Under suitable regularity conditions, there will be a decomposition of the density into products of marginal and conditional densities, and a corresponding decomposition of the score as $S_\theta = S_\theta(x) + S_\theta(w|x) + S_\theta(\tilde{z}|x,w)$, where $S_\theta(x)$ is the score for the marginal density of x and $S_\theta(\tilde{z}|x,w)$ is the score for the conditional density of those elements $\tilde{z}$ of z other than x and w. Also, $E[S_\theta(\tilde{z}|x,w)|x,w] = E[S_\theta(w|x)|x] = 0$, so that $E[G(x)'(w-g_0(x))S_\theta(\tilde{z}|x,w)] = E[G(x)'g_0(x)S_\theta(w|x)] = 0$. Furthermore,

$E[G(x)'(w-g_0(x))S_\theta(x)] = 0$, so that by equation (2.2),

$$\partial\mu(\theta_0)/\partial\theta = E[a_0S_\theta] + E[G(x)'(w-g_0(x))S_\theta(w|x)]$$

$$= E[\{a_0 + G(x)'(w-g_0(x))\}S_\theta].$$

Thus, the functional μ will be differentiable with $d = a_0(z) + G(x)'[w-g_0(x)]$, so that the efficient influence function is

$$(2.3) \qquad d = d - E[d] = a_0(z)-E[a_0] + G(x)'[w-g_0(x)].$$

Then by the previous argument, it should follow that $\psi(z) = d(z)$, i.e.,

$$(2.4) \qquad \sqrt{n}(\hat{\mu}-\mu_0) = \textstyle\sum_{i=1}^n\{a(z_i,g_0(x_i))-E[a(z,g_0(x))]\}/\sqrt{n}$$

$$+ \textstyle\sum_{i=1}^n G(x_i)'[w_i-g_0(x_i)]/\sqrt{n} + o_p(1).$$

In the next section it is verified that this equation does hold for series estimates under appropriate regularity conditions, when $\hat{\eta}$ is not present. In some cases the formula will have to be modified to account for the presence of $\hat{\eta}$.

The form of equation (2.4) is of some interest. If the estimation of g did not affect the limiting distribution of $\hat{\mu}$, then one would expect to find that $\hat{\mu}$ was asymptotically equivalent to the first term following the equality. The second term is an adjustment that accounts for the estimation of $g(x)$. Note that this second term will be zero if $0 = G(x) = \partial E[a(z,g)|x]/\partial g|_{g=g_0(x)}$.

3. Series Nonparametric Regression and Its Properties.

The specific type of nonparametric regression method considered here is series estimation. Such estimates have a long history in statistics, and have recently received attention in econometrics; e.g. Gallant (1981). Series estimates of a conditional expectation $g_0(x) = E[w|x]$, where w is a scalar for now, make use of the first K terms,

$$(3.1) \qquad P^K(x) = (p_1(x),\dots,p_K(x))',$$

of a sequence of functions $(p_1(x), p_2(x), \dots)$. For notational convenience, the terms are restricted to be independent of K and n, although it would be straightforward to weaken this assumption. The estimate is calculated from observations w_i and x_i, $(i = 1, \dots, n)$, as the predicted value obtained from the regression of w_i on $P^K(x_i)$. Let $P = [P^K(x_1),\dots,P^K(x_n)]'$, where the K superscript for P is suppressed for convenience, and let $w = (w_1,\dots w_n)'$. A series estimate takes the form

$$(3.2) \qquad \hat{g}(x) = P^K(x)'(P'P)^{-}P'w,$$

where B^{-} denotes a generalized inverse of a matrix B.

The presence of the generalized inverse allows for perfect multicollinearity among the columns of P. One generalized inverse corresponds to the deletion of redundant columns of P and running the regression on the remainder, as is done by some regression software. It should be noted that $\hat{g}(x)$ of equation (3.2) may not be invariant to the choice of generalized inverse, although $\hat{g}(x_i)$ will be.

One example of $\hat{g}(x)$ is based on power series. Let the dimension of x be r. Let λ denote an r-dimensional vector of nonnegative integers and let $x^\lambda = (x_1)^{\lambda_1}\cdots(x_r)^{\lambda_r}$ denote a product of powers of the components of x.

A basis sequence for power series would take the form

$$(3.3)\qquad p_m(x) \equiv x^{\lambda(m)}, \quad (m = 1, 2, \ldots),$$

with distinct $\lambda(m)$. A more robust alternative, which puts less weight on outlying observations in x, can be obtained by weighting by a function $\omega(x)$ that is small for large values of x and/or replacing each component x_ℓ of x with a one-to-one, bounded function $v(x_\ell)$, such as $v(x_\ell) = \exp(x_\ell)/[1+\exp(x_\ell)]$. For $v(x) = (v(x_1),\ldots,v(x_r))'$ the resulting sequence is

$$(3.4)\qquad p_m(x) \equiv \omega(x)[v(x)^{\lambda(m)}], \quad (m = 1, 2, \ldots).$$

Trigonometric series are another example. Here x may have to be transformed to lie in $(0,2\pi)^r$. See Gallant (1981) for formulas.

The asymptotic distribution results to follow will make use of a convergence rate for the average distance between the estimated and true sample regression values. To obtain such a convergence rate it is essential to impose an approximation rate for the the conditional expectation $g_0(x)$ by the series terms.

Assumption 3.1: There exists π_K such that $(E[|g_0(x)-P^K(x)'\pi_K|^2])^{1/2} = O(K^{-\zeta})$ as $K \longrightarrow \infty$.

Primitive conditions for this assumption will entail restrictions on the support of x and smoothness conditions on $g_0(x)$. In the univariate x case this assumption will hold for power series if the support of x is a compact interval and $g_0(x)$ is ζ times continuously differentiable; see Powell (1981, Theorem 3.2). A literature search has not yet revealed an analogous result for the multivariate case. Nevertheless, if $g_0(x)$ is restricted to be analytic with geometric order bounds on the magnitude of derivatives, then an elementary Taylor expansion argument can be used to obtain an approximation rate. Let $v = v(x)$ and $x^{-1}(v)$ denote the inverse function. Denote the partial derivatives of a function $f(v)$ on the range of $v(x)$ by

$$D^{\lambda}f(v) = (\partial^{\lambda_1}/\partial v_1^{\lambda_1})\cdots(\partial^{\lambda_r}/\partial v_r^{\lambda_r})f(v),$$

where $\lambda = (\lambda_1,\cdots,\lambda_r)$ is a r-vector of nonnegative integers. The order of the derivative is $|\lambda| = \sum_{\ell=1}^{r}|\lambda_\ell|$. Also, let $\mathcal{O}(K) = \max_{m\le K}|\lambda(m)|$ denote the maximum order of the power series terms included in $P^K(x)$. The following result appears as Lemma 3.2 of Newey (1989a):

Suppose that $\mathcal{O}(K) = O(\alpha(K))$ *and that there is a set* $\mathcal{G}$ *such that for all* $g(x) \in \mathcal{G}$, $\mathscr{g}(v) \equiv \omega(x^{-1}(v))^{-1}g(x^{-1}(v))$ *can be taken to have a compact convex domain* $V(g)$ *containing the support of* $v(x_i)$. *Also suppose that there exists bounded* $V \supseteq \cup_{g\in\mathcal{G}}V(g)$ *that* $\omega(v)$ *is bounded and there exists* C *such that for all* λ, $D^{\lambda}\mathscr{g}(v)$ *exists and* $\sup_{g\in\mathcal{G}}\sup_{v\in V(g)}|D^{\lambda}\mathscr{g}(v)| \le C^{|\lambda|}$. *Then* $\lim_{K\to\infty}K^{\zeta}\sup_{g\in\mathcal{G}}\inf_{\eta_K}\{E[(g(x_i)-P^K(x_i)'\eta_K)^2]\}^{1/2} = 0$ *for all* $\zeta > 0$.

Primitive conditions for an approximation rate for Fourier series follow from Corollary 1 of Edmunds and Moscatelli (1977). If the support of x is a compact, convex subset of $(0,2\pi)^r$, and $g_0(x)$ is d times continuously differentiable, then Assumption 3.1 will be satisfied for $\zeta = (d/r) - \epsilon$ for any $\epsilon > 0$.

In what follows, Assumption 3.1 will be taken to be a primitive condition, rather than one of these other smoothness hypotheses. This feature of the paper will allow the results to apply to different types of series and under more general approximation theorems than those currently available.

Assumption 3.1 delivers the following convergence result for $\hat{g}(x_i)$:

Theorem 3.1: If Assumption 3.1 is satisfied and $E[|w|^q] < \infty$ *for* $q > 2$ *then*

$$(3.6) \qquad \sum_{i=1}^n [\hat{g}(x_i)-g_0(x_i)]^2/n = O_p(n^{(2-\nu)/\nu}K) + O_p(K^{-2\zeta}).$$

Furthermore, if $\mathrm{Var}(w_i|x_i)$ *is bounded then the* $(2-\nu)/\nu = (2/\nu)-1$ *term can be replaced by* -1.

All theorems are proven in the Appendix. The two terms following the equality in (3.6) correspond essentially to variance and bias. In the case where $\mathrm{Var}(w_i|x_i)$ is bounded, choosing $K(n) = n^{1/(2\zeta+1)}$ balances the two terms, yielding the best convergence rate $n^{-2\zeta/(2\zeta+1)}$ for $\sum_{i=1}^n [\hat{g}(x_i)-g(x_i)]^2/n$ that is obtainable from equation (3.6).

The asymptotic distribution results also make use of a convergence rate for $\sum_{i=1}^n u_{in}[\hat{g}(x_i)-g_0(x_i)]/n$, where $E[u_{in}|x_i] = 0$. For brevity, details are reserved to the appendix.

Series estimates depend on the choice of the number of terms K, so that it is desirable to choose K based on the data. With a data-based choice of K, series estimates have the flexibility to adjust to conditions in the data. For example, one might choose K by delete one cross validation, by minimizing the sum of squared residuals $\sum_{i=1}^n [w_i-\hat{g}_{-i}(x_i)]^2$, where $\hat{g}_{-i}(x)$ is the estimate of the regression function computed from all the observations but the i^{th}. The results to follow will allow for K to be data-based in the following way:

Assumption 3.2: $\hat{K} = K(z_1,\ldots,z_n,n)$ satisfies $\hat{K}^{-1} = o_p(n^{-\gamma})$ and $\hat{K} = o_p(n^{\Gamma})$, and $\zeta > 1$ in Assumption 3.1.

The values of γ and Γ will be specified in the results to follow. The $\zeta > 1$ hypothesis is useful for obtaining a convergence rate for the series bias under random choice of K; see Lemma A.5 of the Appendix.

There is reason to think that cross-validated $\hat{K}$ can automatically satisfy the growth rate hypotheses in the results to follow. If the rate of Theorem 3.1 is the best attainable, and cross-validated $\hat{K}$ behaves approximately like the optimal one when $\mathrm{Var}(w_i|x_i)$, is bounded, as is true for kernel nonparametric regression, see Hardle et. al. (1988), then $\hat{K}$ converges to zero at the same rate as $1/n^{2\zeta+1}$, implying Assumption 3.2 is satisfied for

$$(3.7) \qquad \gamma = 1/(2\zeta+1) - \epsilon, \quad \Gamma = 1/(2\zeta+1) + \epsilon,$$

for any $\epsilon > 0$. These values will lie within the bounds imposed in the Theorems of Sections 4 and 5 if ζ is sufficiently large and high enough order moments are bounded. Of course, this discussion is speculative, since $\hat{K}$ has not yet been shown to satisfy Assumption 3.2, and verification of this conjecture is beyond the scope of this paper.

4. Asymptotic Distribution Theory for Regression Functionals

A series regression estimator for the functional in equation (2.1) can constructed as follows. Let P be as defined above for a possibly random $K = \hat{K}$, $\hat{w}_i(\beta) = w(z_i,\beta,\hat{\eta})$, and $\hat{w}(\beta) = (\hat{w}_1(\beta),\ldots,\hat{w}_n(\beta))'$. The estimate $\hat{g}(x_i,\beta)$ considered here takes the form

$$(4.1) \qquad \hat{g}(x_i,\beta) = p^{\hat{K}}(x_i)'(P'P)^{-}P'\hat{w}(\beta).$$

Note that this estimate involves no sample splitting. All the observations, including the i^{th}, are used in the computation of $\hat{g}(x_i,\beta)$, and the same data that appear in $A_n(\beta)$ are used in the estimation of g. The results depend on the absence of sample splitting. The use of an in sample regression estimate allows on to use simple properties like Theorem 3.1, that do not require nonsingularity of $P'P$. Also, if the estimate of the functional depends on different data than $\hat{g}(x,\beta)$, then the limiting distribution of $A_n(\beta_0)$ could be different if $G(x) \neq 0$.

In order to present results it is necessary to introduce some regularity conditions. The following condition imposes smoothness restrictions and dominance conditions on $w(z,\beta,\eta)$. Let g, β, and η subscripts denote partial derivatives with respect to each of these arguments, and for a matrix B let $\|B\| = [\mathrm{trace}(B'B)]^{1/2}$.

Assumption 4.1: There are $q > 2$, $d_w(z)$, $\epsilon > 0$, and neighborhoods $\mathcal{B}$, N of β_0, η_0 respectively such that for β, $\tilde{\beta} \in \mathcal{B}$ and $\eta \in N$, $E[\|w(z,\beta,\eta_0)\|^q] < \infty$, $\|w(z,\beta,\eta)\| \le d_w(z)$, $w(z,\beta,\eta)$ is continuously differentiable in η with $\|w_\eta(z,\beta,\eta)\| \le d_w(z)$, $\|w(z,\tilde{\beta},\eta)-w(z,\beta,\eta)\| \le d_w(z)\|\tilde{\beta}-\beta\|^\epsilon$, $\|w_\eta(z,\tilde{\beta},\tilde{\eta})-w_\eta(z,\beta,\eta)\| \le d_w(z)(\|\tilde{\beta}-\beta\|^\epsilon+\|\tilde{\eta}-\eta\|^\epsilon)$, $E[d_w(z)^2] < \infty$. Also, $\sqrt{n}(\hat{\eta}-\eta_0) = O_p(1)$.

In addition to deriving the asymptotic distribution of $\sqrt{n}[A_n(\beta_0)-\bar{A}_n(\beta_0)]$ it will be useful to have conditions for the uniform convergence result

$$(4.2) \qquad \sup_{\beta\in\mathcal{B}}|A_n(\beta)-\bar{A}_n(\beta)| = o_p(1), \quad \bar{A}_n(\beta) \text{ is equicontinuous.}$$

Such results are used to show consistency for estimators of parameters of interest and for asymptotic variances. Here a pair of uniform convergence results will be given, one that restricts $w(z,\beta,\eta)$ to be independent of β and the other of which relaxes this restriction at the expense of stronger smoothness restrictions on $a^n(z,\beta,g)$. These results both use the following condition. Let s denote the dimension of w.

Assumption 4.2: $\mathcal{B}$ is compact, for each $\beta \in \mathcal{B}$, $E[|a^n(z,\beta,g_0(x,\beta))|^{1+\epsilon}] = O(1)$ for $\epsilon > 0$, and $a^n(z,\beta,g)$ is continuously differentiable in $g \in \mathbb{R}^s$.

The first uniform convergence rate depends on the following hypothesis:

Assumption 4.3: $w(z,\beta,\eta)$ does not depend on β and there exists $d_{jn}(z)$, $j = 1,2,3$, $r > 0$, $\nu' > 2$, r', $\epsilon > 0$, such that for all $\beta \in \mathcal{B}$, $\|a^n_g(z,\beta,g)\| \leq d_{1n}(z) + d_{2n}(z)\|g\|$, $|a^n(z,\tilde{\beta},g_0(x))-a^n(z,\beta,g_0(x))| \leq d_{3n}(z)\|\tilde{\beta}-\beta\|^{\epsilon}$, $(E[d_{1n}(z)^2])^{1/2} = O(n^r)$, $(nE[d_{2n}(z)^{\nu'}])^{1/\nu'} = O(n^{r'})$, $E[d_{3n}(z)] = O(1)$.

The growth rate conditions for moments of dominating functions given here will have an impact on the allowed growth rates for $\hat{K}$. Also, it should be noted that a strong condition is imposed on the magnitude of $a^n_g(z,\beta.g)$ in this Assumption, namely that $\|a^n_g(z,\beta.g)\|$ grows no faster than linearly in $\|g\|$. This Assumption rules out some functional forms, e.g. $a^n(z,\beta,g)$ cannot be a cubic function of g. Nevertheless, it would be possible to modify the results to apply to a larger class of functions by imposing a boundedness constraint on $\hat{g}(x_i,\beta)$ that is relaxed as the sample size grows, which is done below for estimation of asymptotic standard errors. Such a generalization will not be given here, because of the additional notational complexity it would require and because it is not needed for the examples. Similar remarks apply to the other hypotheses of this Section.

The following result is a uniform convergence Theorem.

Theorem 4.1: Suppose that Assumptions 4.1 - 4.3 are satisfied, Assumption 3.1 is satisfied for $g_0(x)$ *equal to each element of* $E[w(z,\eta_0)|x]$, *and Assumption 3.2 is satisfied for* $\max\{r,r'\}/(\zeta-1) \leq \gamma < \Gamma \leq 1 - 2/q - 2\max\{r,r'\}$. *Then equation (4.2) holds.*

It should be noted that the conditions on γ and Γ implicitly impose restrictions on q, r, r', and ζ. Positivity of Γ implies $r, r' < 1/2$, and $q > 2$. For γ to be smaller than Γ, ζ must be large enough relative to r and r'. It is possible to weaken the growth rate conditions under further hypotheses. If $\mathrm{var}(w|x)$ is bounded then Γ can be set equal to its limit as q goes to infinity. If $E[d_{1n}(z)^2]$ is bounded then r can be set equal to 0, and if $d_{2n}(z)$ is bounded then r' can be set equal to zero. If all these conditions hold, and in addition with probability approaching one $\hat{K} \in \mathcal{K}(n)$, where the number of elements of $\mathcal{K}(n)$ is uniformly bounded, and $E[\|g_0(x,\beta)'-P^K(x)'\pi_K\|^2] = o(1)$ for each β rather than the stronger Assumption 3.1, then the conditions on $\hat{K}$ can be weakened to

$$(4.3) \qquad \hat{K} \xrightarrow{p} \infty, \quad \hat{K} = o_p(n).$$

Similar remarks also apply to the other results of this Section.

The following condition allows w to depend on β at the expense of strengthening other conditions.

Assumption 4.4: There exists $d_{jn}(z)$, $j = 1,2$, C, $\epsilon > 0$ such that for each $\beta \in \mathcal{B}$, $\|a_g^n(z,\beta,g)\| \le d_{1n}(z) + C\|g\|$, $|a^n(z,\tilde{\beta},g)-a^n(z,\beta,g)| \le \{d_{2n}(z)+C\|g\|^2\}\|\tilde{\beta}-\beta\|^\epsilon$, $E[d_{1n}(z)^2] = O(1)$ and $E[d_{2n}(z)] = O(1)$.

Theorem 4.2: Suppose that Assumptions 4.1, 4.2, and 4.4 are satisfied, that for every $\beta \in \mathcal{B}$ *Assumption 3.1 is satisfied for* $g_0(x)$ *equal to each element of* $E[w(z,\beta,\eta_0)|x]$, *and Assumption 3.2 is satisfied for* $0 < \gamma < \Gamma \le 1 - 2/q$. *Then equation (4.2) holds.*

The following condition is helpful in deriving an asymptotic distribution result for $\sqrt{n}[A_n(\beta_0)-\bar{A}_n(\beta_0)]$.

Assumption 4.5: $E[|a^n(z,\beta_0,g_0(x,\beta_0))|^{2+\epsilon}] = O(1)$ for $\epsilon > 0$, and there exists $d_{1n}(z)$, $d_{2n}(z)$, $\nu > 2$, $\nu' > 1$, r, $r' > 0$ such that for $\beta \in \mathcal{B}$, $a^n(z,\beta,g)$ is continuously differentiable in $g \in \mathbb{R}^s$, $\|a^n_g(z,\beta,g_0(x,\beta))\| \le d_{1n}(z)$, $\|a^n_g(z,\beta,\tilde{g})-a^n_g(z,\beta,g)\| \le d_{2n}(z)\|\tilde{g}-g\|$ for $\tilde{g} \in \mathbb{R}^s$, $(E[nd_{1n}(z)^{\nu}])^{1/\nu} = O(n^r)$, and $(nE[d_{2n}(z)^{\nu'}])^{1/\nu'} = O(n^{r'})$.

When $G(x)$ in equation (2.4) is not equal to zero, an additional condition is important for obtaining an asymptotic distribution result for $\sqrt{n}[A_n(\beta_0)-\bar{A}_n(\beta_0)]$. Let $a^n_g = a^n_g(z,g_0(x))$, where for notational convenience dependence on β_0 is suppressed here (e.g. $g_0(x) = g_0(x,\beta_0)$) and in what follows, and let $G^n(x) = E[a^n_g|x]$.

Assumption 4.6: $E[\|G^n(x)\|^{\nu_G}] = O(1)$ for $\nu_G > 2q/(q-2)$ and q from Assumption 4.1. There exists $\xi > 1$ and $\tilde{\pi}_K$ such that $(E[\|G^n(x)'-P^K(x)'\tilde{\pi}_K\|^2])^{1/2} = O(K^{-\xi})$ uniformly in n.

Note that this condition is not restrictive if $G^n(x) = 0$ for all n large enough. The discussion of primitive conditions for the approximation rate for $g_0(x)$ also apply to the approximation rate for $G^n(x)$. Note in particular that the Taylor theorem approximation result is uniform in a class of functions, implicitly allowing dependence on n.

The following Theorem gives an asymptotic representation result for $\sqrt{n}(A_n-\bar{A}_n)$. Let $g_i = g_0(x_i)$, $\hat{g}_i = \hat{g}(x_i)$, and $w_i = w(z_i,\eta_0)$.

Theorem 4.3: Suppose that Assumptions 4.1, 4.5, and 4.6 are satisfied, that Assumption 3.1 is satisfied for each element of $E[w_i|x]$, *and Assumption 3.2 is satisfied with* $\max\{4r,2r'+1\}/4(\zeta-1) = \gamma < \Gamma = 1/2 - \max\{1/q + r, 2/q + r'\}$. *Also suppose that either* $G^n(x) = 0$ *for large enough* n *or* $\gamma \geq \max\{[(\Gamma/2)+(1/q)]/(\xi-1),\ 1/[2(\xi+\zeta-2)]\}$. *Then*

$$(4.4)\qquad \sqrt{n}[A_n-\bar{A}_n] = \Sigma_{i=1}^{n}\{a^n(z_i,g_i)-E[a^n]\}/\sqrt{n} + \Sigma_{i=1}^{n}G^n(x_i)'(w_i-g_i)/\sqrt{n}$$

$$+ E[G^n(x)'w_\eta(z,\eta_0)]\sqrt{n}(\hat{\eta}-\eta_0) + o_p(1).$$

Similarly to Theorems 4.1 and 4.2, the conditions on γ and Γ implicitly impose restrictions on q, r, r', ζ, and ξ. If $d_{2n}(z)$ is bounded, then r' can be taken equal to zero in the conditions for γ and Γ. If the conditions are changed as discussed following Theorem 4.1, except that Assumption 3.1 is maintained, and $\mathrm{Var}(a_g^n|x)$ is bounded, then the rate conditions for $\hat{K}$ can be weakened to

$$(4.5)\qquad \Gamma = 1/2,\quad \gamma = \max\{1/4\zeta, 1/4\xi, 1/2(\zeta+\xi)\},\quad \zeta,\ \xi > 1/2.$$

This result verifies the conjecture for the limiting distribution of A_n given in the last section for the case where $\hat{\eta}$ is not present and $a^n(z,\beta,g)$ does not depend on n. When $G^n(x) \neq 0$ and $\hat{\eta}$ is present, equation (4.3) contains an additional term that is a correction factor for the estimation of η.

The last result of this Section concerns consistent estimation of the asymptotic variance of A_n, which may be required for asymptotic inference procedures. It is necessary to give a separate result because the asymptotic variance is more complicated than $\bar{A}_n$, so that consistency of the estimated variance cannot be shown by applying Theorem 4.1 or 4.2. Also, it is appropriate to here let $a^n(z,\beta,g)$ be a vector, so that the Theorem concerns consistency of an estimate of the asymptotic variance matrix.

In general it is necessary to be more specific about the properties of $\hat{\eta}$ in order to estimate the asymptotic variance A_n. The following assumption is useful in this respect:

Assumption 4.7: There exists $\psi_\eta(z)$ such that $E[\|\psi_\eta(z)\|^{2+\epsilon}] < \infty$ for $\epsilon > 0$, $E[\psi_\eta(z)] = 0$, and $\sqrt{n}(\hat{\eta}-\eta_0) = \sum_{i=1}^n \psi_\eta(z_i)/\sqrt{n} + o_p(1)$. Also, there exists $\hat{\psi}_{\eta i}$ such that $\sum_{i=1}^n \|\hat{\psi}_{\eta i}-\psi_\eta(z_i)\|^2/n = o_p(1)$.

This Assumption specifies $\psi_\eta(z)$ to be the influence function for η, and $\hat{\psi}_{\eta i}$ to be a consistent estimator of this influence function. The consistency condition for $\hat{\psi}_{\eta i}$ means that $\sum_{i=1}^n \hat{\psi}_{\eta i}\hat{\psi}'_{\eta i}/n$ will be a consistent estimator of the asymptotic variance of $\hat{\eta}$, and will also be useful for showing consistency of the asymptotic variance estimate for A_n.

Assumption 4.7 and Theorem 4.3 lead to a formula for the influence function of A_n. If H_n is bounded then it follows that

$$(4.6) \qquad \sqrt{n}(A_n-\bar{A}_n) = \sum_{i=1}^n \psi^n(z_i)/\sqrt{n} + o_p(1),$$

$$\psi^n(z) = a^n(z,g_0(x)) - E[a^n] + G^n(x)[w-g_0(x)] + H_n\psi_\eta(z).$$

The asymptotic variance of A_n will be $\Sigma_n = E[\psi^n(z)\psi^n(z)']$, which can be estimated by the sample second moment of an estimator of the influence function.

A trimmed estimator of $g_0(x_i)$ is useful for constructing a consistent estimator of this influence function. For a scalar g

$$\tau^\Delta(g) = \begin{cases} \Delta, & g > \Delta \\ g, & |g| \le \Delta \\ -\Delta, & g < -\Delta \end{cases} .$$

Also, for a vector g, let $\tau^\Delta(g)$ be the vector of functions of corresponding components (e.g $\tau^\Delta(g_1,g_2) = (\tau^\Delta(g_1),\tau^\Delta(g_2))$). In some places in the influence function the estimator of $g_0(x_i)$, will be taken to be $\hat{g}_i^{\Delta} \equiv \tau^{\Delta}(\hat{g}_i)$, where $\hat{g}_i$ is calculated from equation (4.1) with $w(z,\hat{\beta},\hat{\eta})$ replacing $w(z,\hat{\eta})$ and $\hat{\beta}$ an estimate of β_0. Note that the influence function implicitly depends on β_0, so that the presence of $\hat{\beta}$ is essential.

The estimate $\hat{\psi}_i^n$ of $\psi^n(z_i)$ is constructed as follows. Let

$$(\hat{G}_i^n)_{\ell} = P^K(x_i)'(P'P)^{-}\textstyle\sum_{j=1}^n P^K(x_j)(a_g(z,\hat{\beta},\hat{g}_j))_{\ell},$$

$$\hat{H}_n = \textstyle\sum_{i=1}^n \hat{G}_i^n w_{\eta}(z_i,\hat{\beta})/n, \quad \hat{a}_i^n = a^n(z_i,\hat{\beta},\hat{g}_i^{\Delta}), \quad \hat{w}_i = w(z_i,\hat{\beta},\hat{\eta}),$$

where $(\circ)_{\ell}$ denotes the ℓ^{th} row of a matrix. Then for $\hat{\psi}_{\eta i}$ from Assumption 4.7, the estimator of $\psi^n(z_i)$ and corresponding estimator of Σ_n are

$$(4.7) \qquad \hat{\psi}_i^n = \hat{a}_i^n - \textstyle\sum_{j=1}^n \hat{a}_j^n/n + \hat{G}_i^n(\hat{w}_i - \hat{g}_i^{\Delta}) + \hat{H}_n\hat{\psi}_{\eta i}. \quad \hat{\Sigma}_n = \textstyle\sum_{i=1}^n \hat{\psi}_i^n \hat{\psi}_i^{n\prime}/n.$$

In cases where $G^n(x)$ is known to be zero, the estimate takes the same form with $\hat{G}_i^n = 0$ and $\hat{H}_n = 0$.

The following pair of regularity conditions are useful for showing consistency of $\hat{\Sigma}_n$. Recall that s is the dimension of $g_0(x)$.

Assumption 4.8: There exists $d_{3n}(z)$, $d_{4n}(z)$, $\nu_3 > 2$, $r_3 > 0$, such that for $\beta \in \mathcal{B}$, $\sup_{\|g\| \le s\Delta} \|a_g^n(z,\beta,g)\| \le d_{3n}(z)$, $\sup_{\|g\| \le \|g_0(x)\|} \|a_g^n(z,\beta,g)\| \le d_{4n}(z)$, $(E[nd_{3n}(z)^{\nu_3}])^{1/\nu_3} = O(n^{r_3})$, $E[d_{4n}(z)^{2q/(q-2)}] = O(1)$ for q from Assumption 4.1.

Assumption 4.9: $\sqrt{n}(\hat{\beta}-\beta_0) = O_p(1)$ and there exist $\epsilon > 0$, $d_{\beta n}(z)$, such that for β, $\tilde{\beta} \in \mathcal{B}$, $\|a^n(z,\tilde{\beta},g_0(x)) - a^n(z,\beta,g_0(x))\| \le d_{\beta n}(z)\|\tilde{\beta}-\beta\|^{\epsilon}$, $\|a_g^n(z,\tilde{\beta},g_0(x)) - a_g^n(z,\beta,g_0(x))\| \le d_{\beta n}(z)\|\tilde{\beta}-\beta\|^{\epsilon}$, $E[d_{\beta n}(z)^2] = O(1)$.

Theorem 4.4: Suppose that $\Delta = \Delta(n) \longrightarrow \infty$, *Assumptions 4.1, 4.8, and 4.9 are satisfied, Assumption 3.1 is satisfied for each element of* $E[w(z,\beta_0,\eta_0)|x]$, *Assumption 3.2 is satisfied with* $r_3/(\zeta-1) \le \gamma < \Gamma \le 1-2r_3-2/q$ *and* $G^n(x) = 0$. *Then*

$$\sum_{i=1}^{n}\hat{\psi}_i^n\hat{\psi}_i^{n\prime}/n - E[\psi_i^n\psi_i^{n\prime}] = o_p(1). \tag{4.8}$$

If $G^n(x) \neq 0$ *but in addition, for some* $1/2 > t > 0$, $\Delta(n) = O(n^{t})$, $E[d_w(z)^{1/t}] < \infty$, *Assumptions 4.5 - 4.7 are satisfied, and* $\max\{(t+r')/(\zeta-1), t/(\xi-1), 1/\nu_G(\zeta-1)\} \le \gamma < \Gamma \le 1-2\max\{t+r'+1/q, t+r, 1/\nu_G+1/q\}$, *then equation (4.8) holds. Furthermore, if the smallest eigenvalue of* $E[\psi_i^n\psi_i^{n\prime}]$ *is bounded away from zero, then*

$$(\sum_{i=1}^{n}\hat{\psi}_i^n\hat{\psi}_i^{n\prime}/n)^{-1/2}\sqrt{n}[A_n(\beta_0)-\bar{A}_n(\beta_0)] \xrightarrow{d} N(0,I).$$

5. Examples

In this Section the results of Section 4 will be used to develop asymptotic distribution theory for a number of examples. The examples are estimating the residual variance from a nonparametric regression, estimation of microeconometric expectations models, correcting for heteroskedasticity of unknown form, best nonlinear two stage least squares, and estimation of additive semiparametric regression models.

5.1 Estimating the Residual Variance

In some circumstances the residual variance

$$(5.1.1)\qquad \sigma^2 = E[\{w - g_0(x)\}^2]$$

may be of interest, where w is a scalar. It can be useful in evaluating the fit of a nonparametric regression. A pair of residual variances could be used to compare the fit of two nonparametric regressions. For inference purposes, it is important to have an asymptotically normal estimator of the residual variance, and a consistent estimator of its asymptotic variance.

A series estimator of the residual variance from a nonparametric regression of w on x is given by

$$(5.1.2)\qquad \hat{\sigma}^2 = \sum_{i=1}^{n}(w_i-\hat{g}_i)^2/(n-\hat{K}),$$

where $\hat{g}_i$ is calculated as in equation (4.1). A consistent estimator of the asymptotic variance of $\hat{\sigma}^2$ can be formed by using the trimmed estimator of g_i discussed in Section 4. Let $\varepsilon_i = w_i-g_i$, $\hat{\varepsilon}_i = w_i-\tau^{\Delta}(\hat{g}_i)$, $\Omega = \mathrm{Var}(\varepsilon^2) = E[\varepsilon^4] - (E[\varepsilon^2])^2$, and

$$(5.1.3)\qquad \hat{\Omega} = \sum_{i=1}^{n}\hat{\varepsilon}_i^4/n - (\sum_{i=1}^{n}\hat{\varepsilon}_i^2/n)^2.$$

Theorem 5.1: Suppose that there exists $q > 4$, $0 < t < 1/2$, *such that* $E[|w|^q] < \infty$, $\Delta(n) \longrightarrow \infty$, *and* $\Delta(n) = O_p(n^t)$. *Also suppose that Assumptions 3.1 and 3.2 are satisfied, with* $0 < \max\{1/q + t,\ 1/4\}/(\zeta-1) \le \gamma < \Gamma \le \min\{1/2 - 2/q,\ 1 - 2t - 4/q\}$. *Then*

$$(5.1.4)\qquad \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0,\ \Omega),\quad \hat{\Omega} \xrightarrow{p} \Omega.$$

It is interesting to note that estimation of $g_0(x)$ does not affect the asymptotic distribution of $\hat{\sigma}^2$, which has the same asymptotic distribution as $\sum_{i=1}^{n}\varepsilon_i^2/n$. This result is not surprising, being familiar from parametric models, and is expected because of the form of the efficient influence function in equation (2.4); note that $G(x) = E[a_g(z,g_0(x))|x] = E[-2(w-g_0(x))|x] = 0$. Indeed, it has been shown by Yatchew (1988) that this result holds quite generally.

It might also be of interest to consider other functionals of nonparametric residuals. For example, a sample cross product $\sum_{i=1}^{n}(w_i-\hat{g}_i)B(X_i)/n$ might be used to test whether $E[w|x] = E[w|X]$. The asymptotic distribution of such functionals will be more complicated than that of $\hat{\sigma}^2$, generally involving the correction term in equation (2.4) for the estimation of $g_0(x)$. Nevertheless, Theorem 4.4 could be used in the construction of a consistent estimator of the asymmtotic variance of such functionals. For brevity, details are omitted.

5.2 Estimation of A Microeconometric Expectations Model

Consider the model

$$(5.2.1)\quad E[\rho(z,g_0(x),\beta_0)|X] = 0,$$

where $\rho(z,g,\beta)$ is residual that depends on $g \in \mathbb{R}^s$ and parameters β, $g_0(x) = E[w|x]$ for vectors w and x of observed variables, and X is a vector of exogenous variables that includes x. This model can be motivated by economic models of individual decision making, where $g_0(x)$ is an expectation of an uncertain outcome on which the decision is based. For example, Manski's (1988) dynamic discrete choice models are more complicated versions of one with

$$(5.2.2) \quad \rho(z,g,\beta) = y - \Phi(f(X)'\beta_1 + g_0(x)'\beta_2),$$

where $y \in \{0,1\}$ is a decision variable and $\Phi(\circ)$ is the standard normal cumulative distribution function. This model is a probit model with conditional expectations as regressors. Another example is a microeconomic risk model of the form,

$$(5.2.3) \quad \rho(z,g,\beta) = y - f(X)'\beta_1 - (g_{10}(x), [g_{20}(x)-\{g_{10}(x)\}^2])\beta_2,$$

where $g_{j0}(x) = E[w^j|x]$, $(j=1,2)$. This is a model with a conditional variance regressor. For example, y might be an indicator for fixed or variable rate telephone charges, w the amount of phone use, and equation (5.2.3) a linear probability model for choice of rate type for a risk averse individual.

These models are microeconometric because of the hypothesis that the data observations are independent. This restriction means that each z_i is best interpreted as an observation for an individual economic unit, and not as an observation for a single time period. Of course, panel data is allowed, where z_i includes data observed at different time periods. For stationary panel data, $g_0(x)$ could be interpreted as an expectation given "lagged" variables x. It should be noted that the expectation $g_0(x)$ is stationary across individuals, a strong assumption that may not be satisfied. See Manski (1988) for further discussion.

Functional form misspecification of $g_0(x)$ could cause inconsistency of estimators of β. Also, when $\rho(z,g,\beta)$ is nonlinear in g, as in the above examples, it is not possible to substitute w for $g_0(x)$ and estimate by instrumental variables, i.e. to use "errors-in-variables" methods. Because of these features, use of a nonparametric estimator of $g_0(x)$ is potentially important. At the same time, nonparametric estimation of $g_0(x)$ will require that the econometrician observe all the variables in x, which is not required by errors in variables methods.

The type of estimator considered here is two-step instrumental variables (IV) estimator, where the first step consists of nonparametric estimation of $g_0(x)$. Let $B(X)$ denote a vector of functions of X, let $\hat{g}_i$ be a series estimator of the elements of $g_0(x_i)$ as in equation (4.1), and define

$$m_n(\beta) = \sum_{i=1}^{n} B(X_i)\rho(z_i,\hat{g}_i,\beta)/n.$$

A nonlinear IV estimator of β_0 can be obtained as the solution to

$$(5.2.4) \qquad \hat{\beta} = \operatorname{argmin}_{\beta\in\mathcal{B}} S_n(\beta), \quad S_n(\beta) = m_n(\beta)'\hat{\Psi}m_n(\beta),$$

where $\hat{\Psi}$ is a symmetric, positive semi-definite matrix.

For regression models such as those of equations (5.2.2) and (5.2.3), two-step nonlinear least squares or maximum likelihood estimation is also possible. The paper focuses on IV, even for these cases, for the technical reason that less stringent smoothness conditions for $\rho(z,g,\beta)$ as a function of g are required for IV than for other methods. It is possible to weaken these hypotheses if a trimmed estimator of $g_0(x)$ is used throughout, which is not done for reasons discussed following Assumption 4.3.

Estimation of the asymptotic variance of $\hat{\beta}$ and of an optimal $\hat{\Psi}$ will require an estimator of the asymptotic variance of $\sqrt{n}m_n(\beta_0)$. In general, estimation of $g_0(x)$ will have to be accounted for, which can be done by applying Theorems 4.3 and 4.4. Note that for this model, $a(z,\beta,g) = B(X)\rho(z,g,\beta)$, so that $G(x) = E[B(X)\rho_g(z,g_0(x),\beta_0)|x]$. Since η is not present, it will follow by the conclusion of Theorem 4.3 that the influence function for $\sqrt{n}m_n(\beta_0)$ is $\psi(z) = B(X)\rho(z,g_0(x),\beta_0) + G(x)[w-g_0(x)]$. Let $\hat{G}_i$ be the estimate of $G(x_i)$ constructed as the predicted values from a regression of $B(X_i)\rho_g(z,\hat{g}_i,\hat{\beta})'$ on P, and let $\hat{g}_i^{\Delta}$ be the trimmed estimator described in Section 4. Then the estimate of the influence function and asymptotic variance $\Sigma = E[\psi\psi']$ of $\sqrt{n}m_n(\beta_0)$ described in equation (4.7) is

$$(5.2.5) \qquad \hat{\psi}_i = B(X_i)\rho(z,\hat{g}_i^{\Delta},\hat{\beta}) + \hat{G}_i(w_i-g_i^{\Delta}), \quad \hat{\Sigma} = \sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i'/n.$$

By the usual method of moments calculation (Hansen, 1982), the asymptotic variance of $\hat{\beta}$ will be $\Omega = (M'\Psi M)^{-1}M'\Psi\Sigma\Psi M(M'\Psi M)^{-1}$, where $M = E[B(X)\rho_\beta(z,g_0(x),\beta_0)] = \text{plim}(\partial m_n(\beta_0)/\partial\beta)$ and $\Psi = \text{plim}(\hat{\Psi})$, which can be estimated by $\hat{\Omega} = (\hat{M}'\hat{\Psi}\hat{M})^{-1}\hat{M}'\hat{\Psi}\hat{\Sigma}\hat{\Psi}\hat{M}(\hat{M}'\hat{\Psi}\hat{M})^{-1}$, where $\hat{M} = \partial m_n(\hat{\beta})/\partial\beta$.

The following result gives the regularity conditions that are sufficient for Theorems 4.3 and 4.4 to apply to this problem.

Theorem 5.2: Suppose that i) β_0 *is an element of the interior of* $\mathcal{B}$, *which is compact and convex;* ii) $\hat{\Psi} = \Psi + o_p(1)$, Ψ *is nonsingular,* and $E[B(X)\rho(z,g_0(x),\beta)] \neq 0$ *for* $\beta \in \mathcal{B}$, $\beta \neq \beta_0$; iii) $E[B(x)\rho_\beta(z,g_0(x),\beta_0)]$ *has full column rank and* Σ *is nonsingular;* iv) $\rho(z,g,\beta)$ *is twice continuously differentiable in* g and β; v) $E[\|w\|^q] < \infty$ *for* $q > 4$ *and there exists* $d_\beta(z)$, $d_{gj}(z)$, $\nu_j \geq 2$, $j=1,2,3$, $\nu_B > 2$, $\nu_\beta > 2\nu_B/(\nu_B-2)$, $\nu_1 \geq 2\nu_B/(\nu_B-2)$, $(1/\nu_B + \max\{1/\nu_1, 1/\nu_2 + 1/q\})^{-1} \geq 2q/(q-2)$, *such that for all* $\beta \in \mathcal{B}$, $\|\rho(z,g_0(x),\beta)\|$, $\|\rho_\beta(z,g_0(x),\beta)\|$, $\|\rho_{\beta\beta}(z,g_0(x),\beta)\| \leq d_\beta(z)$, $\|\rho_g(z,g,\beta)\|$, $\|\rho_{\beta g}(z,g,\beta)\| \leq d_{g1}(z) + d_{g2}(z)\|g\|$, $\|\rho_{gg}(z,g,\beta)\| \leq d_{g3}(z)$, $E[\|B(X)\|^{\nu_B}]$, $E[d_{gj}(z)^{\nu_j}] < \infty$, $j = 1,2,3$; vi) *Assumption 3.1 is satisfied for each element of* $E[w|x]$ *and Assumption 3.6 for* $G(x) = E[B(X)\rho_g(z,g_0(x),\beta_0)'|x]$; vii) $\Delta(n) \longrightarrow \infty$, $\Delta(n) = O(n^t)$, *and Assumption 3.2 is satisfied with* $0 < (1/q + 1/\nu_B + t + \max_{j\leq3}\{1/\nu_j\} + 1/4)/(\zeta-1) \leq \gamma < \Gamma \leq 1/2 - 2/q - 1/\nu_B - t - \max_{j\leq3}\{1/\nu_j\}$, $0 \leq \max\{t/(\xi-1), (\Gamma/2 + 1/q)/(\xi-1), 1/2(\zeta+\xi-2)\}$; *Then*

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

For the above examples, it is easy to give conditions for the dominance hypotheses of this result. For equation (5.2.2), it suffices to take $d_\beta(z) = C(1 + \|f(X)\|^2 + \|g_0(x)\|^2)$, $d_{1n}(z) = C\|f(X)\|$, $d_{2n}(z) = d_{3n}(z) = C$, for some constant C, so that ν_2 and ν_3 can be set as large as desired in the gamma rate conditions. For equation (5.2.3), it suffices to take $d_\beta(z) = C(1 + \|f(X)\| + \|g_0(x)\|^2)$, $d_{1n}(z) = d_{2n}(z) = d_{3n}(z) = C$, so that the ν_j, $j = 1,2,3$ conditions can be ignored in the hypotheses of this result.

Concerning efficiency of $\hat{\beta}$, it follows by Theorem 3.2 of Hansen (1982) that the choice of $\hat{\Psi}$ that minimizes the asymptotic variance of $\hat{\beta}$ is $\hat{\Sigma}^{-1}$. An analogous minimum chi-square two-stage estimator was discussed by Hansen (1985). Little more is known about the efficiency properties of $\hat{\beta}$, although one could derive the semiparametric efficiency bound for this model. Also, it is plausible that the bound should be approximately attained if a sufficient number and variety of functions of X are used in forming the instruments, although verification of this conjecture is beyond the scope of this paper.

5.3 A Series Correction for Heteroskedasticity of Unknown Form

Consider the linear regression model

$$(5.3.1) \qquad y = x'\beta_0 + \varepsilon, \quad E[\varepsilon|x] = 0, \quad E[\varepsilon^2|x] = \sigma^2(x).$$

An asymptotically efficient, linear (in y) estimator of β is the HGLS estimator of β_0, which is weighted least squares with weights $1/\sigma^2(x)$. Carroll (1982) and Robinson (1987) have considered HGLS with $\sigma^2(x)$ replaced by kernel and nearest neighbor estimators respectively. Here, a HGLS estimator with a series estimator of $\sigma^2(x)$ will be given and asymptotic efficiency (in the GLS sense) shown.

To guarantee scale equivariance of HGLS it will be helpful to consider estimation of $g_0(x) = \sigma^2(x)/\phi_0$, where $\phi_0 = E[\varepsilon^2]$. A series estimate of

$g_0(x_i)$ can be constructed as follows. Let $\eta = (\beta', \phi)'$, $w(z,\eta) = (y-x'\beta)^2/\phi$, and let $\hat{\eta}$ be the ordinary least squares (OLS) estimator of η_0; i.e. $\hat{\beta}$ is OLS and $\hat{\phi} = \sum_{i=1}^n (y_i - x_i'\hat{\beta})^2/(n-r)$, where r is the dimension of x. Estimates $\hat{g}_i$ are then obtained from equation (4.1). Note that because $w(z,\eta)$ is linear in $1/\phi$, the resulting estimate is the same as a series estimate of $\sigma^2(x)$ from a regression of the squared residuals on the series terms, divided by $\hat{\phi}$. If a power series is used, then leading linear and quadratic terms would correspond to White's (1980) squared residual regression test for equality of the usual and heteroskedasticity consistent OLS variance estimators.

Because the estimates need not be positive, and for technical reasons involving the nonlinearity of the function $1/\circ$, the asymptotic theory here requires that $\hat{g}_i$ be modified to be positive. Let $\tau^\delta(\circ)$ be a continuously differentiable function with derivative that is Lipschitz uniformly in δ, such that $\tau^\delta(\circ) \geq \delta$ and $\tau^\delta(g) = g$ for $g \geq 2\delta$. An example is

$$(5.3.2) \qquad \tau^\delta(g) = \begin{cases} \delta, & g \leq 0 \\ \delta[1 + (g/2\delta)^2], & 0 < g \leq 2\delta. \\ g, & g > 2\delta \end{cases}$$

An estimate of $g_0(x_i) = \sigma^2(x_i)/\phi_0$ which is constrained to be positive is given by $\hat{g}_i^\delta \equiv \tau^\delta(\hat{g}_i)$.

The weighted least squares least squares estimator with weights $1/\hat{g}_i^\delta$ is

$$(5.3.3) \qquad \hat{\beta} = (\textstyle\sum_{i=1}^n x_i x_i'/\hat{g}_i^\delta)^{-1} \sum_{i=1}^n x_i y_i/\hat{g}_i^\delta.$$

Asymptotic efficiency of $\hat{\beta}$ will require letting δ approach zero as the sample size grows. Note that in the limit as $\delta \longrightarrow 0$ the presence of $\hat{\phi}$ will not affect the estimator; its sole purpose is to make $\hat{\beta}$ scale equivariant.

Under the following conditions, $\hat{\beta}$ is asymptotically efficient.

Theorem 5.3: Suppose that i) $E[xx']$ *and* $E[xx'/\sigma^2(x)]$ *are finite and nonsingular;* ii) $E[\{\|x\|^2/\sigma^2(x)\}^{1+\epsilon}]$, $E[\|x\|^{p}]$, $E[|\varepsilon|^{q}] < \infty$, for $\epsilon > 0$, p, $q > 4$; iii) *Assumption 4.2 is satisfied for* $g_0(x) = \sigma^2(x)$; iv) $\delta = \delta(n) = o(1)$ *such that* $\delta(n)^{-1} = O(n^{t})$ *for* $t > 0$; v) *Assumption 4.3 is satisfied for* $(2/p + 1/q + 2t + 1/4)/(\zeta-1) \le \gamma < \Gamma = 1/2 - 5/q - 2/p - 3t$. *Then for* $\Omega = (E[xx'/\sigma^2(x)])^{-1}$,

$$(5.3.4) \qquad \sqrt{n}(\hat\beta-\beta_0) \xrightarrow{d} N(0,\Omega), \quad \hat\phi(\textstyle\sum_{i=1}^{n} x_i x_i'/\hat{g}_i^{\delta} n)^{-1} \xrightarrow{p} \Omega.$$

The conclusion of this theorem also gives a consistent estimator of the asymptotic variance of $\hat\beta$. The hypotheses of this result are stronger in some respects than those of Carroll (1982) and Robinson (1987), but are weaker in one way. The moment conditions imposed by $\Gamma > 0$ are more severe than they impose, requiring $q > 10$. The conditions can be relaxed if $\mathrm{Var}(\varepsilon^2|x)$ is bounded, in which case $\Gamma \le 1/2 - 1/q - 2/p - 3t$ will suffice. Also, primitive conditions for iii) lead to smoothness restrictions on $\sigma^2(x)$, which are not required by Robinson's (1987) result. The hypothesis that is weaker is that $\sigma^2(x)$ does not have to be bounded away from zero.

5.4 Best Nonlinear Two Stage Least Squares

Consider a model of the form

$$(5.4.1) \qquad E[\rho(z,\beta_0)|x] = 0, \quad E[\rho(z,\beta_0)^2|x] = \sigma^2,$$

where $\rho(z,\beta)$ is a residual and β is a vector of parameters of interest. There are many important examples of this type of model in econometrics and statistics. Nonlinear instrumental variables (IV) estimation of β was considered by Kelejian (1971) and Amemiya (1974). Amemiya showed that the best choice of instruments, in terms of minimizing the asymptotic covariance matrix of an IV estimator, are $D(x) = E[\rho_\beta(z,\beta_0)|x]$. Newey (1989a) gave one-step best IV estimators that use nearest neighbor and series estimates of $D(x)$. Here a corresponding fully iterative efficient estimator is presented and its asymptotic efficiency proven.

One way an efficient estimator might be formed is to use as instrumental variables a set of nonlinear functions of x that becomes richer as the sample size grows. Such a set of instrumental variables is given by $P^K(x)$. Let $\rho(\beta) = (\rho(z_1,\beta),\ldots,\rho(z_n,\beta))'$. The nonlinear two stage least squares estimator with $P^K(x)$ as instrumental variables is

$$(5.4.2) \qquad \hat{\beta} = \text{argmin}_{\beta\in\mathcal{B}}S_n(\beta), \quad S_n(\beta) = \rho(\beta)'P(P'P)^-P'\rho(\beta) = \textstyle\sum_{i=1}^n\hat{g}(x_i,\beta)^2,$$

where $\hat{g}(x_i,\beta)$ is the series estimator of the conditional expectation of $\rho(z,\beta)$ given x_i, and the last equality follows by $P(P'P)^-P'$ idempotent. The previous results can be used to specify a growth rate for K such that this estimator is a best nonlinear instrumental variables estimator.

Theorem 5.4: Suppose that i) β_0 *is an element of the interior of* $\mathcal{B}$, *which is compact and convex;* ii) $E[\rho(z,\beta)|x] \neq 0$ *for* $\beta \in \mathcal{B}$, $\beta \neq \beta_0$; iii) $E[D(x)D(x)']$ *is nonsingular;* iv) $\rho(z,\beta)$ *is twice continuously differentiable and there exists* $q > 2$, $\epsilon > 0$, $d_1^\rho(z)$, $d_2^\rho(z)$ *such that for* β, $\tilde{\beta} \in \mathcal{B}$, $|\rho(z,\beta)| \le d_1^\rho(z)$, $\|\rho_\beta(z,\beta)\| \le d_1^\rho(z)$, $\|\rho_{\beta\beta}(z,\beta)\| \le d_2^\rho(z)$, $\|\rho_{\beta\beta}(z,\tilde{\beta})-\rho_{\beta\beta}(z,\beta)\| \le d_2^\rho(z)\|\tilde{\beta}-\beta\|^\epsilon$, $E[d_1^\rho(z)^q] < \infty$, $E[d_2^\rho(z)^2] < \infty$; v) *For each* $\beta \in \mathcal{B}$, *Assumption 3.1 is satisfied for* $g_0(x)$ *equal to each element of* $g(x,\beta) = E[(\rho(z,\beta),\rho_\beta(z,\beta)')'|x]$; vi) *Assumption 3.2 is satisfied with* $0 < \gamma < \Gamma = (q-2)/2q$. *Then for* $\hat{\sigma}^2 = \sum_{i=1}^n\rho(z_i,\hat{\beta})^2/n$ *and* $\Omega = \sigma^2(E[D(x)D(x)'])^{-1}$,

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,\Omega), \quad \hat{\sigma}^2[\rho_\beta(\hat{\beta})'P(P'P)^-P'\rho_\beta(\hat{\beta})/n]^{-1} \xrightarrow{p} \Omega.$$

The conditions could be weakened somewhat at the expense of additional notational complexity, e.g. by dropping convexity of $\mathcal{B}$ and allowing a number of the dominance conditions to hold only in a neighborhood of β_0. The regularity conditions are similar to those of Newey (1989a), although the identification condition here is weaker; it is the minimal restriction that the conditional expectation of the residual equals zero only at the truth. It is also straightforward to obtain an analogous result for the case where $\rho(z,\beta)$ is a vector, i.e. for systems of nonlinear equations. To avoid additional notational complexity this result is not considered here.

5.5 Additive Semiparametric Regression

Consider a model of the form

$$(5.5.1)\qquad y = f(X,\beta_0) + h_0(x) + \varepsilon, \quad E[\varepsilon|X] = 0, \quad E[\varepsilon^2|X] = \sigma^2,$$

where $f(X,\beta)$ is a known function of exogenous variables X and Euclidean parameters β and $h_0(x)$ is an unknown function of a subvector x of X. This is a nonlinear version of the partially linear model of Engle, Granger, Rice, and Weiss (1984), where $f(X,\beta)$ is linear in β. Estimators for the partially linear model have been considered by N. Heckman (1986), Rice (1986), Schick (1986), Robinson (1988), Chamberlain (1986), and Andrews (1988).

One way an estimator of β might be formed is by a nonlinear least squares regression of y on $f(X,\beta)$ and a set of nonlinear functions of x that becomes richer as the sample size grows. The idea is that the rich set of functions should provide a nonparametric correction for the presence of $h_0(x)$. This estimator is analogous to Chamberlain's (1986) estimator for the partially linear model. Such a rich set of functions is given by $P^K(x)$. Let $w(\beta) = (y_1-f(X_1,\beta),\ldots,y_n-f(X_n,\beta))'$. The nonlinear least squares estimator of β for a regression of y on $f(X,\beta)$ and $P^K(x)$ can be obtained as

(5.5.2) $\hat{\beta} = \operatorname{argmin}_{\beta\in\mathcal{B}} S_n(\beta)$, $S_n(\beta) = w(\beta)'[I-P(P'P)^{-}P']w(\beta)$

$$= \sum_{i=1}^{n}[w(z_i,\beta)-\hat{g}(x_i,\beta)]^2,$$

where $\hat{g}(x_i,\beta)$ is the series estimator of the conditional expectation of $y-f(X,\beta)$ given x_i, and the last equality follows by $P(P'P)^{-}P'$ idempotent. The objective function $S_n(\beta)$ is the sum of squared residuals where the coefficients of $P^K(x)$ have been concentrated out. The previous results can be used to specify a growth rate for K such that this estimator is $\sqrt{n}$-consistent and asymptotically normal.

Theorem 5.5: Suppose that i) β_0 *is an element of the interior of* $\mathcal{B}$, *which is compact and convex;* ii) $f(X,\beta)-E[f(X,\beta)|x] \neq f(X,\beta_0)-E[f(X,\beta_0)|x]$ *for* $\beta \in \mathcal{B}$, $\beta \neq \beta_0$; iii) $E[\mathrm{Var}(f_\beta(X,\beta_0)|x)]$ *is nonsingular;* iv) $f(X,\beta)$ *is twice continuously differentiable and there exists* $q > 4$, $\epsilon > 0$, $d_1^f(z)$, $d_2^f(z)$ *such that for* β, $\tilde{\beta} \in \mathcal{B}$, $|f(X,\beta)| \le d_1^f(z)$, $\|f_\beta(X,\beta)\| \le d_1^f(z)$, $\|f_{\beta\beta}(X,\beta)\| \le d_2^f(z)$, $\|f_{\beta\beta}(X,\tilde{\beta})-f_{\beta\beta}(X,\beta)\| \le d_2^f(z)\|\tilde{\beta}-\beta\|^\epsilon$, $E[d_1^f(z)^q] < \infty$, $E[d_2^f(z)^2] < \infty$, $E[|\varepsilon|^q] < \infty$, $E[|h_0(x)|^q] < \infty$; v) *For every* $\beta \in \mathcal{B}$, *Assumption 3.1 is satisfied for* $g_0(x)$ *equal to each element of* $E[(h_0(x),f(X,\beta),f_\beta(X,\beta)')'|x]$; vi) *Assumption 4.3 is satisfied for* $1/4(\zeta-1) = \gamma < \Gamma = (q-4)/2q$. *Then for* $\hat{\sigma}^2 = S_n(\hat{\beta})/(n-\hat{K})$ *and* $\Omega = \sigma^2\{E[\mathrm{Var}(f_\beta(x,\beta_0)|x)]\}^{-1}$,

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,\Omega), \quad \hat{\sigma}^2[w_\beta(\hat{\beta})'\{I-P(P'P)^{-}P'\}w_\beta(\hat{\beta})/n]^{-1} \xrightarrow{p} \Omega.$$

The conditions could be weakened somewhat, along the lines discussed for Theorem 5.3. Also, if $\mathrm{Var}(f(X,\beta_0)|x)$ and $\mathrm{Var}(f_\beta(X,\beta_0)|x)$ are bounded, it can be shown that $q > 4$, $\Gamma = (q-4)/2q$ can be replaced by $q > 2$ and $\Gamma = (q-2)/2q$ respectively.

For the partially linear model, this result gives $\sqrt{n}$-consistency of Chamberlain's (1986) series estimator (which he did not show) under stronger conditions than he imposed. This estimator is asymptotically equivalent to Robinson's (1988) kernel based estimator. It is of interest to note that the conditions imposed on the distribution of x_i are quite weak; x_i is not required to be continuously distributed, and could even be discrete with infinite support.

As far as efficiency of $\hat{\beta}$ is concerned, for the partially linear model Bickel, Klaassen, Ritov, and Wellner (1989) have shown that Ω is the semiparametric efficiency bound for ε normally distributed. If ε is nonnormal and/or heteroskedastic, then the estimator need not be efficient; see Chamberlain (1987) for the form of the efficiency bound in the heteroskedastic case. Also, if heteroskedasticity is present, the estimator of the asymptotic variance given in the statement of Theorem 5.4 will be inconsistent. It is possible to develop a heteroskedasticity-consistent estimator by using Theorem 4.4. For brevity, this construction will not be discussed here.

6. Extensions

It is easy to relax the identically distributed assumption, which was made here mainly for notational convenience. All of the theorems remain true if the identically distributed assumption is dropped, but all of the conditions on existence and boundedness of moments are replaced by boundedness of average moments of a slightly higher order. Relaxing the independence assumption while retaining the generality of conditions on the series appears to be much more difficult

The general results of Section 4 should be applicable to a number of examples in addition to those in Section 5. It should be possible to combine the HGLS and best nonlinear instrumental variables to construct estimators like those of Newey (1987), that efficiently estimate parameters of conditional moment restriction models. Another example where the results may be useful is Rilstone's (1989) semiparametric missing data model.

Appendix

Throughout the appendix C and ϵ will denote (generic) positive constants that can be different in different uses. Also, $\mathcal{H}$ will denote a reference to the general Holder inequality $E[\Pi_{\ell=1}^{L}|Y_\ell|] \le \Pi_{\ell=1}^{L}(E[|Y_\ell|^{q_\ell}])^{1/q_\ell}$ for $\Sigma_{\ell=1}^{L}1/q_\ell = 1$, and $\mathcal{M}$, $\mathcal{C}$, and $\mathcal{T}$ to the to the Markov, Cauchy-Schwarz, and triangle inequalities, respectively.

Some Lemmas will be useful in proving Theorems 4.1 - 4.4. These Lemmas concern the behavior of various objects for fixed β, and it is convenient to drop the β argument for their statements and proofs. Also, it is convenient to take w and g to be scalars; corresponding results for the vector case follow by applying the results conclusions to each component. Let

(A.1) $$w_i = w(z_i,\eta_0), \quad \hat{w}_i = w(z_i,\hat{\eta}), \quad g_i = g_0(x_i), \quad \hat{g}_i = \hat{g}(x_i),$$

$$w = (w_1,\ldots,w_n)', \quad \hat{w} = (\hat{w}_1,\ldots,\hat{w}_n)', \quad g = (g_1,\ldots,g_n)', \quad \hat{g} = (\hat{g}_1,\ldots,\hat{g}_n)',$$

$$P_i^K = (p_1(x_i),\ldots,p_K(x_i))', \quad P^K = [P_1^K, \ \ldots, \ P_n^K]', \quad P = P^{\hat{K}},$$

$$\tilde{\eta}_K \in \operatorname{argmin}_\eta E[(g_i - P_i^{K\prime}\eta)^2], \quad \bar{g}_K = P^K\tilde{\eta}_K, \quad \bar{g} = \bar{g}_{\hat{K}}.$$

These Lemmas will take as hypotheses Assumptions 4.1, 4.2 and the following condition:

Assumption A.1: i) $\hat{K}^{-1} = o_p(\underline{b}^{-1})$, and $\hat{K} = o_p(b)$ for increasing $\underline{b} = \underline{b}(n)$ and $b = b(n) \longrightarrow \infty$; ii) u_{in}, $(i=1,\ldots,n)$ are i.i.d. random variables satisfying $E[u_{in}|x_i] = 0$, $E[u_{in}^2]$ is finite.

Lemma A.1: There exists nonrandom $\underline{K}$, $\bar{K}$ *such that* $\underline{K} \le \hat{K} \le \bar{K}$ *with probability approaching one,* $\underline{K}^{-1} = o(\underline{b}^{-1})$, *and* $\bar{K} = o(b)$.

Proof: Follows as in Lemma A.8 of Newey (1989a). ∎

Lemma A.2: $\|\hat{w}-w\| = O_p(1)$.

Proof: Follows from Assumption 4.1 as in Lemma A.9 of Newey (1989a). ■

Lemma A.3: For identically distributed random variables $m_{1n}, \ldots, m_{nn}$, *and any* $r > 0$, $\max_{1\le i\le n}|m_{in}| = O_p(\{nE[|m_{in}|^r\}^{1/r})$.

Proof: By the Boole and Markov inequalities,

$$\text{(A.2)} \qquad \text{Prob}(\max_{i\le n}|m_{in}| > \{nE[|m_{in}|^r\}^{1/r}C)$$

$$\le n\circ\text{Prob}(|m_{in}|^r > nE[|m_{in}|^r C^r]) \le 1/C^r. \qquad ■$$

Let $u = (u_{1n},\ldots,u_{nn})'$ and let $Q = P(P'P)^-P'$ denote the matrix of the orthogonal projection onto the space spanned by the columns of P.

Lemma A.4: For any $r > 2$, $\|Qu\| = o_p(b^{1/2}\{nE[u_{in}^r]\}^{1/r})$, *and if* $E[u_{in}^2|x_i]$ *is uniformly bounded, then* $\|Qu\| = o_p(b^{1/2})$.

Proof: Take $\bar{K}$ as in the conclusion of Lemma A.1, and let

$$\text{(A.3)} \qquad \bar{p}_i = (p_1(x_i),\ldots,p_{\bar{K}}(x_i))', \quad \bar{P} = [\bar{p}_1,\ldots\bar{p}_n]', \quad \bar{Q} = \bar{P}(\bar{P}'\bar{P})^- \bar{P}'.$$

By independence of the observations, $u_{1n},\ldots,u_{1n}$ are independent conditional on $X = (x_1,\ldots,x_n)$, implying $E[u|X] = 0$ and $\text{Var}(u|X)$ is a diagonal matrix with i^{th} diagonal element $E[u_{in}^2|x_i]$. Then by Lemma A.3 applied to $m_{in} = E[u_{in}^2|x_i]$ and the Liapunov inequality

$$\text{(A.4)} \qquad E[\|\bar{Q}u\|^2|X] = E[u'\bar{Q}u|X] = \textstyle\sum_{i,j=1}^{n}\bar{Q}_{ij}E[u_{in}u_{jn}|X]$$

$$= \textstyle\sum_{i=1}^{n}\bar{Q}_{ii}E[u_{in}^2|x_i] \le \max_{i\le n}E[u_{in}^2|x_i]\text{trace}(\bar{Q})$$

$$= O_p(\{nE[(E[u_{in}^2|x_i])^{r/2}]\}^{2/r})\text{rank}(\bar{Q}) \le O_p(\{nE[u_{in}^r]\}^{2/r})\text{rank}(\bar{P})$$

$$\le O_p(\{nE[u_{in}^r]\}^{2/r})\bar{K} = o_p(\{nE[u_{in}^r]\}^{2/r}b).$$

It follows by the conditional version of $\mathcal{M}$ and bounded convergence that $\|\bar{Q}u\|^2 = o_p(\{nE[u_{in}^r]\}^{2/r}b)$. Let $\bar{1}$ be the indicator function for the event $\hat{K} \le \bar{K}$, and note that for $\bar{1} = 1$, $\bar{Q} - Q$ is positive semi-definite, so that $\bar{1}\|Qu\|^2 \le \bar{1}\|\bar{Q}u\|^2 \le \|\bar{Q}u\|^2$. Then since $1-\bar{1} = 0$ with probability approaching one, $\|Qu\|^2 = (1-\bar{1})\|Qu\|^2 + \bar{1}\|Qu\|^2 \le (1-\bar{1})\|Qu\|^2 + \|\bar{Q}u\|^2 = o_p(\{nE[u_{in}^r]\}^{2/r}b)$. The conclusion for the case where $E[u_{in}^2|x_i]$ is bounded follows analogously, with a fixed upper bound replacing $\{nE[u_{in}^r]\}^{2/r}$. ■

Lemma A.5: $\|\bar{g}-g\| = o_p(\sqrt{n}\underline{b}^{-\zeta+1})$, and if there is $\mathcal{K}(n)$ such that the number of elements of $\mathcal{K}(n)$ is uniformly bounded and $\hat{K} \in \mathcal{K}(n)$ with probability approaching one then $\|\bar{g}-g\| = o_p(\sqrt{n}\underline{b}^{-\zeta})$.

Proof: Let $\underline{K}$, $\bar{K}$ be as in the conclusion to Lemma A.1. Let $\mathcal{K} = \{\underline{K}, \underline{K}+1, \ldots, \bar{K}\}$, so that $1_{\mathcal{K}} = 1$ with probability approaching one, where $1_{\mathcal{K}}$ is the indicator function for the event $\hat{K} \in \mathcal{K}$. Also,

$$(A.5)\qquad E[1_{\mathcal{K}}\|\bar{g}-g\|] \le E[1_{\mathcal{K}}\circ\max_{\mathcal{K}}\|\bar{g}_K-g\|] \le \sum_{\mathcal{K}}E[\|\bar{g}_K-g\|] \le \sum_{\mathcal{K}}(E[\|\bar{g}_K-g\|^2])^{1/2}$$
$$= \sqrt{n}\sum_{\mathcal{K}}(E[(\bar{g}_{Ki}-g_i)^2])^{1/2}.$$

Also, by Assumption 4.2 and the definition of $\tilde{\pi}_K$,

$$(A.6)\qquad \sqrt{n}\sum_{\mathcal{K}}(E[(\bar{g}_{Ki}-g_i)^2])^{1/2} \le \sqrt{n}\sum_{K=\underline{K}}^{\infty}(E[(\bar{g}_{Ki}-g_i)^2])^{1/2} \le C\sqrt{n}\sum_{K=\underline{K}}^{\infty}K^{-\zeta}$$
$$\le C\sqrt{n}\underline{K}^{-\zeta+1} = o(\sqrt{n}\underline{b}^{-\zeta+1}).$$

It then follows by $\mathcal{M}$ that $1_{\mathcal{K}}\|\bar{g}-g\| = o_p(\sqrt{n}\underline{b}^{-\zeta})$. The first conclusion then follows by the fact that $1-1_{\mathcal{K}} = 0$ with probability approaching one. The second conclusion follows by taking $\mathcal{K} = \mathcal{K}(n)$ and replacing eq. (A.6) with $\sqrt{n}\sum_{\mathcal{K}}(E[(\bar{g}_{Ki}-g_i)^2])^{1/2} \le C\sqrt{n}(E[(\bar{g}_{\underline{K}i}-g_i)^2])^{1/2}$, which holds by $E[(\bar{g}_{Ki}-g_i)^2]$ monotonically decreasing in K and the number of elements of $\mathcal{K}$ being bounded. ■

Lemma A.6: For $r > 2$, $|u'(\bar{g}-g)| = o_p(\sqrt{n}\underline{b}^{-\zeta+1}\{nE[u_{in}^r]\}^{1/r})$. *If* $E[u_{in}^2|x]$ *is uniformly bounded then* $\{nE[u_{in}^r]\}^{1/r}$ *in the conclusion can be replaced by* 1, *and if there is* $\mathcal{K}(n)$ *such that the number of elements of* $\mathcal{K}(n)$ *is uniformly bounded and* $\hat{K} \in \mathcal{K}(n)$ *with probability approaching one then* $-\zeta+1$ *can be replaced by* $-\zeta$. *The same conclusions holds for* $|u'Q(\bar{g}-g)|$.

Proof: Arguing as in the proof of Lemma A.5, it follows by Lemma A.3 that

$$(A.7)\quad E[1_{\mathcal{K}}|u'(\bar{g}-g)||X] \le \Sigma_{\mathcal{K}}E[|u'(\bar{g}_K-g)||X] \le \Sigma_{\mathcal{K}}(E[|u'(\bar{g}_K-g)|^2|X])^{1/2}$$

$$= \Sigma_{\mathcal{K}}((\bar{g}_K-g)'E[uu'|X](\bar{g}_K-g))^{1/2} \le \max_{i\le n}(E[u_{in}^2|x_i])^{1/2}\Sigma_{\mathcal{K}}\|\bar{g}_K-g\|$$

$$= O_p(\{nE[u_{in}^r]\}^{1/r})\Sigma_{\mathcal{K}}\|\bar{g}_K-g\|.$$

The conclusion for $|u'(\bar{g}-g)|$ then follows by arguing as in the proofs of Lemmas A.4 and A.5. Let $P^K = [P^K(x_1)\ldots,P^K(x_n)]'$ and $Q_K = P^K(P^{K\prime}P^K)^-P^{K\prime}$. Then by Q_K idempotent,

$$(A.8)\quad E[|u'Q_K(\bar{g}_K-g)|^2|X] = (\bar{g}_K-g)'Q_KE[uu'|X]Q_K(\bar{g}_K-g)$$

$$\le \max_{i\le n}E[u_{in}^2|x_i]\|Q_K(\bar{g}_K-g)\|^2 \le \max_{i\le n}E[u_{in}^2|x_i]\|\bar{g}_K-g\|^2,$$

so that the conclusion for $|u'Q(\bar{g}-g)|$ follows by an analogous argument. ■

Lemma A.7: For $r > 2$, $\|\hat{g}-g\| = o_p(b^{1/2}\{nE[|w_i|^q]\}^{1/q}) + o_p(\sqrt{n}\underline{b}^{-\zeta+1})$. *If* $E[w_i^2|x]$ *is uniformly bounded then* $\{nE[|w_i|^q]\}^{1/q}$ *in the conclusion can be replaced by* 1, *and if there is* $\mathcal{K}(n)$ *such that the number of elements of* $\mathcal{K}(n)$ *is uniformly bounded and* $\hat{K} \in \mathcal{K}(n)$ *with probability approaching one then* $-\zeta+1$ *can be replaced by* $-\zeta$.

Proof: Note that $\bar{g} = Q\bar{g}$ by Q the projection matrix for P and $\bar{g}$ a linear combination of P. The conclusion then follows by Lemmas A.2, A.4

(with $u = w-g$), and A.5, since by $\mathcal{T}$ and Q idempotent, $\|\hat{g}-g\| \le \|Q(\hat{w}-w)\| + \|Q(w-g)\| + \|Q(g-\bar{g})\| + \|\bar{g}-g\| \le \|\hat{w}-w\| + \|Q(w-g)\| + 2\|\bar{g}-g\|$. ■

Lemma A.8: For $\nu > 2$, $|u'(\hat{g}-g)| = o_p(\bar{b}\{nE[u_{in}^{\nu}]\}^{1/\nu}\{nE[|w_i|^q]\}^{1/q}) + o_p(\sqrt{n}\underline{b}^{-\zeta+1}\{nE[u_{in}^{\nu}]\}^{1/\nu})$. If $E[w_i^2|x]$ is uniformly bounded then $\{nE[\|w_i\|^q]\}^{1/q}$ in the conclusion can be replaced by 1, if $E[u_{in}^2|x]$ is uniformly bounded then $\{nE[u_{in}^{\nu}]\}^{1/\nu}$ in the conclusion can be replaced by 1, and if there is $\mathcal{K}(n)$ such that the number of elements of $\mathcal{K}(n)$ is uniformly bounded and $\hat{K} \in \mathcal{K}(n)$ with probability approaching one then $-\zeta+1$ can be replaced by $-\zeta$.

Proof: The conclusion follows from Lemmas A.2, A.4, A.5, and A.6, since by $\mathcal{T}$ and $\mathcal{E}$, and Q idempotent,

$$(A.9) \qquad |u'(\hat{g}-g)| \le |u'Q(\hat{w}-w)| + |u'Q(w-g)| + |u'Q(g-\bar{g})| + |u'(\bar{g}-g)|$$

$$\le \|Qu\|\|\hat{w}-w\| + \|Qu\|\|Q(w-g)\| + |u'Q(g-\bar{g})| + |u'(\bar{g}-g)|. \quad ■$$

Lemma A.9: For $A^n(x)$ suppose that there exists $\xi > 1$ and η_K^n such that for $A_i^n = A^n(x_i)$, $(E[(A_i^n - p_i^{K\prime}\eta_K^n)^2])^{1/2} = O(K^{-\xi})$ uniformly in n. Then

$$\sum_{i=1}^{n} A_i^n(\hat{w}_i-\hat{g}_i)/\sqrt{n} = o_p(\underline{b}^{-\xi+1}b^{1/2}(nE[\|w_i\|^q])^{1/q}) + o_p(\sqrt{n}\underline{b}^{-\xi-\zeta+2}).$$

If $E[w_i^2|x]$ is uniformly bounded then $\{nE[\|w_i\|^q]\}^{1/q}$ in the conclusion can be replaced by 1 and if there is $\mathcal{K}(n)$ such that the number of elements of $\mathcal{K}(n)$ is uniformly bounded and $\hat{K} \in \mathcal{K}(n)$ with probability approaching one then $-\zeta+1$ and $-\xi+1$ can be replaced by $-\zeta$ and $-\xi$ respectively.

Proof: Let $A = (A_1^n,\ldots,A_n^n)'$, $\bar{A} = P\eta_K^n$, and note that by Lemma A.5 applied to A, $\|A-\bar{A}\| = o_p(\sqrt{n}\underline{b}^{-\xi+1})$. Also, $P'(\hat{w}-\hat{g}) = 0$ holds by orthogonality of least squares residuals. The conclusion then follows from Lemmas A.2, A.6 (with $u = w-g$), and A.7, since by $\mathcal{T}$ and $\mathcal{E}$, (A.10) $\quad |A'(\hat{w}-\hat{g})| = |(A-\bar{A})'(\hat{w}-\hat{g})| \le \|A-\bar{A}\|(\|\hat{w}-w\|+\|\hat{g}-g\|) + |(A-\bar{A})'(w-g)|.$ ∎

Proof of Theorem 3.1: The conclusion follows immediately from Lemma A.7 and $\sum_{i=1}^{n}[\hat{g}_i-g_i]^2/n = \|\hat{g}-g\|^2/n$. ∎

Proof of Theorem 4.1: By a Taylor expansion and Assumptions 4.2 and 4.3,

$$(A.11) \qquad |a(z,\beta,\tilde{g})-a(z,\beta,g)| = |a_g(z,\beta,\bar{g})'(\tilde{g}-g)| \le [d_{1n}(z)+d_{2n}(z)\|\bar{g}\|]\|\tilde{g}-g\|$$

$$\le [d_{1n}(z)+d_{2n}(z)(\|\tilde{g}\|+\|g\|)]\|\tilde{g}-g\|,$$

where g, $\tilde{g}$, $\tilde{g} \in \mathbb{R}^s$, $\bar{g}$ is the mean value, and the final inequality follows by $\bar{g}$ on the line joining g and $\tilde{g}$. Let $\tilde{A}_n(\beta) = \sum_{i=1}^{n}a^n(z_i,\beta,g_0(x_i))/n$. By Q idempotent and Assumption 4.1, $\|\hat{g}\| \le \|d_w\|$ with probability approaching one, for $d_w = (d_w(z_1),\ldots,d_w(z_n))'$. Then by $\mathcal{E}$ and $\mathcal{M}$, eq. (A.11), Lemma A.3, and Lemma A.7, with $\underline{b} = n^{\gamma}$ and $b = n^{\Gamma}$,

$$(A.12) \qquad \sup_{\beta\in\mathcal{B}}|A_n(\beta)-\tilde{A}_n(\beta)| \le C(\|d_1\|+\max_{1\le i\le n}d_{2n}(z_i)\{\|d_w\|+\|g\|\})\|\hat{g}-g\|/n$$

$$= O_p(n^{r} + n^{r'})O_p(1)o_p(n^{\Gamma/2+1/q-1/2} + n^{-\gamma(\zeta-1)}) = o_p(1).$$

where $d_1 = (d_1(z_1),\ldots,d_1(z_n))'$. Also, by Assumption 4.3,

$$(A.13) \qquad |\tilde{A}_n(\tilde{\beta})-\tilde{A}_n(\beta)| \le (\textstyle\sum_{i=1}^{n}d_{3n}(z_i)/n)\|\tilde{\beta}-\beta\|^{\epsilon}.$$

Then by $\bar{A}_n(\beta) = E[\tilde{A}_n(\beta)]$, $|\bar{A}_n(\tilde{\beta})-\bar{A}_n(\beta)| \le E[|\tilde{A}_n(\tilde{\beta})-\tilde{A}_n(\beta)|] \le E[d_{3n}(z)]\|\tilde{\beta}-\beta\|^{\epsilon}$ $\le C\|\tilde{\beta}-\beta\|^{\epsilon}$, so that $\bar{A}_n(\beta)$ is equicontinuous. Also, by Assumption (4.2) and Markov's weak law of large numbers,

$$(A.14) \qquad |\tilde{A}_n(\beta)-\bar{A}_n(\beta)| = o_p(1), \quad \text{for each } \beta \in \mathcal{B}.$$

It then follows by eqs. (A.13) and (A.14) and Corollary 1 of Newey (1989c) that $\sup_{\beta\in\mathcal{B}}|\tilde{A}_n(\beta)-\bar{A}_n(\beta)| = o_p(1)$. The conclusion then follows from eq.

(A.12) and $\mathcal{T}$. ■

For the proof of Theorem 4.2 the β argument will be explicitly stated for the objects of eq. (A.1); e.g. $\hat{g}(\beta) = (\hat{g}(x_1,\beta),\ldots,\hat{g}_n(x,\beta))'$.

Proof of Theorem 4.2: Note that by Assumptions 4.2 and 4.4, eq. (A.11) holds with $d_{2n}(z)$ there replaced by C. Note that Assumption 4.1 implies $E[\|g_0(x,\beta)\|^2] \le E[\|w(z,\beta,\eta_0)\|^2] \le E[d_w(z)^2]$ and $E[\|g_0(x,\tilde{\beta})-g_0(x,\beta)\|^2] \le E[\|w(z,\tilde{\beta},\eta_0)-w(z,\beta,\eta_0)\|^2] \le E[d_w(z)^2]\|\tilde{\beta}-\beta\|^{2\epsilon}$. Then for $\tilde{\beta}$, $\beta \in \mathcal{B}$, it follows by Assumption 4.4 that

$$\text{(A.15)}\qquad |\bar{A}_n(\tilde{\beta})-\bar{A}_n(\beta)| \le E[|a^n(z,\tilde{\beta},g_0(x,\tilde{\beta}))-a^n(z,\tilde{\beta},g_0(x,\beta))|]$$

$$+ E[|a^n(z,\tilde{\beta},g_0(x,\beta))-a^n(z,\beta,g_0(x,\beta))|]$$

$$\le \{(E[d_{1n}(z)^2])^{1/2}+CE[d_w(z)^2])^{1/2}\}(E[d_w(z)^2])^{1/2}\|\tilde{\beta}-\beta\|^{\epsilon}$$

$$+ (E[d_{2n}(z)]+CE[\|g_0(x,\beta)\|^2])\|\tilde{\beta}-\beta\|^{\epsilon} \le C\|\tilde{\beta}-\beta\|^{\epsilon}.$$

Thus, $\bar{A}_n(\beta)$ is equicontinuous. Similarly, by $\|\hat{g}(\beta)\| \le \|d_w\|$ with probability approaching one and $\|\hat{g}(\tilde{\beta})-\hat{g}(\beta)\| = \|Q[\hat{w}(\tilde{\beta})-\hat{w}(\beta)]\| \le \|\hat{w}(\tilde{\beta})-\hat{w}(\beta)\| \le \|d_w\|\|\tilde{\beta}-\beta\|^{\epsilon}$ with probability approaching one, and by Assumption 4.4,

$$\text{(A.16)}\qquad |A_n(\tilde{\beta})-A_n(\beta)| \le \textstyle\sum_{i=1}^n|a^n(z_i,\tilde{\beta},\hat{g}(x_i,\tilde{\beta}))-a^n(z_i,\tilde{\beta},\hat{g}(x_i,\beta))|/n$$

$$+ \textstyle\sum_{i=1}^n|a^n(z_i,\tilde{\beta},\hat{g}(x_i,\beta))-a^n(z_i,\beta,\hat{g}(x_i,\beta))|/n$$

$$\le \{(\|d_1\| + C\|d_w\|)\|d_w\|/n\}\|\tilde{\beta}-\beta\|^{\epsilon}$$

$$+ \{\textstyle\sum_{i=1}^n d_{2n}(z_i)/n + C\|d_w\|^2/n\}\|\tilde{\beta}-\beta\|^{\epsilon} \le O_p(1)\|\tilde{\beta}-\beta\|^{\epsilon},$$

where the $O_p(1)$ term following the last equality is independent of β. Let $\tilde{A}_n(\beta) = \sum_{i=1}^n a^n(z_i,\beta,g_0(x_i,\beta))/n$. By Assumption 4.2 and Markov's law of large numbers, $\tilde{A}_n(\beta)-\bar{A}_n(\beta) = o_p(1)$ for each $\beta \in \mathcal{B}$. Also, it follows by

Assumptions 4.1, 4.2, 4.4, and Lemma A.7 that for each $\beta \in \mathcal{B}$,

$$(A.17)\qquad |A_n(\beta)-\tilde{A}_n(\beta)| \le \{\|d_1\| + C\|\hat{g}(\beta)\| + C\|g(\beta)\|)\}\|\hat{g}(\beta)-g(\beta)\|/n$$

$$\le \{\|d_1\| + C\|d_w\| + C(\textstyle\sum_{i=1}^{n}E[d_w(z)^2|x_i])^{1/2})\}\|\hat{g}(\beta)-g(\beta)\|/n$$

$$= O_p(1)\|\hat{g}(\beta)-g(\beta)\|/\sqrt{n} = o_p(n^{\Gamma/2+1/q-1/2} + n^{-\gamma(\zeta-1)}) = o_p(1).$$

Then by $\mathcal{T}$, $|A_n(\beta)-\bar{A}_n(\beta)| = o_p(1)$. The conclusion then follows by Corollary 1 of Newey (1989c). ■

Proof of Theorem 4.3: Consider first the scalar w and g case. A Taylor expansion gives

$$(A.18)\qquad \textstyle\sum_{i=1}^{n}a^n(z_i,\hat{g}_i)/\sqrt{n} = \sum_{i=1}^{n}a_i^n/\sqrt{n} + \sum_{i=1}^{n}a_{gi}^n(\hat{g}_i-g_i)/\sqrt{n}$$

$$+ \textstyle\sum_{i=1}^{n}[a_g^n(z_i,\tilde{g}_i)-a_{gi}^n](\hat{g}_i-g_i)/\sqrt{n} \equiv T_0 + T_1 + T_2,$$

where $\tilde{g}_i$ lies between $\hat{g}_i$ and g_i, $a_i^n = a^n(z_i,g_i)$, and $a_{gi}^n = a_g^n(z_i,g_i)$. It follows by Lemmas A.3 and A.7 with $\underline{b} = n^{\gamma}$ and $b = n^{\Gamma}$, Assumption 4.5, and the conditions on γ and Γ that

$$(A.19)\qquad |T_2| \le (\max_{1\le i\le n}d_{2n}(z_i))\|\hat{g}-g\|^2/\sqrt{n}$$

$$= O_p(\{nE[d_{2n}(z)^{\nu'}]\}^{1/\nu'})[o_p(n^{\Gamma+2/q-1/2})+o_p(n^{-2\gamma(\zeta-1)+1/2})]$$

$$= o_p(n^{r'+\Gamma+2/q-1/2} + n^{-2\gamma(\zeta-1)+1/2+r'}) = o_p(O(1)) = o_p(1).$$

Also, note that for $G_i^n = E[a_{gi}^n|x_i]$, G_i^n

$$(A.20)\qquad T_1 = \textstyle\sum_{i=1}^{n}(a_{gi}^n-G_i^n)(\hat{g}_i-g_i)/\sqrt{n} + \sum_{i=1}^{n}G_i^n(\hat{g}_i-\hat{w}_i)/\sqrt{n} +$$

$$\textstyle\sum_{i=1}^{n}G_i^n(w_i-g_i)/\sqrt{n} + \sum_{i=1}^{n}G_i^n(\hat{w}_i-w_i)/\sqrt{n} \equiv T_{11} + T_{12} + T_{13} + T_{14}.$$

By Lemma A.8, with $u_{in} = a_{gi}^n - G_i^n$, and Assumption 4.5,

$$(A.21)\qquad |T_{11}| = o_p(n^{\Gamma+r+1/q-1/2}) + o_p(n^{-\gamma(\zeta-1)+r}) = o_p(O(1)) = o_p(1).$$

Note that if $G_i^n = 0$, then $T_{12} = T_{13} = T_{14} = 0$, so that the conclusion follows from eqs. (A.20) and (A.21). For the other case it follows by Lemma A.9 with $A_i^n = G_i^n$ and Assumption 4.6 that

$$(A.22)\qquad T_{12} = o_p(n^{-\gamma(\xi-1)}n^{\Gamma/2+1/q}) + o_p(n^{-\gamma(\xi+\zeta-2)+1/2}) = o_p(1).$$

Next, note that $E[\|G_i^n\|^{\nu_G}]$ is bounded by Assumption 4.6, so that for $\eta \in N$, $\|G_i^{n\prime}w(z_i,\eta_0)\| \le \|G_i^n\|d_w(z_i)$ and $E[\{\|G_i^n\|d_w(z_i)\}^{1+C}]$ is bounded for some $C > 0$, so that by Markov's law of large numbers, $\sum_{i=1}^n G_i^n w_\eta(z_i,\eta_0)/n - E[G_i^n w_{\eta i}] = o_p(1)$. Then by a Taylor expansion,

$$\begin{aligned}(A.23)\qquad T_{24} &= \textstyle\sum_{i=1}^n G_i^n(\hat{w}_i-w_i)/\sqrt{n} - E[G_i^n w_\eta(z_i,\eta_0)]\sqrt{n}(\hat{\eta}-\eta_0) \\ &= [\textstyle\sum_{i=1}^n G_i^n\{w_\eta(z_i,\tilde{\eta})-w_\eta(z_i,\eta_0)\}/n]\sqrt{n}(\hat{\eta}-\eta_0) + o_p(1) \\ &= \{O_p(E[G_i^n d_w(z_i)])\|\tilde{\eta}-\eta\|^{\epsilon}\sqrt{n}\|\hat{\eta}-\eta_0\| + o_p(1) = o_p(1)\end{aligned}$$

where $\tilde{\eta}$ denotes the mean value. The conclusion then follows by eqs. (A.20) - (A.23). The conclusion for the vector g case then follows in this exact way by applying the above argument to a expansion in the vector $g(x)$. ■

Proof of Theorem 4.4: By $\mathcal{T}$ (in ℓ^2) and $\mathcal{C}$,

$$\begin{aligned}(A.24)\qquad \textstyle\sum_{i=1}^n\|\hat{\psi}_i^n-\psi_i^n\|^2/n &\le C\{\textstyle\sum_{i=1}^n\|(\hat{a}_i^n-\sum_{j=1}^n\hat{a}_j^n/n) - (a_i^n-\sum_{j=1}^n a_j^n/n)\|^2/n \\ &\quad + \textstyle\sum_{i=1}^n\|\hat{G}_i^n(\hat{w}_i-\hat{g}_i^\Delta) - G_i^n(w_i-g_i)\|^2/n + \|\hat{H}_n\|^2\sum_{i=1}^n\|\hat{\psi}_{ni}^\eta-\psi_{ni}^\eta\|^2/n \\ &\quad + \|\hat{H}_n-H_n\|^2\textstyle\sum_{i=1}^n\|\psi_{ni}^\eta\|^2/n\} \equiv T_1 + T_2 + T_3 + T_4,\end{aligned}$$

where $T_2 = T_3 = T_4 = 0$ if $G^n(x_i) = 0$. Let $g_i^\Delta = \tau^\Delta(g_i)$. Note that $\tau^\Delta(\circ)$ is uniformly Lipschitz, so that $\|\tau^\Delta(\tilde{g})-\tau^\Delta(g)\| \le C\|\tilde{g}-g\|$. Then by a Taylor expansion, Assumption 4.8, Lemma A.3, and Lemma A.7,

$$\text{(A.25)}\qquad \textstyle\sum_{i=1}^{n}\|\hat{a}_i^n-a^n(z_i,\hat{\beta},g_i^{\Delta})\|^2/n = \sum_{i=1}^{n}\|a_g^n(z_i,\hat{\beta},\bar{g}_i^{\Delta})(\hat{g}_i^{\Delta}-g_i^{\Delta})\|^2/n$$

$$\le C[\max_{1\le i\le n}d_{3n}(z_i)^2]\|\hat{g}-g\|^2/n = o_p(n^{2r_3+\Gamma+2/q-1}+n^{2r_3-2\gamma(\zeta-1)}) = o_p(1),$$

where the intermediate value $\bar{g}_i^{\Delta}$ lies on the line joining $\hat{g}_i^{\Delta}$ and g_i^{Δ}. Note that $g_i^{\Delta} = g_i$ for $\|g_i\| \le C\Delta$ and that $\|g_i^{\Delta}\| \le \|g_i\|$. Then by another Taylor expansion, Assumption 4.8, $\Delta \longrightarrow \infty$, and $\mathcal{M}$ and $\mathcal{H}$,

$$\text{(A.26)}\qquad \textstyle\sum_{i=1}^{n}[a^n(z_i,\hat{\beta},g_i^{\Delta})-a^n(z_i,\hat{\beta},g_i)]^2/n$$

$$\le \textstyle\sum_{i=1}^{n}\sup_{\beta\in\mathcal{B},\,\|g\|\le\|g_i\|}\|a_g^n(z_i,\beta,g)\|^2\|g_i^{\Delta}-g_i\|^2/n$$

$$\le C\textstyle\sum_{i=1}^{n}d_{4n}(z_i)^2\|g_i\|^2 1(\|g_i\|\ge C\Delta)/n$$

$$= O_p(\{E[d_{4n}(z)^{2q/(q-2)}]\}^{(q-2)/q}\{E[1(\|g_i\|\ge C\Delta)\|g_i\|^q]\}^{2/q})$$

$$= O_p(o(1)) = o_p(1),$$

where the next to last equality follows by $\Delta \longrightarrow \infty$ and Assumption 4.1. In addition, it follows by Assumption 4.9 that with probability approaching one,

$$\text{(A.27)}\qquad \textstyle\sum_{i=1}^{n}\|a^n(z_i,\hat{\beta},g_i)-a_i^n\|^2/n \le [\sum_{i=1}^{n}d_{\beta n}(z_i)^2/n]\|\hat{\beta}-\beta_0\|^{2\epsilon} = O_p(1)o_p(1)$$

$$= o_p(1).$$

Then by eqs. (A.25) - (A.27) and $\mathcal{T}$ and $\mathcal{E}$ $T_1 \le \sum_{i=1}^{n}\|\hat{a}_i^n-a_i^n\|^2/n = o_p(1)$, so that in the $G^n(x) = 0$ case,

$$\text{(A.28)}\qquad \textstyle\sum_{i=1}^{n}\|\hat{\psi}_i^n-\psi_i^n\|^2/n = o_p(1).$$

Next, let $\tilde{G}_i^n$ be the matrix with ℓ^{th} row $(\tilde{G}_i^n)_{\ell} = p^K(x_i)'(P'P)^{-}\sum_{j=1}^{n}p^K(x_j)(a_g^n(z,\hat{\beta},g_j))_{\ell}$. By Q idempotent, a Taylor expansion, and Assumption 4.5,

$$\text{(A.29)}\qquad \textstyle\sum_{i=1}^{n}\|\hat{G}_i^n-\tilde{G}_i^n\|^2/n \le \sum_{i=1}^{n}\|a_g^n(z,\hat{\beta},\hat{g}_i)-a_g^n(z,\hat{\beta},g_i)\|^2/n$$

$$\le [\max_{1\le i\le n}d_{2n}(z_i)^2]\textstyle\sum_{i=1}^{n}\|\hat{g}_i-g_i\|^2/n$$

$$= o_p(n^{2r'+\Gamma+2/q-1} + n^{2r'-2\gamma(\zeta-1)}) = o_p(1).$$

Note that $\tilde{G}_i^n$ is a series conditional expectation estimate of G_i^n for $w_n(z,\eta) = a_g^n(z,\beta,g_0(x))$ and $\eta = \beta$. Note that $\|a_g^n(z,\tilde{\beta},g_0(x))-a_g^n(z,\beta,g_0(x))\| \le d_{2n}^{\beta}(z)\|\tilde{\beta}-\beta\|$, $E[d_{2n}^{\beta}(z)^2]$ bounded, $(nE[\|a_g^n(z,\beta_0,g_0(x))\|^{\nu}])^{1/\nu} = O(n^{r})$, and Assumption 4.6 correspond to Assumptions 3.1 and 4.1 applied to $w_n(z,\eta)$. It then follows as in Lemma A.7 with $q = \nu$ that

$$\text{(A.30)}\qquad \textstyle\sum_{i=1}^{n}\|\tilde{G}_i^n-G_i^n\|^2/n = o_p(n^{\Gamma+2r-1}+n^{-2\gamma(\xi-1)}) = o_p(1).$$

Then by eqs. (A.29), (A.30), $\mathcal{T}$, $E[d_w(z_i)^{1/t}] < \infty$, Lemma A.2, and $\Delta = O(n^{t})$,

$$\text{(A.31)}\qquad \textstyle\sum_{i=1}^{n}\|(\hat{G}_i^n-G_i^n)(\hat{w}_i-\hat{g}_i^{\Delta})\|^2/n \le C(\max_{1\le i\le n}d_w(z_i)^2+\Delta^2)\sum_{i=1}^{n}\|\hat{G}_i^n-G_i^n\|^2/n$$

$$= o_p(n^{2t+2r'+\Gamma+2/q-1} + n^{2t+2r'-2\gamma(\zeta-1)}) + o_p(n^{2t+\Gamma+2r-1}+n^{2t-2\gamma(\xi-1)})$$

$$= o_p(1).$$

Also by $\mathcal{T}$, Lemmas A.2 and A.3, and a Taylor expansion, it follows that with probability approaching one,

$$\text{(A.32)}\qquad \textstyle\sum_{i=1}^{n}\|G_i^n[(\hat{w}_i-\hat{g}_i^{\Delta})-(w_i-g_i^{\Delta})\|^2/n$$

$$\le C\max_{1\le i\le n}\|G_i^n\|^2\{\|\hat{w}-w\|^2/n + C\|\hat{g}-g\|^2/n\}$$

$$= O_p(n^{2/\nu_G})[O_p(n^{-1}) + o_p(n^{\Gamma+2/q-1}+n^{-2\gamma(\zeta-1)})]$$

$$= o_p(n^{2/\nu_G+\Gamma+2/q-1}+n^{2/\nu_G-2\gamma(\zeta-1)}) = o_p(1).$$

Also by $\nu_G > 2q/(q-2)$, $E[\|G_i^n\|^{2q/(q-2)}]$ is bounded. Then by $\mathcal{M}$ and $\mathcal{H}$,

$$(A.33)\qquad \sum_{i=1}^n \|G_i^n(g_i^\Delta - g_i)\|^2/n \le \sum_{i=1}^n \|G_i^n\|^2\|g_i\|^2 1(\|g_i\|>C\Delta)/n$$

$$= O_p(\{E[\|G_i^n\|^{2q/(q-2)}]\}^{(q-2)/q}\{E[\|g_i\|>C\Delta\|g_i\|^q]\}^{2/q}) = o_p(1).$$

$T_2 = o_p(1)$ then follows by eqs. (A.31) - (A.33) and $\mathcal{T}$.

Next, note that with probability approaching one, $\sum_{i=1}^n \|w_\eta(z_i,\hat{\beta},\hat{\eta})-w_\eta(z_i,\beta_0,\eta_0)\|^2/n \le [\sum_{i=1}^n d_w(z_i)^2/n](\|\hat{\beta}-\beta_0\|^2+\|\hat{\eta}-\eta_0\|^2) = O_p(1)o_p(1) = o_p(1)$. Also note that by $\nu_G > 2$ and $E[d_w(z)^2]$ finite there exists $C > 0$ such that $E[\|G^n(x)'w_\eta(z,\beta_0,\eta_0)\|^{1+C}] \le E[\{\|G^n(x)\|d_w(z)\}^{1+C}]$ is bounded. It then follows by Markov's weak law of large numbers, eqs. (A.29) and (A.30), and $\mathcal{T}$ and $\mathcal{C}$ that

$$(A.34)\qquad \|\hat{H}_n-H_n\| \le \sum_{i=1}^n \|\hat{G}_i^{n\prime} w_\eta(z_i,\hat{\beta},\hat{\eta})-G_i^{n\prime} w_\eta(z_i,\beta_0,\eta_0)\|/n$$

$$+ \|\sum_{i=1}^n G_i^n w_\eta(z_i,\beta_0,\eta_0)/n - H_n\|$$

$$\le (\sum_{i=1}^n \|\hat{G}_i^n-G_i^n\|^2/n)^{1/2}\|d_w\|/\sqrt{n}$$

$$+ (\sum_{i=1}^n \|G_i^n\|^2/n)^{1/2}(\sum_{i=1}^n \|w_\eta(z_i,\hat{\beta},\hat{\eta})-w_\eta(z_i,\beta_0,\eta_0)\|^2/n)^{1/2} + o_p(1)$$

$$= o_p(1).$$

Since $E[\|\psi_n^\eta(z)\|^2]$ is bounded, it follows that $T_4 = o_p(1)$. Also, by H_n bounded and eq. (A.34), $\|\hat{H}_n\|^2 = O_p(1)$, so that $T_3 = o_p(1)$. Therefore, by eq. (A.24) and $\mathcal{T}$, it follows that eq. (A.28) also holds for the case where $G^n(x) \ne 0$. The first conclusion then follows by $\mathcal{T}$ and $\mathcal{M}$ as well as Markov's law of large numbers, since

$$\text{(A.35)}\quad \|\hat{\Sigma}_n-\Sigma_n\| \le \sum_{i=1}^n\|\hat{\psi}_i^n\hat{\psi}_i^{n\prime}-\psi_i^n\psi_i^{n\prime}\|/n + \|\sum_{i=1}^n\psi_i^n\psi_i^{n\prime}/n - E[\psi_i^n\psi_i^{n\prime}]\|/n$$

$$\le 2\sum_{i=1}^n\|\psi_i^n\|\|\hat{\psi}_i^n-\psi_i^n\|/n + \sum_{i=1}^n\|\hat{\psi}_i^n-\psi_i^n\|^2/n + o_p(1)$$

$$\le 2(\sum_{i=1}^n\|\psi_i^n\|^2/n)^{1/2}(\sum_{i=1}^n\|\hat{\psi}_i^n-\psi_i^n\|^2/n)^{1/2} + o_p(1) = o_p(1).$$

Next, note that by Assumptions 4.1 and 4.5 - 4.7 and $\mathcal{T}$, there exists $\epsilon > 0$ such that $E[\|\psi_i^n\|^{2+\epsilon}]$ is bounded. The second conclusion then follows from Theorem 4.3 by application of the Liapunov central limit theorem and Slutzky's theorem. ■

Proof of Theorem 5.1: Consider $a(z,g) = (w-g)^2$, so that $a_g(z,g) = -2(w-g)$. Note that η is nonexistent, so that Assumption 4.1 is satisfied for q as given in the statement of Theorem 5.1. Also, Assumption 4.5 is satisfied by $q > 4$, with $\nu = q$, $r = 1/q$, and $d_{2n}(z)$ bounded. By the remarks following Theorem 4.3, r' can be taken equal to zero in the hypotheses of Theorem 4.3, so that the conditions on γ and Γ are $\gamma = \max\{1/q, 1/4\}/(\zeta-1)$, $\Gamma = 1/2 - \max\{1/q + 1/q,\ 2/q + 0\} = 1/2 - 2/q$. It then follows by Theorem 4.3 that $\sqrt{n}[\sum_{i=1}^n(w_i-\hat{g}_i)^2/n - \sigma^2] \xrightarrow{d} N(0,\Omega)$. Furthermore, by the hypotheses on Γ, $\sqrt{n}[\sum_{i=1}^n(w_i-\hat{g}_i)^2/n - \hat{\sigma}^2] = \sqrt{n}[(n-\hat{K})/n - 1]\hat{\sigma}^2 = (-\hat{K}/\sqrt{n})\hat{\sigma}^2 = o_p(1)$, giving the first conclusion.

Next, Assumption 4.8 is satisfied with $d_{3n}(z) = |w| + C\Delta$, $\nu_3 = q$, $r_3 = 1/q + t$, and $d_{4n}(z) = |w| + |g_0(x)|$, since $E[d_{4n}(z)^{2q/(q-2)}] = E[(|w| + |g_0(x)|)^{2q/(q-2)}]$ is finite by implying $2q/(q-2) < q$ (which is implied by $q > 4$). The conditions on γ and Γ in the hypotheses of Theorem 4.4 then become $\gamma \ge (1/q + t)/(\zeta-1)$, $\Gamma \le 1 - 2(1/q + t) - 2/q = 1 - 2t - 4/q$. The second conclusion then follows by Theorem 4.4. ■

Proof of Theorem 5.2: First, $\hat{\beta} = \beta_0 + o_p(1)$ will be shown. By v), Assumption 4.1 is satisfied with $d_w(z) = 0$. Consider $a(z,g,\beta)$ equal to an element of $B(X)\rho(z,g,\beta)$. By iv) and v) and $\mathcal{H}$, Assumption 4.2 is satisfied for $\epsilon = 2$. By a Taylor expansion and convexity of $\mathcal{B}$,

$$\text{(A.36)}\quad |a(z,\tilde{\beta},g_0(x))-a(z,\beta,g_0(x))| = |a_\beta(z,\bar{\beta},g_0(x))'(\tilde{\beta}-\beta)| \le \|B(X)\|d_\beta(z)\|\tilde{\beta}-\beta\|.$$

Then by v) and $\mathcal{H}$, Assumption 4.3 is satisfied for $d_{jn}(z) = \|B(X)\|d_{gj}(z)$, $j=1,2$, $d_{3n}(z) = \|B(X)\|d_\beta(z)$, $r = 0$, $r' = 1/\nu' = 1/\nu_B + \max\{1/\nu_1, 1/\nu_2\}$. It follows by vii) that $\gamma > r'/(\zeta-1)$ and $\Gamma < 2(1/2 - 1/q - r')$. Then by the conclusion of Theorem 4.3 applied to each element of $B(X)\rho(z,g,\beta)$, $m_n(\beta)$ converges uniformly in probability to $\bar{m}(\beta) = E[B(X)\rho(z,g_0(x),\beta)]$, which is continuous. It then follows by a standard argument (e.g. see Hansen (1982)) that $S_n(\beta)$ converges uniformly in probability to $S_0(\beta) = \bar{m}(\beta)'\Psi\bar{m}(\beta)$, which is continuous in β, and has a unique minimum at β_0 by ii). Consistency of $\hat{\beta}$ then follows by Lemma 3 of Amemiya (1973).

Next, it follows by v) and the identical argument to that given above with $a(z,\beta,g)$ an element of $B(X)\rho_\beta(z,g,\beta)'$ that $M_n(\beta) = \partial m_n(\beta)/\partial\beta$ converges uniformly in probability to $\bar{M}(\beta) = E[B(X)\rho_\beta(z,g,\beta)']$, which is continuous in β. Then by Lemma 4 of Amemiya (1973), for any $\tilde{\beta} = \beta_0 + o_p(1)$,

$$\text{(A.37)}\qquad M_n(\tilde{\beta}) = M + o_p(1).$$

Next, let $a(z,\beta,g)$ be an element of $B(X)\rho(z,g,\beta)$ again. By $\mathcal{H}$ and $\nu_\beta > 2\nu_B/(\nu_B-2)$ there exists $\epsilon > 0$ such that the first hypothesis of Assumption 4.5 is satisfied. Also, by v), a Taylor expansion argument like eq. (A.36), and $\mathcal{H}$, the other hypotheses are satisfied with $d_{1n}(z) = \|B(X)\|(d_1(z) + d_2(z)\|g_0(x)\|)$, $d_{2n}(z) = d_3(z)$, $r = 1/\nu = 1/\nu_B + \max\{1/\nu_1, 1/\nu_2 + 1/q\}$, $r' = 1/\nu' = 1/\nu_3$. Also, by $\|a_g\| \le \|B(X)\|(d_1(z) + d_2(z)\|g_0(x)\|)$, Assumption 4.6 is satisfied with $\nu_G = \nu$. Furthermore, by vii), the hypotheses on γ and Γ of Theorem 4.3 are satisfied, so that by its conclusion applied to each element of $\sqrt{n}m_n(\beta_0)$, (A.38) $\sqrt{n}m_n(\beta_0) = \sum_{i=1}^n\psi_i/\sqrt{n}$, $\psi_i = B(X_i)\rho(z_i,g_i,\beta_0) + G(x_i)[w_i-g_i]$.

The convergence in distribution result now follows by a standard Taylor expansion argument, vis; eq. (A.37), ii), and iii) imply that for $\tilde{\beta} = \beta_0 + o_p(1)$, $[M_n(\hat{\beta})'\hat{\Psi}M_n(\tilde{\beta})]^{-1} = (M'\Psi M)^{-1} + o_p(1)$, while the central limit theorem and eq. (A.38), $\sqrt{n}m_n(\beta_0) = O_p(1)$, so that since β_0 interior and $\hat{\beta}$ consistent implies $0 =(1/2)\partial S_n(\hat{\beta})/\partial\beta = M_n(\hat{\beta})'\hat{\Psi}m_n(\hat{\beta})$ with probability approaching one, expanding $m_n(\hat{\beta})$ around β_0 and solving for $\hat{\beta}$,

(A.39) $$\sqrt{n}(\hat{\beta}-\beta_0) = [M_n(\hat{\beta})'\hat{\Psi}M_n(\bar{\beta})]^{-1}M_n(\hat{\beta})'\hat{\Psi}\sqrt{n}m_n(\beta_0) + o_p(1),$$

$$= (M'\Psi M)^{-1}M'\Psi\sqrt{n}m_n(\beta_0) \ + o_p(1) \xrightarrow{d} N(0,\ (M'\Psi M)^{-1}M'\Psi\Sigma\Psi M(M'\Psi M)^{-1}),$$

where the second equality follows by $M_n(\hat{\beta})'\hat{\Psi} = M'\Psi + o_p(1)$ and Slutzky's theorem.

Next, since $[M_n(\hat{\beta})'\hat{\Psi}M_n(\hat{\beta})]^{-1}M_n(\hat{\beta})'\hat{\Psi} = (M'\Psi M)^{-1}M'\Psi + o_p(1)$ follows as above, to finish the proof it suffices to show that $\hat{\Sigma} = \Sigma + o_p(1)$, via Theorem 4.4. By v), Assumption 4.8 is satisfied with $d_{3n}(z) = \|B(X)\|[d_1(z)+d_2(z)\Delta(n)]$, $d_{4n}(z) = \|B(X)\|[d_1(z)+d_2(z)\|g_0(x)\|]$, $r_3 = 1/\nu_3 + t = 1/\nu_B + \max\{1/\nu_1,1/\nu_2\} + t$. Also, since $\sqrt{n}(\hat{\beta}-\beta_0) = O_p(1)$ holds by eq. (A.39), Assumption 4.9 is also satisfied by v). Noting that the gamma conditions are satisfied by v), the conclusion then follows by eq. (4.8). ■

Proof of Theorem 5.3: Note $\sqrt{n}$-consistency of $\hat{\eta}$ follows by the usual argument for OLS. For $w(z,\eta)$ as specified in the text and $g_0(x) = \sigma^2(x)/\phi_0$, note that for an compact neighborhood N of (β_0',ϕ_0) on which ϕ is bounded away from zero, Assumption 4.1 is satisfied with $q = \mathit{q}/2$, $d_w(z) = C(|\varepsilon|^2+\|x\|^2)$. Also note that Assumptions 3.1 and 3.2 are satisfied. First it will be shown that $\hat{R} = R + o_p(1)$ for $\hat{R} = \sum_{i=1}^n x_ix_i'/\hat{g}_i^{\delta}n$, $R =$

$\phi E[xx'/\sigma^2(x)]$. Assume for the moment that x is a scalar, and let $a^n(z,g) = x^2/\tau^\delta(g)$. Note that $E[\|a^n(z,g_0(x))\|^{1+\epsilon}] \le E[\{\|x\|^2/\sigma^2(x)\}^{1+\epsilon}] = O(1)$, by $\tau^\delta(g_0(x))^{-1} \le g_0(x)^{-1}$. Also, $\|a^n_g(z,g_0(x))\| = \|x^2/[\tau^\delta(g_0(x))]^{-2}\tau^\delta_g(g_0(x))\| \le \|x\|^2\delta^{-2}$. Thus, Assumption 4.3 is satisfied with $\|x\|^2\delta^{-2} = d_{1n}(z)$, $\nu = p/2$, $r = 2/p + 2t$, $d_{2n}(z) = 0$, and $d_{3n}(z) = 0$. Note that

$$(A.40) \qquad \gamma \ge 2(t + 1/p)/(\zeta-1) = r/(\zeta-1)$$

$$\Gamma \le 1 - 4/q - 4/p - 4t = 1 - 2/q - 2r = 2(1/2 - 2/q - 2/p - 2t).$$

It follows by Theorem 4.1 that $\hat{R} - E[x^2/\tau^\delta(g_0(x))] = o_p(1)$. Furthermore, by the dominated convergence theorem, $E[x^2/\tau^\delta(g_0(x))] - R = o(1)$, so that $\hat{R} = R + o_p(1)$ follows by $\mathscr{C}$. The same conclusion also follows in the vector x case by applying this argument to each element of $\hat{R}$. Next, it will be shown that $\sum_{i=1}^n x_i\varepsilon_i/\hat{g}_i^\delta\sqrt{n} = \sum_{i=1}^n x_i\varepsilon_i/g_i\sqrt{n} + o_p(1)$. Take $a^n(z,g) = x\varepsilon/\tau^\delta(g)$, and note that it follows as above that Assumption 4.5 is satisfied with $d_{1n}(z) = C\|x\||\varepsilon|\delta^{-2}$, $d_{2n}(z) = C\|x\||\varepsilon|\delta^{-3}$, $\nu = \nu' = pq/(p+q)$, $r = 1/\nu + 2t = 1/p + 1/q + 2t < r' = 1/p + 1/q + 3t$. Note that

$$(A.41) \qquad \gamma \ge \max\{1/p + 1/q + 2t, 1/2p + 1/2q + 3t/2 + 1/4\}/(\zeta-1)$$

$$\Gamma \le 1/2 - 2/q - r' = 1/2 - 5/q - 1/p - 3t.$$

Then by Theorem 4.3, it follows that $\sum_{i=1}^n x_i\varepsilon_i/\hat{g}_i^\delta\sqrt{n} = \sum_{i=1}^n x_i\varepsilon_i/\hat{g}_i^\delta\sqrt{n} + o_p(1)$. Note also that $\tau^\delta(g_0(x)) \ge g_0(x)$ so that by the dominated convergence theorem

$$(A.42) \qquad E[\|x\varepsilon\|^2\{\tau^\delta(g_0(x))^{-1} - g_0(x)^{-1}\}^2]$$

$$= E[\|x\|^2\sigma^2(x)\{1 - [g_0(x)/\tau^\delta(g_0(x))]\}^2/g_0(x)\}^2]$$

$$= \phi_0^2 E[\{\|x\|^2/\sigma^2(x)\}\{1 - [g_0(x)/\tau^\delta(g_0(x))]\}^2] = o(1).$$

Then by independence of the observations and $E[\varepsilon|x_i] = 0$, $E[\{\sum_{i=1}^n x_i\varepsilon_i[\tau^\delta(g_i)^{-1}-g_i{}^{-1}]/\sqrt{n}\}^2] = o_p(1)$. It then follows by the Chebyshev and triangle inequalities that $\sum_{i=1}^n x_i\varepsilon_i/\hat{g}_i^\delta\sqrt{n} = \sum_{i=1}^n x_i\varepsilon_i/g_i\sqrt{n} + o_p(1)$. The conclusion is then immediate from the usual least squares calculation. ■

Proof of Theorem 5.4: First, $\hat{\beta} = \beta_0+o_p(1)$ will be shown. Let $w(z,\beta) = \rho(z,\beta)$ and $a(z,g) = g^2$. Assumption 4.1 follows by iv), since a Taylor expansion gives $\|w(z,\tilde{\beta})-w(z,\beta)\| \le \|\rho_\beta(z,\beta)\|\|\tilde{\beta}-\beta\| \le d_1^\rho(z)\|\tilde{\beta}-\beta\|$. Also, Assumption 4.4 holds with $d_{1n}(z) = d_{2n}(z) = 0$, and Assumptions 3.1 and 3.2 hold by v) and vi). Then by Theorem 4.2, eq. (4.2) holds for $A_n(\beta) = S_n(\beta)$, $\bar{A}_n(\beta) = E[\{E[\rho(z,\beta)|x]\}^2]$. Then $\hat{\beta} = \beta_0+o_p(1)$ follows by $\mathcal{B}$ compact, $\bar{A}_n(\beta)$ having a unique minimum (of zero) at β, and Lemma 3 of Amemiya (1973).

Next, suppose for the moment that β is a scalar and consider

(A.43) $$\partial^2S_n(\beta)/\partial\beta^2/2n = \rho_\beta(\beta)'Q\rho_\beta(\beta)/n + \rho(\beta)'Q\rho_{\beta\beta}(\beta)/n$$
$$= \sum_{i=1}^n\hat{g}_2(x_i,\beta)^2/n + \sum_{i=1}^n\hat{g}_1(x_i,\beta)\rho_{\beta\beta}(z_i,\beta)/n,$$

where $\hat{g}_1(x,\beta)$ corresponds to $w(z,\beta) = \rho(z,\beta)$, $g_2(x,\beta)$ to $w(z,\beta) = \rho_\beta(z,\beta)$, and the second equality follows by Q idempotent. It follows analogously to the proof of consistency of $\hat{\beta}$ that eq. (4.2) holds for $A_n(\beta) = \sum_{i=1}^n\hat{g}_2(x_i,\beta)^2/n$ and $\bar{A}_n(\beta) = E[\{E[\rho_\beta(z,\beta)|x]\}^2]$. Also, for $a(z,\beta,g) = g\circ\rho_{\beta\beta}(z,\beta)$, Assumption 4.4 is satisfied with $d_{1n}(z) = d_2^\rho(z)$ and $d_{2n}(z) = d_2^\rho(z)^2$, so that by Theorem 4.2 eq. (4.2) holds for $A_n(\beta) = \sum_{i=1}^n\hat{g}_1(x_i,\beta)\rho_{\beta\beta}(z_i,\beta)/n$, $\bar{A}_n(\beta) = E[E[\rho(z,\beta)|x]\rho_{\beta\beta}(z,\beta)]$. Therefore, by eq. (A.43) it follows that for any $\tilde{\beta} = \beta_0+o_p(1)$,

(A.44) $$\partial^2 S_n(\tilde{\beta})/\partial\beta^2/2n = E[D(x)D(x)'] + E[E[\rho(z,\beta_0)|x]\rho_{\beta\beta}(z,\beta_0)] + o_p(1)$$
$$= E[D(x)D(x)'] + o_p(1).$$

Consider

$$\partial S_n(\beta_0)/\partial\beta/2\sqrt{n} = \rho_\beta(\beta_0)'Q\rho(\beta_0)/\sqrt{n} = \Sigma_{i=1}^n \hat{g}(x_i)\rho(z_i,\beta_0)/\sqrt{n},$$

where $\hat{g}(x)$ is the series estimate of $D(x)$ corresponding to $w(z) = \rho_\beta(z,\beta_0)$. For $a(z,g) = g\circ\rho(z,\beta_0)$, Assumptions 4.1, 4.5, and 4.6 are satisfied with $d_w(z) = d_1^\rho(z)+d_2^\rho(z)$, $\mathrm{Var}(a_g(z,g_0(x))) = \mathrm{Var}(\rho(z,\beta_0)|x) = \sigma^2$ bounded, and $a_{gg} = 0$. It then follows by Theorem 4.3 and the remarks following Theorem 4.3 that

(A.45) $$\partial S_n(\beta_0)/\partial\beta/2\sqrt{n} = \Sigma_{i=1}^n D(x_i)\rho(z_i,\beta_0)/\sqrt{n} + o_p(1).$$

Both of eqs. (A.41) and (A.45) can also be shown to hold in the vector β case, by analogous arguments for each element of the Hessian and gradient. Then by β_0 an interior point, $\partial S_n(\hat{\beta})/\partial\beta = 0$ with probability approaching one, so that by the usual Taylor expansion argument and the Lindbergh-Levy central limit theorem,

(A.46) $$\sqrt{n}(\hat{\beta}-\beta_0) = [\partial^2 S_n(\tilde{\beta})/\partial\beta^2]^{-1}\sqrt{n}\partial S_n(\beta_0)/\partial\beta + o_p(1)$$
$$= (E[D(x)D(x)'])^{-1}\Sigma_{i=1}^n D(x_i)\rho(z_i,\beta_0)/\sqrt{n} + o_p(1)$$
$$\xrightarrow{d} (E[D(x)D(x)'])^{-1}N(0,\sigma^2 E[D(x)D(x)']) \overset{d}{=} N(0,\Omega),$$

where $\tilde{\beta}$ is the mean value, giving the first conclusion. The second conclusion follows from consistency of $\hat{\beta}$ and Theorem 4.2 applied to $w(z,\beta) = \rho_\beta(z,\beta)$ and $a(z,g) = g^2$, as in the argument for eq. (A.44). ■

Proof of Theorem 5.5: First, $\hat{\beta} = \beta_0+o_p(1)$ will be shown. Note that for $w(z,\beta) = y - f(X,\beta)$, $E[w(z,\beta)|x] = h_0(x) + E[f(X,\beta_0)|x] - E[f(X,\beta)|x]$, so that Assumption 3.1 holds by v). Then it follows exactly as in the proof of Theorem 5.4 that eq. (4.2) is satisfied for $A_n(\beta) = w(\beta)'Qw(\beta)/n$, $A_n(\beta) = E[\{E[w(z,\beta)|x]\}^2]$. Also, by iv) eq. (4.2) holds for $A_n(\beta) = w(\beta)'w(\beta)/n$, $A_n(\beta) = E[w(z,\beta)^2]$. Therefore, $S_n(\beta)$ converges uniformly in probability to the continuous function

$$\text{(A.47)} \qquad S_0(\beta) = E[w(z,\beta)^2] - E[\{E[w(z,\beta)|x]\}^2] = E[\{w(z,\beta)-E[w(z,\beta)|x]\}^2]$$

$$= E[\{\varepsilon + f(X,\beta_0)-E[f(X,\beta_0)|x] - (f(X,\beta)-E[f(X,\beta)|x])\}^2]$$

$$= \sigma^2 + E[\{f(X,\beta_0)-E[f(X,\beta_0)|x] - (f(X,\beta)-E[f(X,\beta)|x])\}^2],$$

which has a unique minimum at β_0 by ii), giving consistency of $\hat{\beta}$.

Next, suppose for the moment that β is a scalar and consider

$$\partial^2 S_n(\beta)/\partial\beta^2/2n = w_\beta(\beta)'(I-Q)w_\beta(\beta)/n + w(\beta)'(I-Q)w_{\beta\beta}(\beta)/n.$$

It follows as in the proof of Theorem 5.4, and by a standard argument for the terms not involving Q, that for any $\tilde{\beta} = \beta + o_p(1)$,

$$\text{(A.48)} \qquad \partial^2 S_n(\tilde{\beta})/\partial\beta^2/2n = E[f_\beta f_\beta'] - E[E[f_\beta|x]E[f_\beta|x]'] + E[(y-f)f_{\beta\beta}] -$$

$$E[E[(y-f)|x]f_{\beta\beta}] + o_p(1) = E[Var(f_\beta|x)] + E[\varepsilon\circ f_{\beta\beta}] + o_p(1)$$

$$= E[Var(f_\beta|x)] + o_p(1).$$

where $f = f(z,\beta_0)$, $f_\beta = f_\beta(z,\beta_0)$, and $f_{\beta\beta} = f_{\beta\beta}(z,\beta_0)$. Consider

$$\text{(A.49)} \qquad \partial S_n(\beta_0)/\partial\beta/2\sqrt{n} = w_\beta(\beta_0)'(I-Q)w(\beta_0)/\sqrt{n}$$

$$= \textstyle\sum_{i=1}^n [w_2(z_i)-\hat{g}_2(x_i)][w_1(z_i)-\hat{g}_1(x_i)]/\sqrt{n},$$

where $\hat{g}_1(x)$ and $\hat{g}_2(x)$ are the series estimates corresponding to $w_1(z) = y-f$ and $w_2(z) = -f_\beta$ respectively, and the last equality follows by I-Q idempotent. For $a(z,g_1,g_2) = [w_2(z)-g_2][w_1(z)-g_1]$ Assumptions 4.1 and 4.5 are satisfied with $d_w(z) = d_1^f(z)+|h_0(x)|+|\varepsilon|$, a_{gg} is bounded and $\nu = q$, $r = 1/q$. It then follows by Theorem 4.3 and the remarks following it that

$$\text{(A.50)} \qquad \partial S_n(\beta_0)/\partial\beta/2\sqrt{n} = \textstyle\sum_{i=1}^n [w_2(z_i)-g_2(x_i)][w_1(z_i)-g_1(x_i)]/\sqrt{n} + o_p(1).$$

Both of eqs. (A.48) and (A.50) can also be shown to hold in the vector β case, by analogous arguments for each element of the Hessian and gradient. The first conclusion then follows by arguing analogously to the proof of Theorem 5.4. The second conclusion follows from consistency of $\hat{\beta}$, uniform convergence in probability of $S_n(\beta)/n$ to $S_0(\beta)$, $S_0(\beta_0) = \sigma^2$, $\hat{K}/n = o_p(1)$, and eq. (A.48). ■

References

Amemiya, T. (1973): "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, 41, 997-1016.

Amemiya, T. (1974): "The Nonlinear Two-Stage Least-Squares Estimator," *Journal of Econometrics*, 2, 105-110.

Andrews, D.W.K. (1988): "Asymptotic Normality of Series Estimators for Various Nonparametric and Semiparametric Models," mimeo, Yale University.

Bickel P., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1989): "Efficient and Adaptive Inference in Semiparametric Models" monograph, Johns Hopkins University Press, forthcoming.

Carroll, R.J. (1982): "Adapting for Heteroskedasticity in Linear Models," *Annals of Statistics*, 10, 1224-1233.

Chamberlain, G. (1986): "Notes on Semiparametric Regression," mimeo, Department of Economics, Harvard University.

Chamberlain, G. (1987): "Efficiency Bounds for Semiparametric Regression," mimeo, Department of Economics, Harvard University.

Edmunds, D.E. and V.B. Moscatelli (1977): "Fourier Approximation and Embedding in Sobolev Space," *Dissertationes Mathematicae,* 145, 1-46.

Engle, R.F., C.W.J. Granger, J. Rice, and A. Weiss (1986): "Semiparametric Estimates of the Relation Between Weather and Electricity Sales." *Journal of the American Statistical Association*, 81, 310-320.

Gallant, A.R. (1981): "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics*, 15, 211-245.

Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

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

Heckman, N.E. (1986): "Spline Smoothing in a Partly Linear Model," *Journal of the Royal Statistical Society*, Series B, 48, 244-248.

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

Kelejian, H. (1971): "Two-Stage Least Squares and Econometric Systems Linear in Parameters and Nonlinear in Endogenous Variables," *Journal of the American Statistical Association*, 66, 373-374.

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

Manski, C. (1988): "Path Utility Analysis of Dynamic Choice," SSRI Working Paper No. 8805, University of Wisconsin.

Newey, W.K. (1987): "Efficient Estimation of Models with Conditional Moment Restrictions," mimeo, Department of Economics, Princeton University.

Newey, W.K. (1988): "Two Step Series Estimation of Sample Selection Models," mimeo, Department of Economics, Princeton University.

Newey, W.K. (1989a): "Efficient Instrumental Variables Estimation of Nonlinear Models," mimeo, Department of Economics, Princeton University.

Newey, W.K. (1989b): "Semiparametric Efficiency Bounds," mimeo, Department of Economics, Princeton University.

Newey, W.K. (1989c): "Uniform Convergence in Probability and Uniform Stochastic Equicontinuity," mimeo, Department of Economics, Princeton University.

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

Pitman, E.J.G. (1979): *Some Basic Theory for Statistical Inference*, London: Chapman and Hall.

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

Rice, J. (1986): "Convergence Rates for Partially Splined Estimates," *Statistics and Probability Letters*, 4, 203-208.

Rilstone, P. (1989): "Semiparametric Estimation of Missing Data Models," mimeo, Department of Economics, Laval University.

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

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

Schick, A. (1986): "On Asymptotically Efficient Estimation in Semiparametric Models," *Annals of Statistics*, 14, 1139-1151.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press.

Yatchew, A. (1988): "Nonparametric Regression Tests Based on an Infinite Dimensional Least Squares Procedure," mimeo, Department of Economics, University of Toronto.

White, H. (1980): "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.